On aggregation for heavy-tailed classes

Shahar Mendelson

Abstract. We introduce an alternative to the notion of ‘fast rate’ in Learning Theory, which coincides with the optimal error rate when the given class happens to be convex and regular in some sense. While it is well known that such a rate cannot always be attained by a learning procedure (i.e., a procedure that selects a function in the given class), we introduce an aggregation procedure that attains that rate under rather minimal assumptions—for example, that the $L_q$ and $L_2$ norms are equivalent on the linear span of the class for some $q > 2$, and the target random variable is square-integrable. The key components in the proof include a two-sided isomorphic estimator on distances between class members, which is based on the median-of-means; and an almost isometric lower bound of the form $\frac{1}{N} \sum_{i=1}^{N} f^2(X_i) \geq (1 - \zeta) E f^2$ which holds uniformly in the class. Both results only require that the $L_q$ and $L_2$ norms are equivalent on the linear span of the class for some $q > 2$.

Mathematics Subject Classification 62G99 · 60G25 · 68T01

1 Introduction

The focus of this article is on the question of Prediction: let $\mathcal{F}$ be a class of functions defined on a probability space $(\Omega, \mu)$ and let $X$ be distributed according to $\mu$. Given an unknown target random variable $Y$, one would like to find some $f \in \mathcal{F}$ for which,
on average, predicting $f(X)$ instead of $Y$ is the most ‘cost effective’. If the pointwise cost is measured according to the squared loss, that is, if the price of predicting $f(X)$ instead of $Y$ is $(f(X) - Y)^2$, the goal is to identify, or at least approximate in some sense, the behaviour of the function that minimizes in $F$ the risk $E(f(X) - Y)^2$, where the expectation is taken with respect to the joint distribution of $X$ and $Y$ on $\Omega \times \mathbb{R}$. With that in mind, set

$$f^* = \arg\min_{f \in F} E(f(X) - Y)^2,$$

and assume, for the sake of simplicity, that the minimizer exists.

It should also be noted that there are other reasonable choices for the pointwise cost of predicting $f(X)$ instead of $Y$, and although our results are presented only for the squared loss, they may be extended to other convex loss functions, following the path of [19].

Unlike standard questions in Approximation Theory, in the prediction framework one has limited information: a random sample $(X_i, Y_i)_{i=1}^N$, selected independently according to the joint distribution of $X$ and $Y$. The hope is that a typical sample may be used to produce a (random) function in $F$ that has almost the same ‘predictive capabilities’ as the minimizer $f^*$.

**Definition 1.1** For every integer $N$ and a base class $F$, a learning procedure$^1$ is a function $\Psi : (\Omega \times \mathbb{R})^N \rightarrow F$.

Setting $\tilde{f} = \Psi((X_i, Y_i)_{i=1}^N)$, and given $0 < \delta < 1$ and a set of potential targets $\mathcal{Y}$, the learning procedure $\Psi$ performs with an error rate of $\mathcal{E}_p(F, N, \delta)$ if for every reasonable$^2$ class of functions $F \subset L_2(\mu)$ and $Y \in \mathcal{Y}$,

$$E((\tilde{f}(X) - Y)^2|(X_i, Y_i)_{i=1}^N) \leq E(f^*(X) - Y)^2 + \mathcal{E}_p(F, N, \delta)$$

with probability at least $1 - \delta$ relative to the samples $(X_i, Y_i)_{i=1}^N$ (i.e., with respect to the $N$-product of the joint distribution of $X$ and $Y$ endowed on $(\Omega \times \mathbb{R})^N$).

One would like to identify the best learning procedure $\Psi$ in the sense that the error rate $\mathcal{E}_p$ is as small as possible; find what features of $F$ and $\mathcal{Y}$ govern $\mathcal{E}_p$; and study the way in which $\mathcal{E}_p$ scales with the sample size $N$.

Although it is not obvious from Definition 1.1, the effect of the set of admissible targets $\mathcal{Y}$ has on the error rate $\mathcal{E}_p$ is rather small. In standard scenarios, $\mathcal{Y}$ consists of all random variables that are bounded by 1 or, alternatively, that have rapidly decaying tails (e.g., subgaussian or subexponential). However, as will be explained later, this type of condition can be relaxed considerably, and $\mathcal{Y}$ may be as large as the $L_2$ unit ball on the underlying probability space, rather than the $L_\infty$ one.

$^1$ This is sometimes called a proper learning procedure.

$^2$ Naturally, there is no chance that a nontrivial error rate would be true without any assumption on the class $F$. The aim is to obtain such an estimate under minimal assumptions on the class—as will be specified in what follows.
1.1 Fast and slow rates

One frequently encounters in literature the terms ‘fast rate’ and ‘slow rate’, used to describe the behaviour of a learning procedure as a function of the sample size $N$. Unfortunately, the meaning of the two is somewhat ambiguous, and is often misinterpreted.

A common misapprehension is that ‘fast rate’ means that $E_p$ scales as $1/N$, and that a ‘slow rate’ implies that $E_p$ is of the order of $1/\sqrt{N}$, though in reality, the situation is different. Indeed, on one hand, it is straightforward to construct examples of classes that are simply too rich for a rate of $1/N$ (or even of $1/\sqrt{N}$, for that matter), even when the problem is realizable, i.e., when the target $Y$ happens to belong to $\mathcal{F}$. On the other, the ‘size’ of $\mathcal{F}$ does not capture the correct behaviour of $E_p$: if $\mathcal{F} = \{f_1, f_2\}$ and $Y$ happens to be a noisy $1/\sqrt{N}$ perturbation of the mid-point $(f_1 + f_2)/2$, no learning procedure can achieve an error rate that is better than $c/\sqrt{N}$ with probability at least $3/4$ using $N$ sample points and for a suitable absolute constant $c$ (see, e.g., [1] for a more precise statement).

Thus, a reasonable definition of the terms ‘fast rate’ and ‘slow rate’ must reflect the fact that the error rate is highly affected by the ‘location’ of the target, as well as by the ‘complexity’ of $\mathcal{F}$.

To avoid potential ambiguity, we will refrain from using the terms ‘fast rate’ and ‘slow rate’ in what follows. Instead, we will adopt the notion of ‘optimistic rate’, which is, roughly put, the rate one encounters when the location of the target is favourable, and should be considered as a more accurate version of the intuitive ‘fast rate’ (see Sect. 1.2 for the definition). For example, if $\mathcal{F} = \{f_1, f_2\}$, the optimistic rate is of the order of $1/N$ rather than $1/\sqrt{N}$, seemingly ignoring the possibility that $Y$ is a perturbation of the mid-point $(f_1 + f_2)/2$ as above.

This example shows that the optimistic rate may be, at times, unreachable by any learning procedure. Hence, if there is any hope of constructing a procedure that always attains the optimistic rate regardless of the location of the target, that procedure must be allowed the flexibility of selecting functions that are outside the given base class $\mathcal{F}$. Such procedures belong to the model selection aggregation framework.

Definition 1.2 For an integer $N$, an aggregation procedure is a map $\Psi : (\Omega \times \mathbb{R})^N \to L_2(\mu)$. The procedure has an error rate of $\epsilon^{\text{agg}}_p(\mathcal{F}, N, \delta)$ if for every reasonable class of functions $\mathcal{F}$ and every target $Y \in \mathcal{Y}$,

$$
\mathbb{E}\left((\hat{f}(X) - Y)^2|(X_i, Y_i)_{i=1}^N\right) \leq \mathbb{E}(f^*(X) - Y)^2 + \epsilon^{\text{agg}}_p(\mathcal{F}, N, \delta)
$$

with probability at least $1 - \delta$ relative to the $N$-product of the joint distribution of $X$ and $Y$, and for $\hat{f} = \Psi((X_i, Y_i)_{i=1}^N)$.

Detailed surveys on the aggregation framework in a broad context may be found in [5, 22, 25].

Therefore, rather than restricting one to a learning procedure, i.e., forcing one to select functions from $\mathcal{F}$, the goal here is to construct an aggregation procedure that attains the optimistic rate under minimal assumptions on $\mathcal{F}$ and $\mathcal{Y}$.
The analysis of the aggregation procedure we will introduce below requires the use of some auxiliary classes that are connected to the given base class $\mathcal{F}$; those will be denoted by $U$, $V$ and $H$. To avoid confusion, in what follows we will denote ‘generic’ function classes by $\mathcal{F}$ and $\mathcal{K}$.

1.2 The optimistic rate

The definition of the optimistic rate is based on the small-ball method, which was developed in [18,19] for the analysis of the Empirical Risk Minimization (ERM) procedure. We will outline the essentials of the method in what follows, and refer the reader to [18,19] for a more detailed description of the parameters involved, their role in the analysis of ERM and the way in which they may be controlled in specific applications.

Definition 1.3 Given a sample $(X_i, Y_i)_{i=1}^N$ and a base class $\mathcal{F}$, the empirical minimizer in $\mathcal{F}$ is

$$\hat{f} \in \arg\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} (f(X_i) - Y_i)^2,$$

assuming, of course, that a minimizer exists.

From here on we will denote by $P_N h$ the empirical mean $\frac{1}{N} \sum_{i=1}^{N} h(X_i, Y_i)$.

Recall that $f^* = \arg\min_{f \in \mathcal{F}} \mathbb{E} (f(X) - Y)^2$, consider the excess squared loss functional relative to $\mathcal{F}$ and $Y$,

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

and observe that the minimizer in $\mathcal{F}$ of $P_N (f - Y)^2$ is also a minimizer in $\mathcal{F}$ of $P_N L_f$. Thus, $P_N L_f \leq 0$, simply because $L_{f^*} = 0$, and, in particular,

$$\hat{f} \in \{ f \in \mathcal{F} : P_N L_f \leq 0 \}.$$

It follows that if $(X_i, Y_i)_{i=1}^N$ is a sample for which

$$\{ f \in \mathcal{F} : \mathbb{E} L_f \geq \eta \} \subset \{ f \in \mathcal{F} : P_N L_f > 0 \},$$

then

$$\mathbb{E} \left( (\hat{f}(X) - Y)^2 | (X_i, Y_i)_{i=1}^N \right) \leq \mathbb{E} (f^*(X) - Y)^2 + \eta,$$

which is the type of result one is looking for.
To obtain (1.1), note that for every $f \in \mathcal{F}$ and every sample $(X_i, Y_i)_{i=1}^N$,

$$P_N \mathcal{L}_f \geq \frac{1}{N} \sum_{i=1}^N (f - f^*)^2(X_i) + 2\mathbb{E}(f^*(X) - Y)(f - f^*)(X) - 2 \left| \frac{1}{N} \sum_{i=1}^N (f^*(X_i) - Y_i)(f - f^*)(X_i) - \mathbb{E}(f^*(X) - Y)(f - f^*)(X) \right|,$$  \hspace{1cm} (1.2)

i.e., $P_N \mathcal{L}_f$ is lower bounded by a sum of (random) quadratic and multiplier components, and a deterministic term, $2\mathbb{E}(f^*(X) - Y)(f - f^*)(X)$, which calibrates the ‘location’ of the target $Y$ relative to $\mathcal{F}$.

The optimistic rate is defined based on the belief that the location of $Y$ is favourable in the sense that for every $f \in \mathcal{F}$,

$$\mathbb{E}(f^*(X) - Y)(f - f^*)(X) \geq 0. \hspace{1cm} (1.3)$$

It is straightforward to verify that (1.3) is satisfied in two important cases. Firstly, when $\mathcal{F} \subset L_2(\mu)$ happens to be closed and convex, in which case, (1.3) follows from the characterization of the metric projection onto a closed, convex set in an inner-product space. Secondly, for an arbitrary class $\mathcal{F}$ and a target $Y = f^*(X) + \xi$, where $f^* \in \mathcal{F}$ and $\xi$ is mean-zero and independent of $X$.

It turns out that a lower bound on the quadratic component in (1.2) follows from a rather minimal assumption:

**Definition 1.4** A class $\mathcal{F} \subset L_2(\mu)$ satisfies a small-ball condition with constants $\kappa_0$ and $\varepsilon$, if for every $f_1, f_2 \in \mathcal{F} \cup \{0\}$,

$$Pr(\|f_1 - f_2\| \geq \kappa_0 \|f_1 - f_2\|_{L_2(\mu)}) \geq \varepsilon. \hspace{1cm} (1.4)$$

The small-ball condition is simply a uniform lower estimate on the probability that $\|f_1 - f_2\|/\|f_1 - f_2\|_{L_2(\mu)}$ is sufficiently far from zero for every pair of distinct functions $f_1, f_2 \in \mathcal{F} \cup \{0\}$.

One may find in [18,19] several examples of classes that satisfy a small-ball condition. For our purposes, the most significant example is when $q > 2$ and the $L_q$ and $L_2$ norms are $L$-equivalent on $\mathcal{F}$, in the sense that for every $f_1, f_2 \in \mathcal{F} \cup \{0\}$, $\|f_1 - f_2\|_{L_q} \leq L\|f_1 - f_2\|_{L_2}$. In such a case, the Paley-Zygmund inequality [6] shows that (1.4) holds for constants $\kappa_0$ and $\varepsilon$ that depend only on $q$ and $L$.

Let $D$ be the unit ball in $L_2(\mu)$. Given a class of functions $\mathcal{F} \subset L_2(\mu)$, the canonical gaussian process indexed by $\mathcal{F}$ is $\{G_f : f \in \mathcal{F}\}$, where each $G_f$ is a gaussian variable, and the covariance structure of the process is endowed by the inner product in $L_2(\mu)$; thus, for every $f, h \in \mathcal{F}$, $\mathbb{E}(G_f - G_h)^2 = \|f - h\|^2_{L_2(\mu)}$. Let

$$\mathbb{E}\|G\|_{\mathcal{F}} = \sup\{\mathbb{E} \sup_{f \in \mathcal{F}'} G_f : \mathcal{F} \subset \mathcal{F}', \mathcal{F}' \text{ is finite}\}.$$
Definition 1.5 For $\mathcal{F} \subset L_2(\mu)$, let $\text{star}(\mathcal{F}) = \{ \lambda f : 0 \leq \lambda \leq 1, f \in \mathcal{F} \}$ be the star-shaped hull of $\mathcal{F}$ around 0, and set $\mathcal{F} - \mathcal{F} = \{ f - h : f, h \in \mathcal{F} \}$. Let

$$U = \left\{ \frac{f_1 + f_2}{2} : f_1, f_2 \in \mathcal{F} \right\}$$

and set $H = \text{star}(U - U)$.

Finally, for $\zeta > 0$, let

$$r_{Q,1}(\mathcal{F}, \zeta) = \inf \left\{ r > 0 : \mathbb{E} \| G \|_{(H-H) \cap rD} \leq \zeta r \sqrt{N} \right\},$$

and

$$r_{Q,2}(\mathcal{F}, \zeta) = \inf \left\{ r > 0 : \mathbb{E} \sup_{w \in (H-H) \cap rD} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i w(X_i) \right| \leq \zeta r \sqrt{N} \right\},$$

where $(\varepsilon_i)_{i=1}^{N}$ are independent, symmetric $\{-1, 1\}$-valued random variables that are independent of $(X_i)_{i=1}^{N}$, and the expectation is taken with respect to both $(\varepsilon_i)_{i=1}^{N}$ and $(X_i)_{i=1}^{N}$.

Note that $U$ is only slightly richer than $\mathcal{F}$; it contains $\mathcal{F}$ and all the midpoints of intervals whose ends belong to $\mathcal{F}$. If $\mathcal{F}$ happens to be convex, then $U = \mathcal{F}$, but in general, $U$ is much smaller than the convex hull of $\mathcal{F}$. Also, $H = \text{star}(U - U)$ is star-shaped around 0, centrally symmetric, and contains $\mathcal{F} - \mathcal{F}$; hence, both $\mathcal{F}$ and $\mathcal{F} - \mathcal{F}$ belong to $H - H$.

The parameters $r_{Q,1}$ and $r_{Q,2}$ measure the ‘local’ complexity of the indexing class, and from a statistical point of view, the two capture the correlation of the indexing class with various forms of random noise. The reader may find a more detailed explanation of their role in [18] and [19].

It should be noted that the definitions of $r_{Q,1}$ and $r_{Q,2}$ in [19] appear to be slightly different from the ones defined above. However, the reason for the difference is that in [19] one considers a convex base class, while here $\mathcal{F}$ need not be convex. If $\mathcal{F}$ happens to be convex then $U = \mathcal{F}$, $U - U = \mathcal{F} - \mathcal{F}$ is convex and centrally symmetric, and $H = \mathcal{F} - \mathcal{F}$; therefore $H - H = 2H = 2(\mathcal{F} - \mathcal{F})$ and the definitions above coincide with the ones from [19] up to a factor of 2, which is only an issue of normalization.

The third and final complexity parameter is also a minor modification of a similar parameter from [18,19]. It will be used to study the multiplier component in the decomposition (1.2) of the excess squared-loss functional.

Definition 1.6 Let $\mathcal{F} \subset L_2(\mu)$ be the given base class and set $U$ and $H$ as above. For every $u_0 \in U$ consider the random function

$$\phi_{\mathcal{F},N,u_0}(r) = \frac{1}{\sqrt{N}} \sup_{w \in \text{star}(U - u_0) \cap rD} \left| \sum_{i=1}^{N} \varepsilon_i (u_0(X_i) - Y_i) w(X_i) \right|.$$
and set
\[ r_M(\mathcal{F}, \zeta, \delta, u_0) = \inf \left\{ r > 0 : \Pr \left( \phi_{\mathcal{F}, N, u_0}(r) \leq r^2 \zeta \sqrt{N} \right) \geq 1 - \delta \right\}. \]

The importance of \( r_M \) and the way it may be used to upper bound the multiplier component can be seen in the next lemma from [19]:

**Lemma 1.7** Let \( u_0 \in U, 0 < \delta < 1, \kappa > 0 \) and set \( r = 2r_M(\mathcal{F}, \kappa/4, \delta/2, u_0) \). Put \( \xi = u_0(X) - Y \), and given a sample \((X_i, Y_i)_{i=1}^N\) set \( \xi_i = u_0(X_i) - Y_i \). Then, with probability at least \( 1 - \delta \), for every \( u \in U \), one has
\[
\left| \frac{1}{N} \sum_{i=1}^N \xi_i (u - u_0)(X_i) - \mathbb{E} \xi(u - u_0)(X) \right| \leq \kappa \max \{ \|u - u_0\|_{L^2(\mu)}^2, r^2 \}.
\]

With all the complexity parameters in place, one may derive an estimate on the error rate of ERM, performed in any subset of \( U \) that satisfies the ‘optimistic’ assumption. Indeed, the following is a minor modification of Lemma 5.2 from [19], originally formulated for a convex base class, though it is straightforward to verify that the convexity condition may be relaxed.

In the setup we are interested in, \( \mathcal{F} \subset L_2(\mu) \) is the given base class, \( U \) and \( H \) are defined as above and \( H \) satisfies a small-ball condition with constants \( \kappa_0 \) and \( \varepsilon \). Fix \( Y \in L_2 \) and \( V \subset U \), put \( v^* = \arg\min_{v \in V} \|v - Y\|_{L^2(\mu)} \) and let \( \hat{v} \) be the empirical minimizer in \( V \) of the squared loss functional.

**Theorem 1.8** There exists an absolute constant \( c_0 \) and for every \( \kappa_0 > 0 \) and \( 0 < \varepsilon < 1 \) there exist constants \( c_1, c_2 \) and \( c_3 \) that depend only on \( \kappa_0 \) and \( \varepsilon \) for which the following holds. If for every \( v \in V \), \( \mathbb{E}(v^*(X) - Y)(v - v^*)(X) \geq 0 \), then probability at least \( 1 - \delta - 2 \exp(-c_0 \varepsilon^2 N) \),
\[
\mathbb{E} \left( (\hat{v}(X) - Y)^2 \middle| (X_i, Y_i)_{i=1}^N \right) \leq \mathbb{E}(v^*(X) - Y)^2 + r^2(v^*),
\]
where
\[
r(v^*) = 2 \max \left\{ r_M(\mathcal{F}, c_1, \delta/2, v^*), r_{Q,1}(\mathcal{F}, c_2), r_{Q,2}(\mathcal{F}, c_3) \right\}.
\]

Since \( v^* \) is not known, one has to use a uniform version of \( r(v^*) \) as a complexity parameter. This uniform version is the optimistic rate:

**Definition 1.9** Given a base class \( \mathcal{F} \), a set of targets \( \mathcal{Y} \) and for constants \( c_1, c_2, c_3 \) and \( \delta \), the optimistic rate in \( \mathcal{F} \) is defined by
\[
r_{\text{opt}} = 2 \sup_{Y \in \mathcal{Y}} \sup_{u_0 \in U_Y} \max \left\{ r_M^2(\mathcal{F}, c_1, \delta/2, u_0), r_{Q,1}^2(\mathcal{F}, c_2), r_{Q,2}^2(\mathcal{F}, c_3) \right\},
\] (1.5)
where for every \( Y \in \mathcal{Y} \),
\[
U_Y = \{ u \in U : \mathbb{E}(u(X) - Y)^2 \leq \mathbb{E}(f^*(X) - Y)^2 \}.
\] (1.6)
In what follows, \( H = \text{star}(U - U) \) will satisfy the small-ball condition with constants \( \kappa_0 \) and \( \epsilon \), and \( c_1, c_2 \) and \( c_3 \) will be chosen as constants that depend only on \( \kappa_0 \) and \( \epsilon \). To avoid cumbersome notation, we will not specify in what follows that \( r_{\text{opt}} \) depends on \( \mathcal{F} \), \( \delta \) and the constants \( c_1, c_2, c_3 \), but their choice will be made clear.

When \( \mathcal{F} \) happens to be convex, \( U = \mathcal{F} \), and upon selecting \( V = \mathcal{F} \) it follows that \( \mathbb{E}(f^*(X) - Y)(f - f^*)(X) \geq 0 \) for every \( f \in \mathcal{F} \). Thus, Theorem 1.8 extends the main result from [18] on the performance of ERM in a convex class and relative to the squared loss. Moreover, since the ‘optimistic assumption’ includes the choice of \( Y = f^*(X) + \xi \) for \( f^* \in \mathcal{F} \) and \( \xi \) that is mean-zero and independent of \( X \), the results from [12,18] indicate that \( r_{\text{opt}} \) captures the minimax rate\(^3\) in \( \mathcal{F} \) under mild structural assumptions on that class.

With this in mind, the optimistic rate \( r_{\text{opt}} \) is defined as what is essentially the best possible rate that any learning procedure may achieve in \( \mathcal{F} \) when the target \( Y \) is in a ‘good location’ relative to \( \mathcal{F} \) in the sense of (1.3). And, when \( \mathcal{F} \) happens to be convex, every target \( Y \in L_2 \) is in a ‘good location’.

Having said that, let us emphasize once again that the problem we wish to address occurs when the location of the target is less favourable, and in which case no learning procedure can achieve the optimistic rate.

We will show that there is an aggregation procedure that always achieves the optimistic rate when the \( L_2(\mu) \) and \( L_q(\mu) \) norms are equivalent on \( \text{span}(\mathcal{F}) \) for some \( q > 2 \), and \( Y \in L_2 \) – thus overcoming the possible problem that may occur when the target \( Y \) is not in a ‘good location’ relative to the given class. The point is, of course, that the procedure takes values anywhere in \( L_2(\mu) \), rather than just in \( \mathcal{F} \).

**Theorem 1.10** For every \( L \geq 1 \) and \( q > 2 \) there are constants \( c_0, c_1, c_2 \) and \( c_3 \) that depend only on \( L \) and \( q \) for which the following holds. Let \( \mathcal{F} \subseteq L_2(\mu) \) be the given base class and let \( U \) and \( H \) be as above. Assume that for every \( w \in H - H \), \( \|w\|_{L_q(\mu)} \leq L \|w\|_{L_2(\mu)} \). Then, there is an aggregation procedure \( \Psi : (\Omega \times \mathbb{R})^N \to L_2(\mu) \) for which, for every \( Y \in L_2 \), with probability at least \( 1 - \delta - 2 \exp(-c_0 N) \),

\[
\mathbb{E}(\hat{f}(X) - Y)^2|\{(X_i, Y_i)_{i=1}^N\}) \leq \mathbb{E}(f^*(X) - Y)^2 + r_{\text{opt}},
\]

where \( \hat{f} = \Psi((X_i, Y_i)_{i=1}^N) \) and

\[
r_{\text{opt}} = 2 \sup_{Y \in \mathcal{Y}} \sup_{u_0 \in U_Y} \max \left\{ r_M^2(\mathcal{F}, c_1, \delta/4, u_0), r_{Q,1}^2(\mathcal{F}, c_2), r_{Q,2}^2(\mathcal{F}, c_3) \right\}.
\]

To put Theorem 1.10 is some perspective, note that in the standard framework of aggregation, \( \mathcal{F} \) is a finite dictionary and both the dictionary and the target are bounded in \( L_\infty \) (see, for example, [10,25] and references therein). Within that framework one has the following:

**Theorem 1.11** [11] There exists an aggregation procedure \( \Psi \) for which the following holds. Assume that \( \mathcal{F} \) is a finite dictionary consisting of functions that are bounded by

---

\(^3\) Roughly put, the minimax rate is the best possible error rate one may achieve by any learning procedure, i.e., by any \( \Psi : (\Omega \times \mathbb{R})^N \to \mathcal{F} \).
1, and assume that the target \( Y \) is bounded by 1 as well. Then, for every \( x > 0 \), with probability at least \( 1 - 2 \exp(-x) \),
\[
\mathbb{E} \left( (\tilde{f}(X) - Y)^2 | (X_i, Y_i)_{i=1}^N \right) \leq \mathbb{E} (f^*(X) - Y)^2 + c(1 + x) \frac{\log |\mathcal{F}|}{N},
\]
where \( \tilde{f} = \Psi((X_i, Y_i)_{i=1}^N) \).

In comparison, when applied to a finite dictionary, Theorem 1.10 leads to the following:

**Corollary 1.12** Let \( \mathcal{F} \) be a finite dictionary and set \( H \) as above. Assume that \( H - H \) is \( L \)-subgaussian, in the sense that for every \( w_1, w_2 \in H - H \), and every \( p \geq 2 \),
\[
\|w_1 - w_2\|_{L_p(\mu)} \leq L \sqrt{p} \|w_1 - w_2\|_{L_2(\mu)}.
\]
If \( Y \in L_q \) for some \( q > 2 \) then with probability at least \( 1 - \delta \),
\[
\mathbb{E} ((\tilde{f}(X) - Y)^2 | (X_i, Y_i)_{i=1}^N) \leq \mathbb{E} (f^*(X) - Y)^2 + c_1 \|f^* - Y\|_{L_q}^2 \frac{\log |\mathcal{F}|}{N}, \tag{1.8}
\]
where \( c_1 \) depend only on \( q, L \) and \( \delta \).

The proof of Corollary 1.12 will be presented in Sect. 5.1. It is well known that the best error rate a learning procedure may attain for a finite dictionary is of the order of \( \sqrt{\log |\mathcal{F}|/N} \) (see, e.g., [10,25]), which is not remotely close to \( r_{opt} \), as the latter is of the order of \( (\log |\mathcal{F}|)/N \). Moreover, the error rate in (1.8) scales well with the noise-level\(^4 \) \( \|f^* - Y\|_{L_q} \); it tends to zero as the problem becomes ‘more realizable’, in which case one expects a zero-error when \( N \geq c \log |\mathcal{F}| \).

The aggregation procedure we will introduce here is a ‘close family member’ of the one from [11], but with many significant and unavoidable changes.

**Remark 1.13** It should be noted that Audibert obtained in [2] the same estimate as in Theorem 1.11, but using a different aggregation procedure—the *empirical star algorithm*, and it is not clear whether it is possible to obtain a version of Theorem 1.10 using an analog of that algorithm. Moreover, the empirical star algorithm involves running ERM on the star-hull of \( \mathcal{F} \), which is an infinite class even if the dictionary is finite. In contrast, the procedure suggested here uses ERM on a well-chosen \( V \subset \mathcal{U} \); hence, if \( \mathcal{F} \) is finite, so is \( V \).

Unlike the bounded case, aggregation in unbounded situations was not fully understood. The benchmark result in that direction is due to Audibert [3] and independently to Juditsky, Rigollet and Tsybakov [8], who obtained the following estimate on the expected risk when the class is bounded but the target may be unbounded:

\(^4\) The reason for calling \( f^*(X) - Y \) ‘noise level’ is the independent noise case, in which \( Y = f^*(X) - \xi \) and \( \xi \) is mean-zero and independent of \( X \). Thus \( f^*(X) - Y \) is indeed the noise, and its \( L_q \) norm calibrates the noise level of the problem.
Theorem 1.14 There is an aggregation procedure \( \Psi : (\Omega \times \mathbb{R})^N \to L_2(\mu) \) for which the following holds. Assume that \( \mathcal{F} \) is a finite dictionary consisting of functions bounded by 1 and assume that \( Y \in L_q \) for \( q \geq 2 \). Then setting \( \tilde{f} = \Psi((X_i, Y_i)_{i=1}^N) \),

\[
\mathbb{E}
\left( \mathbb{E} \left( (\tilde{f}(X) - Y)^2 | (X_i, Y_i)_{i=1}^N \right) \right) \leq \mathbb{E}(f^*(X) - Y)^2 + C \left( \frac{\log |\mathcal{F}|}{N} \right)^{2/(q+2)}
\]

for a constant \( C = C(q, \|Y\|_{L_q}) \).

Moreover, this estimate is optimal—up to the constant \( C \).

It is interesting to see the subtle differences between the assumptions used in Theorem 1.14 and the ones from Corollary 1.12. In the former, the class is assumed to be bounded in \( L_\infty \), while in the latter, the class is \( L \)-subgaussian, which is a different type of condition: it implies norm equivalence rather than having a bounded diameter with respect to some (possibly strong) norm.

As noted in [13], statistical procedures may behave in a very different way when one assumes even a weak norm equivalence rather than an \( L_\infty \) bound, and the same phenomenon is true here as well: although the dictionary may consist of unbounded functions, the norm equivalence gives sufficient information to ensure an error rate of \( N^{-1} \log |\mathcal{F}| \) rather than much slower \( (N^{-1} \log |\mathcal{F}|)^{2/(q+2)} \). Moreover, the error rates in Theorems 1.11 and 1.14 do not scale well with the ‘noise level’ and do not improve even when the problem is arbitrarily close to being realizable.

The proof of Theorem 1.10 is based on two key facts that are of independent interest and which will be presented in the next two sections. The first is an almost isometric (rather than isomorphic) lower bound of the form

\[
\inf_{\{f \in \mathcal{F}, \mathbb{E}f^2 \geq r(\xi)\}} \frac{N^{-1} \sum_{i=1}^N f^2(X_i)}{\mathbb{E}f^2} \geq 1 - \xi
\]

for any \( 0 < \xi < 1 \); the second is a satisfactory alternative to the empirical means \( N^{-1} \sum_{i=1}^N (f - h)^2(X_i) \) as two-sided estimators for the distances \( \|f - h\|_{L_2}^2 \) — uniformly for every \( f, h \in \mathcal{F} \). Both facts require a rather minimal assumption: that the \( L_q \) and \( L_2 \) norms are equivalent on \( \text{span}(\mathcal{F}) \) for some \( q > 2 \).

The need for an alternative to the empirical mean as a two-sided estimator of \( L_2 \)-distances is that its performance is poor in heavy-tailed situations. Indeed, one ‘bad’ sample point can make the empirical mean \( N^{-1} \sum_{i=1}^N (f - h)^2(X_i) \) very large – much larger than \( \|f - h\|_{L_2(\mu)}^2 \). And, if \( f - h \) is heavy tailed, encountering such a sample point is likely, even for a single difference \( f - h \), let alone if one is looking for an estimate that holds uniformly in the class.

We will show that the alternative, based on the median-of-means, can be used as an isomorphic, two-sided estimator for the \( L_2 \) distances uniformly in a class, even in a heavy-tailed situation.

We believe that besides being the crucial components in the proof of Theorem 1.10, the two facts are likely to have many other applications.

We end this introduction with some notation. Throughout, absolute constants are denoted by \( c, c_1, \ldots \), etc. Their value may change from line to line. When a constant
depends on a parameter $\alpha$ it will be denoted by $c(\alpha)$. $A \preceq B$ means that $A \leq cB$ for an absolute constant $c$, and the analogous two-sided inequality is denoted by $A \sim B$.

For a set $A$, let $\mathbb{1}_A$ be its indicator function and put $|A|$ to be its cardinality.

Finally, let us mention that we will abuse notation and write $\|z\|_{L_2}$ for the $L_2$ norm of the function $z$, without specifying the exact probability space on which the integration is performed. For example, $\|f - Y\|_{L_2^2} = \mathbb{E}(f(X) - Y)^2$, while $\|f - f^*\|_{L_2}^2 = \mathbb{E}(f - f^*)^2(X)$. We will denote the unit ball in $L_2$ by $D$ and the unit sphere by $S(L_2)$, again, without specifying the underlying space.

2 An almost isometric lower estimate

The main result of this section is an ‘almost isometric’ lower bound on

$$\inf_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} f^2(X_i).$$

The small-ball method, introduced in [9,17–19], can be used to show that if $\mathcal{F}$ satisfies the small-ball condition with constants $\kappa_0$ and $\varepsilon$ then for an appropriate choice of $r$, 

$$\inf_{\{f \in \mathcal{F}, \|f\|_{L_2} \geq r\}} \frac{1}{N} \sum_{i=1}^{N} \left( \frac{f(X_i)}{\|f\|_{L_2}} \right)^2 \geq c_0,$$

but $c_0 = c_0(\kappa_0, \varepsilon)$ is a constant that need not be close to 1. To obtain an almost isometric result rather than an isomorphic one, a slightly stronger assumption than the small-ball condition is required.

**Theorem 2.1** For every $2 < q \leq 4$ and $L \geq 1$ there exist constants $c_1$ and $c_2$ that depend only on $q$ and $L$ for which the following holds. Let $\mathcal{K} = \text{star}(\mathcal{F})$ and assume that for every $h \in \mathcal{K} - \mathcal{K}$, $\|h\|_{L_q} \leq L\|h\|_{L_2}$. Set $\gamma_1 = q/2(q - 1)$ and $\gamma_2 = (q - 2)/(2(q - 1))$, and for $0 < \zeta < 1$ let $r$ that satisfies

$$\mathbb{E}\|G\|_{(\mathcal{K} - \mathcal{K}) \cap rD} \leq \zeta \sqrt{Nr}, \quad \sup_{h \in (\mathcal{K} - \mathcal{K}) \cap rD} \frac{1}{\sqrt{N}} \left| \sum_{i=1}^{N} \varepsilon_i h(X_i) \right| \leq \zeta \sqrt{Nr}. \quad (2.1)$$

Then, with probability at least $1 - 2N \exp(-c_1 \zeta \gamma_1 N)$, if $h \in \mathcal{K} - \mathcal{K}$ and $\|h\|_{L_2} \geq r$, 

$$\frac{1}{N} \sum_{i=1}^{N} h^2(X_i) \geq \|h\|_{L_2}^2 \left( 1 - c_2 \zeta^{\gamma_2} \right). \quad (2.2)$$

It is highly likely that the exponents $\gamma_1$ and $\gamma_2$ are not optimal. For example, when $q = 4$ one would expect an estimate of $(1 - c_2 \zeta^{1/2})\|h\|_{L_2}^2$ rather than $(1 - c_2 \zeta^{1/3})\|h\|_{L_2}^2$ that follows from Theorem 2.1. Fortunately, this gap has little effect on the proof of
Theorem 1.10; as will be explained later, the value of $\zeta$ that is required in the proof is a small but fixed constant that depends only on $q$ and $L$.

The first step in the proof of Theorem 2.1 is an almost isometric lower estimate for a finite set of functions. In the general case, that set is selected to be an appropriate net that approximates $\mathcal{F}$, and the final step in the proof is an upper estimate on the empirical ‘approximation errors’.

2.1 An estimate for a single function

Given integers $N$ and $m$ and a function $f \in L^q$ for some $2 < q \leq 4$, set

$$\phi(f) = \begin{cases} f & \text{if } |f| \leq \left(\frac{N}{m}\right)^{1/q} \|f\|_{L^q} \\ \left(\frac{N}{m}\right)^{1/q} \text{sgn}(f) & \text{if } |f| > \left(\frac{N}{m}\right)^{1/q} \|f\|_{L^q}. \end{cases}$$

Hence, $\phi$ is a truncation of $f$ at a level that is selected according to the $L^q$ space to which $f$ belongs, the sample size $N$ and a parameter $m$ that will be used to calibrate the probability estimate.

Observe that pointwise, $|\phi(f)| \leq |f|$, and given a sample $(X_i)_{i=1}^N$, set

$$I_f = \{i : (\phi(f))(X_i) = f(X_i)\} = \left\{ i : |f(X_i)| \leq (N/m)^{1/q} \|f\|_{L^q} \right\}.$$

Theorem 2.2 There exist absolute constants $c_0, c_1, c_2$ and for $2 < q \leq 4$ there are constants $c_3, c_4$ that depend only on $q$ for which the following holds. If $N \geq c_0 m$, then with probability at least $1 - 2 \exp(-c_1 m)$,

1. $|I_f| \geq N - c_2 m$, and
2. For every $J \subset \{1, \ldots, N\}$ with $|J| \leq 4m$,

$$\left( \|f\|_{L^2}^2 - c_3 \|f\|_{L^q}^2 \left(\frac{m}{N}\right)^{1-(2/q)} \right) \leq \frac{1}{N} \sum_{i \in I_f \setminus J} f^2(X_i) \leq \left( \|f\|_{L^2}^2 + c_4 \|f\|_{L^q}^2 \left(\frac{m}{N}\right)^{1-(2/q)} \right).$$

Proof Observe that $Pr(|f| \geq (N/m)^{1/q} \|f\|_{L^q}) \leq m/N$. Using a standard binomial estimate applied to the event $\{|f| \geq (N/m)\|f\|_{L^q}\}$, it follows that for $0 < u < N/4m$,

$$Pr(|I_f| \geq u m) \leq \binom{N}{u m} \left(\frac{m}{N}\right)^{u m} \leq \left(\frac{e}{u}\right)^{u m},$$

and the first claim follows.

Springer
Next, consider $h = (\phi(f))^2$. Since $\|h\|_{L_\infty} \leq (N/m)^{2/q} \|f\|_{L_q}^2$ and $|\phi(f)| \leq |f|$, 

$$\|h\|_{L_2}^2 = \mathbb{E}(\phi(f))^q \cdot (\phi(f))^{4-q} \leq \|f\|_{L_q}^q \cdot \left(\frac{N}{m}\right)^{-1+4/q} \|f\|_{L_q}^{4-q}$$

$$= \left(\frac{N}{m}\right)^{-1+4/q} \cdot \|f\|_{L_q}^4.$$ 

Hence, by Bernstein’s inequality (see, e.g., [26]) for $h = (\phi(f))^2$, 

$$\left| \frac{1}{N} \sum_{i=1}^{N} (\phi(f))^2(X_i) - \mathbb{E}(\phi(f))^2 \right| \leq \left(\frac{m}{N}\right)^{1-2/q} \|f\|_{L_q}^2 \equiv u$$

with probability at least 

$$1 - 2 \exp\left(-c_2 N \min\left\{ \frac{u^2}{\|h\|_{L_2}^2}, \frac{u}{\|h\|_{L_\infty}} \right\}\right) = 1 - 2 \exp(-c_2 m).$$

Recall that if $i \in I_f^c$ then $|\phi(f)(X_i)| = (N/m)^{1/q} \|f\|_{L_q}$, and by the first part of the claim, $|I_f^c| \leq c_3 m$. Therefore, 

$$\frac{1}{N} \sum_{i \in I_f^c} (\phi(f))^2(X_i) \leq \frac{|I_f^c|}{N} \cdot \left(\frac{N}{m}\right)^{2/q} \|f\|_{L_q}^2 \leq c_4 \left(\frac{m}{N}\right)^{1-2/q} \|f\|_{L_q}^2.$$ 

Also, 

$$\mathbb{E}(\phi(f))^2 \geq \mathbb{E}f^2 1_{|f| \leq (N/m)^{1/q} \|f\|_{L_q}} = \mathbb{E}f^2 - \mathbb{E}f^2 1_{|f| > (N/m)^{1/q} \|f\|_{L_q}},$$

and

$$\mathbb{E}f^2 1_{|f| > (N/m)^{1/q} \|f\|_{L_q}} = \int_0^\infty 2t Pr\left(|f| 1_{|f| > (N/m)^{1/q} \|f\|_{L_q}} > t\right) dt$$

$$\leq \left(\frac{N}{m}\right)^{2/q} \|f\|_{L_q}^2 Pr(|f| > (N/m)^{1/q} \|f\|_{L_q})$$

$$+ \int_{(N/m)^{1/q} \|f\|_{L_q}}^\infty 2t Pr(|f| > t) dt$$

$$\leq \frac{q}{q-2} \left(\frac{m}{N}\right)^{1-2/q} \|f\|_{L_q}^2;$$

thus,

$$\mathbb{E}(\phi(f))^2 \geq \mathbb{E}f^2 - \frac{q}{q-2} \left(\frac{m}{N}\right)^{1-2/q} \|f\|_{L_q}^2.$$
Combining these observations, with probability at least $1 - 2\exp(-c_2m)$,

$$\frac{1}{N} \sum_{i \in I_f} f^2(X_i) = \frac{1}{N} \sum_{i \in I_f} (\phi(f))^2(X_i) = \frac{1}{N} \sum_{i=1}^{N} (\phi(f))^2(X_i) - \frac{1}{N} \sum_{i \in I'_f} (\phi(f))^2(X_i) \geq \mathbb{E}(\phi(f))^2 - c_4 \left( \frac{m}{N} \right)^{1-2/q} \|f\|_{L_q}^2$$

and,

$$\frac{1}{N} \sum_{i \in I_f} f^2(X_i) = \frac{1}{N} \sum_{i \in I_f} (\phi(f))^2(X_i) \leq \mathbb{E}f^2 - c(q) \left( \frac{m}{N} \right)^{1-2/q} \|f\|_{L_q}^2,$$

for a constant $c(q)$ that depends only on $q$.

Finally, note that if $i \in I_f$ then $|f(X_i)| \leq (N/m)^{1/q} \|f\|_{L_q}$. Hence, for every $J \subset I_f$,

$$\frac{1}{N} \sum_{j \in J} f^2(X_i) \leq \frac{|J|}{N} \left( \frac{N}{m} \right)^{2/q} \|f\|_{L_q}^2,$$

which completes the proof.

2.2 A uniform lower bound

Let $\mathcal{F} \subset L_2$ and $\eta > 0$, and set

$$\Phi_N(\mathcal{F}, \eta) = \mathbb{E} \sup_{h \in (\mathcal{F} - \mathcal{F}) \cap \eta D} \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i h(X_i) \right| + \eta.$$

When the underlying class $\mathcal{F}$ or the sample size $N$ are clear, we will abuse notation and write $\Phi(\eta)$ instead of $\Phi_N(\mathcal{F}, \eta)$.

Let $\mathcal{N}(\varepsilon, \mathcal{F}, L_2)$ be the minimal number of open $\varepsilon$-balls with respect to the $L_2$ norm that are needed to cover $\mathcal{F}$ and set

$$e_m(\mathcal{F}) = \inf\{ \varepsilon > 0 : \log \mathcal{N}(\varepsilon, \mathcal{F}, L_2) \leq m \}$$

to be the $m$-th entropy number of $\mathcal{F}$; the centres of the balls are called a minimal cover of $\mathcal{F}$.

Lemma 2.3 Let $\mathcal{F} \subset rS(L_2)$, set $2 < q \leq 4$ and assume that for every $f_1, f_2 \in \mathcal{F} \cup \{0\}$, $\|f_1 - f_2\|_{L_q} \leq L \|f_1 - f_2\|_{L_2}$. If $1 \leq m \leq N$ and $e_m(\mathcal{F}) \leq \eta$ then with probability at least $1 - 2N \exp(-c_1m)$,
\[
\inf_{\mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} f^2(X_i) \geq r^2 \left( 1 - c_2(q, L) \left( \left( \frac{N}{m} \right)^{1/2} \frac{\Phi(\eta)}{r} + \left( \frac{m}{N} \right)^{1-2/q} \right) \right),
\]

where \(c_1\) is an absolute constant and \(c_2\) is a constant that depend only on \(L\) and \(q\).

**Proof** Fix an integer \(m\) and let \(\mathcal{F}'\) be a minimal \(\eta\)-cover of \(\mathcal{F}\) with respect to the \(L_2\) norm. Since \(\eta \geq e_m(\mathcal{F})\) it follows that \(\log |\mathcal{F}'| \leq m\).

Let \(f \in \mathcal{F}\) and set \(\pi f \in \mathcal{F}'\) for which \(\|f - \pi f\|_{L_2} \leq \eta\). Put \(v_j = \pi f(X_j)\) and \(u_j = (f - \pi f)(X_j)\), and observe that if \(I \subset \{1, \ldots, N\}\) then

\[
\sum_{i=1}^{N} f^2(X_i) \geq \sum_{i \in I} f^2(X_i) = \sum_{i \in I} (\pi f(X_i) + (f - \pi f)(X_i))^2 \\
\geq \sum_{i \in I} v_i^2 - 2 \left( \sum_{i \in I} v_i^2 \right)^{1/2} \cdot \left( \sum_{i \in I} u_i^2 \right)^{1/2}.
\]

Let \(I_{\pi f} = \{i : \pi f(X_i) = (\phi(\pi f))(X_i)\}\) be as in Theorem 2.2; set \(J_f \subset \{1, \ldots, N\}\) to be the union of the set of the largest \(2m\) coordinates of \((|f - \pi f|(X_i))_{i=1}^{N} = (|u_i|)_{i=1}^{N}\) and the set of the largest \(2m\) coordinates of \((|\pi f(X_i)|)_{i=1}^{N} = (|v_i|)_{i=1}^{N}\). Applying Theorem 2.2, the union bound and the \(L_q\)-\(L_2\) norm equivalence, there is an absolute constant \(c_1\) and a constant \(c_2\) that depends only on \(q\) for which, with probability at least \(1 - 2 \exp(-c_1 m)\), for every \(\pi f \in \mathcal{F}'\),

\[
\left( 1 - c_2 L \left( \frac{m}{N} \right)^{1-2/q} \right) \leq \frac{1}{N} \sum_{i \in I_{\pi f} \setminus J_f} \left( \frac{\pi f(X_i)}{\|\pi f\|_{L_2}} \right)^2 \leq \left( 1 + c_2 L \left( \frac{m}{N} \right)^{1-2/q} \right).
\]

Next, one has to obtain a high probability estimate on the ‘coordinate distribution’ of the vector \((|u_i|)_{i=1}^{N}\). To that end, fix \(t > 0\) and observe that by symmetrization and contraction arguments (see, e.g., [15, 26]),

\[
t \mathbb{E} \sup_{f \in \mathcal{F}} |\{i : |f - \pi f|(X_i) \geq t\}| \leq \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{N} |f - \pi f|(X_i) \\
\leq 2 \mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{i=1}^{N} \varepsilon_i (f - \pi f)(X_i) \right| + N \sup_{f \in \mathcal{F}} \mathbb{E} |f - \pi f| \leq 2N \Phi(\eta). \quad (2.3)
\]

Fix \(t_j\) to be named later and apply Talagrand’s concentration inequality for bounded empirical processes [4, 14, 24] to the class of indicator functions \(\{1_{|f - \pi f| \geq t_j} : f \in \mathcal{F}\}\). Setting

\[
W_j = \sup_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} 1_{|f - \pi f| \geq t_j}(X_i),
\]
with probability at least $1 - 2 \exp(-m)$, for every $f \in \mathcal{F}$,

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{1}_{|f - \pi f| \geq t_j}(X_i) \leq c_3 \left( \mathbb{E}W_j + \sqrt{\frac{m}{N}} \sigma_j + \frac{m}{N} \right) = (\ast)_j,$$

where $\sigma_j = \sup_{f \in \mathcal{F}} \Pr_{1/2}(|f - \pi f| \geq t_j)$. By the $L_q$ and $L_2$ norm equivalence,

$$\sigma_j^2 \leq \sup_{f \in \mathcal{F}} \frac{\mathbb{E}|f - \pi f|^q}{t_j^q} \leq \left( \frac{L \eta}{t_j} \right)^q,$$

and by (2.3), $\mathbb{E}W_j \leq 2N \Phi(\eta)/t_j$. Therefore, if $j \geq 2m$ and $t_j = c_4(q, L) \Phi(\eta) N/j$,

$$(\ast)_j \leq c_3 \left( \frac{2\Phi(\eta)}{t_j} + \sqrt{\frac{m}{N}} \cdot \left( \frac{L \eta}{t_j} \right)^{q/2} + \frac{m}{N} \right) \leq \frac{j}{N},$$

because $\Phi(\eta) \geq \eta$.

Summing for $2m \leq j \leq N$, with probability at least $1 - 2N \exp(-m)$, for every $j \geq 2m$ and every $f \in \mathcal{F}$,

$$\left| \left\{ i : |(f - \pi f)(X_i)| \geq c_4 \Phi(\eta) \frac{N}{j} \right\} \right| \leq j.$$

And, on that event, for $j \geq 2m$,

$$u^*_j \leq c_4 \Phi(\eta) \frac{N}{j},$$

where $(u^*_i)_{i=1}^{N}$ denotes a non-increasing rearrangement of $(|u_i|)_{i=1}^{N}$.

Hence,

$$\left( \sum_{j > 2m} (u^*_j)^2 \right)^{1/2} \leq c_4 \Phi(\eta) N \left( \sum_{j > 2m} j^{-2} \right)^{1/2} \leq c_5(q, L) \Phi(\eta) N / \sqrt{m},$$

and setting $I = I_{\pi f \setminus J_f}$,

$$\left( \sum_{i \in I} v_i^2 \right)^{1/2} \cdot \left( \sum_{i \in I} u_i^2 \right)^{1/2} \leq c_6(q, L) N^{3/2} r \Phi(\eta) / \sqrt{m}.$$

Recalling the lower estimate on $\sum_{i \in I} v_i^2$, it follows that

$$\frac{1}{N} \sum_{i \in I} f^2(X_i) \geq r^2 \cdot \left( 1 - c(L, q) \left( \frac{m}{N} \right)^{1-(2/q)} + \frac{\Phi(\eta)}{r} \cdot \left( \frac{N}{m} \right)^{1/2} \right).$$

$\square$
Proof of Theorem 2.1} Let \( 0 < \zeta < 1 \) and set \( r \) for which (2.1) holds. Recall that \( \mathcal{K} = \text{star}(\mathcal{F}) \) and thus \( \mathcal{K} - \mathcal{K} \) is star-shaped around 0. Hence, it is standard to verify that if \( r' \geq r \) then

\[
\mathbb{E} \sup_{w \in (\mathcal{K} - \mathcal{K}) \cap r' D} \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i w(X_i) \right| \leq \zeta r'
\]

and

\[
\mathbb{E} \|G\|_{(\mathcal{K} - \mathcal{K}) \cap r' D} \leq \zeta \sqrt{Nr'}.
\]

Consider the class \( \mathcal{F}_r = \text{star}(\mathcal{F}) \cap r S(L_2) = \mathcal{K} \cap r S(L_2) \). Set \( \eta = c \mathbb{E} \|G\|_{\mathcal{F}_r} / \sqrt{m} \) and note that by Sudakov’s minoration (see, e.g., [15, 23]), \( \eta \geq e_m(\mathcal{F}_r) \), provided that \( c \) is a well-chosen absolute constant. Therefore,

\[
\Phi_N(\mathcal{F}_r, \eta) \leq \mathbb{E} \sup_{w \in (\mathcal{K} - \mathcal{K}) \cap 2r D} \left| \frac{1}{N} \sum_{i=1}^{N} \varepsilon_i w(X_i) \right| + \eta \leq 2 \zeta r + \frac{c \mathbb{E} \|G\|_{\mathcal{F}_r}}{\sqrt{m}}.
\]

Set \( m = \theta N \) for a constant \( 0 < \theta < 1 \) to be specified later; thus,

\[
\frac{\Phi_N(\mathcal{F}_r, \eta)}{r} \leq 2 \zeta + \frac{c}{\sqrt{\theta}} \cdot \frac{\mathbb{E} \|G\|_{\mathcal{F}_r}}{r \sqrt{N}}.
\]

Moreover, \( \mathcal{F}_r \subset (\mathcal{K} - \mathcal{K}) \cap r D \). By the choice of \( r \), \( \mathbb{E} \|G\|_{\mathcal{F}_r} \leq \mathbb{E} \|G\|_{(\mathcal{K} - \mathcal{K}) \cap r D} \leq \zeta r \sqrt{N} \), and

\[
\frac{\Phi_N(\mathcal{F}_r, \eta)}{r} \leq \frac{c_1 \zeta}{\sqrt{\theta}}.
\]

Thanks to Lemma 2.3, with probability at least \( 1 - 2N \exp(-c_2 \theta N) \), for every \( f \in \mathcal{F}_r \),

\[
\frac{1}{N} \sum_{i=1}^{N} f^2(X_i) \geq r^2 \left( 1 - c(q, L) \left( \left( \frac{m}{N} \right)^{-2/q} + \left( \frac{N}{m} \right)^{1/2} \frac{\Phi(\eta)}{r} \right) \right)
\]

\[
\geq r^2 \left( 1 - c(q, L) \left( \theta^{1-2/q} + \frac{c \zeta}{\theta} \right) \right) \quad (2.4)
\]

Setting \( \theta = \zeta^{q/2(q-1)} \), the claim follows for \( f \in \mathcal{F}_r = \text{star}(\mathcal{F}) \cap r S(L_2) \).

Finally, since (2.4) is positive homogeneous and \( \text{star}(\mathcal{F}) \) is star-shaped around 0, it also holds on the same event when \( f \in \text{star}(\mathcal{F}) \) and \( \|f\|_{L_2} > r \). □
3 The median of means as a crude measure of distances

As noted previously, the results of [9,17–19] show that a small-ball condition suffices for ensuring that with probability at least $1 - 2 \exp(-cN)$,

$$
\alpha^2 \| f - h \|^2_{L_2} \leq \frac{1}{N} \sum_{i=1}^{N} (f - h)^2(X_i)
$$

for well chosen constants $\alpha$ and $c$ that depend only on the small-ball condition in $\mathcal{F}$, and for every $f, h \in \mathcal{F}$ whose $L_2$-distance is not ‘too small’. However, if class members do not have well-behaved tails, the probability that

$$
\alpha^2 \| f - h \|^2_{L_2} \leq \frac{1}{N} \sum_{i=1}^{N} (f - h)^2(X_i) \leq \beta^2 \| f - h \|^2_{L_2}
$$

even for a single pair $f, h \in \mathcal{F}$ may be rather small; certainly not of the order of $1 - 2 \exp(-cN)$. Unfortunately, this means that the empirical mean is a poor two-sided estimator of distances, as it lacks stability: if $f - h$ is a heavy-tailed function, there will be at least one very large value of $|(f - h)(X_i)|$, and that will destroy any hope of having

$$
\frac{1}{N} \sum_{i=1}^{N} (f - h)^2(X_i) \leq \beta^2 \| f - h \|^2_{L_2}
$$

unless $\beta$ is very large.

To bypass this obstacle we will use the more stable median-of-means functional (see Definition 3.2, below).

Our starting point is that very little ‘mixing’ is needed for an empirical mean to satisfy a small-ball estimate with a rather high (constant) probability.

**Lemma 3.1** For every $q > 2$ and $L \geq 1$ there are constants $\ell$ and $\kappa_0$ that depend only on $q$ and $L$ for which the following holds. If $\| Z \|_{L_q} \leq L \| Z \|_{L_2}$ and $Z_1, \ldots, Z_\ell$ are independent copies of $Z$, then

$$
\Pr \left( \frac{1}{\ell} \sum_{i=1}^{\ell} |Z_i| \geq \kappa_0 \| Z \|_{L_2} \right) \geq \frac{3}{4}.
$$

**Proof** Since $\| Z \|_{L_q} \leq L \| Z \|_{L_2}$, it follows from a standard application of the Paley-Zygmund inequality (see, e.g., [6]) that $Z$ satisfies a small-ball condition with constants $c_1$ and $c_2$ that depend only of $q$ and $L$. Therefore,

$$
\| Z \|_{L_1} \geq c_1 \| Z \|_{L_2} \Pr(|Z| \geq c_1 \| Z \|_{L_2}) \geq c_1 c_2 \| Z \|_{L_2}.
$$

(3.1)
By an appropriate version of the Berry-Esseen inequality for independent copies of $Z \in L_q$ for $q > 2$ [21], if $\ell \geq c_3(q, L)$ then

$$\sup_{t \in \mathbb{R}} \left| Pr \left( \frac{1}{\ell} \sum_{i=1}^{\ell} |Z_i| \geq \mathbb{E}|Z| + \frac{t\||Z| - \mathbb{E}|Z|\|_{L_2}}{\sqrt{\ell}} \right) - Pr \left( g \geq t \right) \right| \leq 0.05.$$ 

Take $t < 0$ to be the largest for which $Pr \left( g \geq t \right) \geq 0.8$. Applying (3.1), if $\ell \geq 4t^2/(c_1c_2)^2$ then $\mathbb{E}|Z| \geq 2t\||Z|\|_{L_2}/\sqrt{\ell}$ and

$$\mathbb{E}|Z| + \frac{t\||Z| - \mathbb{E}|Z|\|_{L_2}}{\sqrt{\ell}} \geq (c_1c_2/2)\|Z\|_{L_2}.$$ 

Therefore, setting $\kappa_0 = c_1c_2/2$ (which depends only on $q$ and $L$),

$$Pr \left( \frac{1}{\ell} \sum_{i=1}^{N} |Z_i| \geq \kappa_0\|Z\|_{L_2} \right) \geq \frac{3}{4}. \quad \Box$$

What will assume the role of the empirical mean $\frac{1}{N} \sum_{i=1}^{N} (f - h)^2(X_i)$ as a way of estimating $L_2$ distances is the following median-of-means functional, which is more stable than the empirical mean when dealing with heavy-tailed functions:

**Definition 3.2** Let $1 \leq \ell \leq N$ and set $I_j = \{\ell j + 1, \ldots, \ell (j + 1)\} \subset \{1, \ldots, N\}$ for $0 \leq j \leq \lfloor N/\ell \rfloor \equiv M - 1$. For $v \in \mathbb{R}^N$ let $\text{Med}_\ell(v)$ to be the median of the vector of means $(\ell^{-1} \sum_{i \in I_j} v_i)_{j=0}^{M-1} \in \mathbb{R}^M$.

Thus, $I_0, \ldots, I_{M-1}$ are disjoint subsets of $\{1, \ldots, N\}$, each of cardinality $\ell$, and $\text{Med}_\ell(v)$ is the median of the means taken over the ‘blocks’ $I_j$.

Fix $2 < q \leq 4$ and $L \geq 1$, and set $\ell$ and $\kappa_0$ as in Lemma 3.1. Without loss of generality, assume that $N = \ell M$ for an integer $M$, and following Definition 3.2, given $v \in \mathbb{R}^N$, set $\text{Med}_\ell(v)$ to be the median of the vector of means performed in the $M$ blocks $I_0, \ldots, I_{M-1}$.

**Theorem 3.3** For every $2 < q \leq 4$ and $L \geq 1$ there exists constants $c_1, c_2, c_3$ and $\alpha < 1 < \beta$ that depend only on $q$ and $L$ for which the following holds. Let $\mathcal{F} \subset L_2$, put $\mathcal{K} = \text{star}(\mathcal{F})$ and assume that for every $w \in \mathcal{K}, \|w\|_{L_q} \leq L\|w\|_{L_2}$. Set $r > 0$ that satisfies

$$\mathbb{E}\| G \|_{(\mathcal{K} - \mathcal{K}) \cap \mathcal{D}} \leq c_1 \sqrt{N} r, \quad \mathbb{E} \sup_{h \in (\mathcal{K} - \mathcal{K}) \cap \mathcal{D}} \frac{1}{\sqrt{N}} \left| \sum_{i=1}^{N} e_i h(X_i) \right| \leq c_2 \sqrt{N} r.$$ 

Then, with probability at least $1 - 2 \exp(-c_3 N)$, for every $w \in \mathcal{K}$ for which $\|w\|_{L_2} \geq r$,

$$\alpha \|w\|_{L_2} \leq \text{Med}_\ell \left( \|w(X_i)\|_{1=1}^{N} \right) \leq \beta \|w\|_{L_2}.$$
Moreover, on the same event, if \( \|w\|_{L^2} \leq r \) then

\[
\text{Med}_\ell (|w(X_i)|)_{i=1}^N \leq \beta r.
\]

The proof of Theorem 3.3 is based on a similar argument to the one used in Theorem 4.3 from [19]. Both are based on an estimate on the distribution of the coordinates of the vectors \( (|w(X_i)|)_{i=1}^N \). First, one shows that the majority of the coordinates of each \( (|w(X_i)|)_{i=1}^N \) in an appropriate net are large, and then that the ‘oscillation terms’ cannot have ‘many’ large coordinates.

The first step in that direction is the following observation.

**Lemma 3.4** There are absolute constants \( c_1 \) and \( c_2 \) for which the following holds.

Consider \( Z \in L^2 \) that satisfies a small-ball condition with constants \( \kappa_0 \) and \( \varepsilon \). If \( Z_1, \ldots, Z_N \) are independent copies of \( Z \), then with probability at least \( 1 - 2\exp(-c_1\delta^2\varepsilon N) \) there is a subset \( I \subset \{1, \ldots, N\} \), \( |I| \geq (1 - \delta)\varepsilon N \), and for every \( i \in I \),

\[
\kappa_0 \|Z\|_{L^2} \leq |Z_i| \leq c_2 \|Z\|_{L^2}/\sqrt{\delta \varepsilon}.
\]

**Proof** Fix \( 0 < \delta < 1 \) and let \( A = \{ \kappa_0 \|Z\|_{L^2} \leq |Z| \leq 3\|Z\|_{L^2}/\sqrt{\delta \varepsilon} \} \). Combining the small-ball condition and Chebyshev’s inequality, \( \Pr(A) \geq (1 - \delta/9)\varepsilon \). Let \( \eta \) to be a selector (i.e., a \{0, 1\}-valued random variable) with mean \( (1 - \delta/9)\varepsilon \) and set \( \eta_1, \ldots, \eta_N \) to be independent copies of \( \eta \). A straightforward concentration argument shows that with probability at least \( 1 - 2\exp(-c_1\delta^2\varepsilon N) \),

\[
|i : \eta_i = 1| = \sum_{i=1}^N \eta_i \geq (1 - \delta)\varepsilon N,
\]

and the claim follows. \( \square \)

**Proof of Theorem 3.3** Let \( \ell \) and \( \kappa_0 \) be as in Lemma 3.1 and recall that the two constants depend only on \( q \) and \( L \). Assume, without loss of generality, that \( M = N/\ell \) is an integer, set \( \varepsilon = 3/4 \) and fix \( 0 < \delta < 1 \) for which \((1 - \delta)\varepsilon = 0.6\). Set \( \mathcal{K} = \text{star}(\mathcal{F}) \), let \( \xi_1 \) and \( \xi_2 \) to be named later and put \( r > 0 \) that satisfies

\[
E \|G\|_{\mathcal{K} \cap rD} \leq \xi_1 \sqrt{N}r \quad \text{and} \quad \mathbb{E} \sup_{h \in (\mathcal{K} - \mathcal{K}) \cap rD} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^N \xi_i h(X_i) \right| \leq \xi_2 \sqrt{N}r.
\]

Let \( v \in \mathcal{K} \) and set

\[
\mathcal{M}_v = \frac{1}{\ell} \sum_{i=1}^\ell |v|(X_i);
\]

\( \square \) Springer
thus \((M_v, j)_{j=0}^{M-1} = \left( \ell^{-1} \sum_{i \in I_j} |v|(X_i) \right)_{j=0}^{M-1}\) are \(M\) independent copies of the random variable \(M_v\), which, by Lemma 3.1 satisfies the small-ball condition with constants \(\kappa_0\) and \(\varepsilon = 3/4\).

One may verify that

\[
\frac{3}{4} \kappa_0 \|v\|_{L_2} \leq \|M_v\|_{L_2} \leq \|v\|_{L_2},
\]

and thus, by Lemma 3.4, with probability at least \(1 - 2 \exp(-c_1 \delta^2 \varepsilon M) = 1 - 2 \exp(-c_2 N)\), there is \(J \subset \{0, \ldots, M-1\}, |J| \geq (1 - \delta/2)\varepsilon M\), and for every \(j \in J\),

\[
\frac{3}{4} \kappa_0^2 \|v\|_{L_2} \leq M_v, j \leq c_3 \|v\|_{L_2} / \sqrt{\delta \varepsilon} = c_4 \|v\|_{L_2}.
\]

Thanks to the high-probability estimate, the same assertion holds uniformly for \(\exp(c_2 N/2)\) random variables of the form \(M_v\). And, in particular, for every \(v \in V_r \subset K \cap rS(L_2) \equiv K_r\), which is a maximal \(\eta\)-separated set for a choice of \(\eta\) large enough to ensure that \(|V_r| \leq \exp(c_2 N/2)\).

Therefore, with probability at least \(1 - 2 \exp(-c_2 N/2)\), for every \(v \in V_r\) there is a subset \(J_v \subset \{0, \ldots, M-1\}\) of cardinality at least \((1 - \delta/2)\varepsilon M\) and for every \(j \in J_v\),

\[
\frac{3}{4} \kappa_0^2 r \leq M_v, j \leq c_4 r.
\] (3.2)

By Sudakov’s inequality applied to the set \(K_r\) and using the choice of \(r\), one may select

\[
\eta = c_5 \frac{\|G\|_{K_r}}{\sqrt{c_2 N/2}} \leq c_6 \zeta_1 r.
\]

Next, consider the empirical oscillation term: for every \(f \in K_r\), let \(\pi f\) be the best approximation with respect to the \(L_2\) distance of \(f\) in \(V_r\). Set \(u_f = \mathbb{I}_{\{|f - \pi f| > 3\kappa_0^2 r / 8\}}\), consider the class of indicator functions \(U_r = \{u_f : f \in K_r\}\) and let

\[
\psi(X_1, \ldots, X_N) = \sup_{u_f \in U_r} \frac{1}{N} \sum_{i=1}^N u_f(X_i).
\]

By the bounded differences inequality (see, for example, [4]), with probability at least \(1 - \exp(-c_7 t^2)\),

\[
\psi(X_1, \ldots, X_N) \leq \mathbb{E} \psi + \frac{t}{\sqrt{N}}.
\]

To estimate \(\mathbb{E} \psi\) from above, set \(\phi(t) = t / (3\kappa_0^2 r / 8)\). Observe that for every \(u_f \in U_r\), \(u_f(X) \leq \phi(|f - \pi f|(X))\), and that by the Giné-Zinn symmetrization theorem [7,15] and the choice of \(r\),

\[\text{ Springer}\]
Thus, on at least \((1 - \delta)\varepsilon M = 0.6M\) of the coordinate blocks \(I_j\) contain a sample point \(X_i\) for which \(|f - \pi f|(X_i) \geq (3\kappa_0^2/8)r\); on the remaining \((1 - \delta)\varepsilon M = 0.6M\) blocks,

\[
\frac{1}{\ell} \sum_{i \in I_j} |f - \pi f|(X_i) \leq (3\kappa_0^2/8)r.
\]

Hence, with probability at least \(1 - 2 \exp(-c_9(q, L)N)\), for every \(f \in \text{star}(\mathcal{F}) \cap rS(L_2)\) there is a subset of \([0, \ldots, M - 1]\) of cardinality at least \((1 - \delta)\varepsilon M = 0.6M\), on which

\[
|M_{f, j}| \geq |M_{\pi f, j}| - |M_{f - \pi f, j}| \geq (3\kappa_0^2/4)r - (3\kappa_0^2/8)r = (3\kappa_0^2/8)\|f\|_{L_2}, \quad \text{and}
\]

\[
|M_{f, j}| \leq |M_{\pi f, j}| + |M_{f - \pi f, j}| \leq (c_4 + 3\kappa_0^2/8)\|f\|_{L_2}. \tag{3.3}
\]

Moreover, since the estimates are positive homogeneous and \(\text{star}(\mathcal{F})\) is star-shaped around 0, (3.3) is true on the same event when \(f \in \text{star}(\mathcal{F})\) and \(\|f\|_{L_2} \geq r\). The claim follows by recalling that \(\ell\) and \(\kappa_0\) depend only on \(q\) and \(L\), and selecting \(0 < \alpha < 3\kappa_0/8\) and \(\beta \geq c_4 + 3\kappa_0^2/8\).

The proof of the second part is almost identical: \(\mathcal{V}_r\) is defined exactly as above, and for every \(f \in \text{star}(\mathcal{F}) \cap rD\), \(\pi f\) is the best approximation in \(\mathcal{V}_r\); thus, \(\|f - \pi f\|_{L_2} \leq 2r\). Just as in the proof of the first part, with probability at least \(1 - 2 \exp(-c_9(q, L)N)\), for every \(f \in \text{star}(\mathcal{F}) \cap rD\),

\[
|\{i : |f - \pi f|(X_i) \geq 3r\}| \leq \delta \varepsilon M/2.
\]

Thus, on at least \((1 - \delta)\varepsilon M = 0.6M\) of the coordinate blocks,

\[
|M_{\pi f, j}| + |M_{f - \pi f, j}| \leq c_{10}(q, L)r,
\]

and one may choose \(\beta = \max\{c_4 + 3\kappa_0^2/8, c_{10}\}\). \hfill \Box
4 The aggregation procedure

After developing the necessary machinery that is needed for the proof of Theorem 1.10, let us turn to the definition of the aggregation procedure.

The aggregation procedure follows the general path of [11]—though with many essential modifications. The core difference between the method of proof used in [11] and the one we employ here is unavoidable, as the former is based on a two-sided concentration estimate on empirical means which is simply false for heavy-tailed functions. Most notably, two-sided empirical estimates on $L_2$ distances play a central role in [11] and the alternatives to those concentration-based bounds are the empirical ‘isomorphic’ upper estimate and the almost isometric lower bound established above.

The aggregation procedure consists of two stages. Given a base class $\mathcal{F}$, one must first identify a subset $V \subset \mathcal{F}$ that is selected in a data-dependent way. The set $V$ will consists of well-behaved functions in a sense that will be clarified below, and will be shown to contain $f^*$. Then, in the second stage, one applies ERM to the set of midpoints of pairs of elements in $V$ (a set that contains $V$ as well), i.e., to

$$W = \left\{ \frac{v_1 + v_2}{2} : v_1, v_2 \in V \right\} \subset U,$$

using a second, independent sample.

The first step, i.e., the selection of $V$, is essential to the success of the procedure. Tempting as it is, one cannot just take the entire class, add to it all the midpoints and apply ERM to the resulting set. For example, if $\mathcal{F} = \{-f, 3f\}$ then by adding the mid-point $f$, one reverts to the two-point problem mentioned in the introduction.

We will show that if the diameter of the random set $V$ is small enough, then ERM performed in $V$ will result in the choice of a function with a small error. On the other hand, if $V$ has a larger diameter, a mid-point of two ‘almost diametric’ points improves the approximation error considerably, and ERM is likely to select such a function. Since one compares the performance of the function chosen to that of the best in $\mathcal{F}$, the improved approximation error is the dominant factor in the bound, thus leading to the desired result.

**Lemma 4.1** Let $C \geq 1$, $r > 0$ and $0 < \theta \leq 1/32$, and consider $V \subset \mathcal{F}$ that satisfies the following:

- $f^* \in V$ (where, as always, $f^* = \arg \min_{f \in \mathcal{F}} \|f - Y\|_{L_2}$).
- For every $v \in V$, $\|v - Y\|_{L_2}^2 \leq \|f^* - Y\|_{L_2}^2 + \max\{Cr^2, \theta \text{diam}^2(V, L_2)\}$.

Let $W = \{(v_1 + v_2)/2 : v_1, v_2 \in V\}$, put $w^* = \arg \min_{w \in W} \|w - Y\|_{L_2}$ and set $\hat{w}$ to be the empirical minimizer in $W$. If $(X_i, Y_i)_{i=1}^N$ is a sample for which, for every $w \in W$,

$$\|w - Y\|_{L_2}^2 - \|w^* - Y\|_{L_2}^2 \leq \frac{1}{N} \sum_{i=1}^N \left( (w(X_i) - Y_i)^2 - (w^*(X_i) - Y_i)^2 \right) + \max\{Cr^2, \theta \|w - w^*\|_{L_2}^2\},$$

(4.1)
then
\[
\mathbb{E}\left((\hat{w} - Y)^2|(X_i, Y_i)_{i=1}^N\right) \leq \mathbb{E}(f^*(X) - Y)^2 + 2Cr^2.
\]

Lemma 4.1 implies that if \( V \) consists of functions whose excess risk (relative to \( \mathcal{F} \)) is either small (\( \leq Cr^2 \)), or, alternatively, at least smaller than a fixed proportion of the square of the diameter of \( V \), and if one is given a sample for which the oracle type inequality (4.1) holds, then ERM performed in \( W \) using that sample selects a function whose excess risk is at most \( 2Cr^2 \).

Naturally, at this point Lemma 4.1 is somewhat speculative, as it contains two substantial ‘if’s’. For the lemma to be of any use, one has to construct \( V \) using a random sample and without knowing the identity of \( f^* \), and then to establish the oracle type inequality in \( W \) using a second, independent sample.

**Proof** Set \( d_V = \text{diam}(V, L_2) \), note that \( \text{diam}(W, L_2) = d_V \) and that
\[
\frac{1}{N} \sum_{i=1}^{N} (\hat{w}(X_i) - Y_i)^2 - \frac{1}{N} \sum_{i=1}^{N} (w^*(X_i) - Y_i)^2 \leq 0. \tag{4.2}
\]

Consider two cases: firstly, if \( d_V \leq \sqrt{Cr} \) then by (4.1)
\[
\mathbb{E}\left((\hat{w}(X) - Y)^2|(X_i, Y_i)_{i=1}^N\right) \leq \mathbb{E}(w^*(X) - Y)^2 + Cr^2 \leq \mathbb{E}(f^*(X) - Y)^2 + Cr^2
\]
because \( f^* \in V \subset W \).

Secondly, assume that \( d_V \geq \sqrt{Cr} \). Since \( f^* \in V \), there is some \( v \in V \) for which \( \|v - f^*\|_{L_2} \geq d_V/2 \). Set \( w = (v + f^*)/2 \in W \) and observe that by the uniform convexity of the \( L_2 \) norm and the definition of \( V \),
\[
\|w^* - Y\|_{L_2}^2 \leq \|w - Y\|_{L_2}^2 = \frac{1}{2}\|v - Y\|_{L_2}^2 + \frac{1}{2}\|f^* - Y\|_{L_2}^2 - \frac{1}{4}\|v - f^*\|_{L_2}^2
\]
\[
\leq \|f^* - Y\|_{L_2}^2 + \max\left\{Cr^2, \theta d_V^2\right\} - \frac{d_V^2}{16}
\]
\[
\leq \|f^* - Y\|_{L_2}^2 + Cr^2 - \left(\frac{1}{16} - \theta\right)d_V^2.
\]

Combining this with (4.1) applied to \( \hat{w} \), and with (4.2), and recalling that \( \theta \leq 1/32 \),
\[
\mathbb{E}\left((\hat{w}(X) - Y)^2|(X_i, Y_i)_{i=1}^N\right)
\]
\[
\leq \mathbb{E}(f^*(X) - Y)^2 + Cr^2 + \theta d_V^2 + \left(\mathbb{E}(w^*(X) - Y)^2 - \mathbb{E}(f^*(X) - Y)^2\right)
\]
\[
\leq \mathbb{E}(f^*(X) - Y)^2 + 2Cr^2 - \left(\frac{1}{16} - 2\theta\right)d_V^2
\]
\[
\leq \mathbb{E}(f^*(X) - Y)^2 + 2Cr^2.
\]

\( \square \)
Next, we shall identify sufficient conditions that allow one to construct the set $V$ as in Lemma 4.1 in a data-dependent way.

### 4.1 The construction of $V$

Given a class $\mathcal{F}$, recall that $U = \{(f_1 + f_2)/2 : f_1, f_2 \in \mathcal{F}\}$; that for $1 \leq \ell \leq N$ and $0 \leq j \leq \lfloor N/\ell \rfloor \equiv M - 1$, $I_j = \{\ell j + 1, \ldots, \ell(j + 1)\} \subset \{1, \ldots, N\}$; and that for $v \in \mathbb{R}^N$, $\text{Med}_\ell(v)$ is the median of the vector of means $(\ell^{-1} \sum_{i \in I_j} v_i)_{j=0}^{M-1} \in \mathbb{R}^M$.

**Definition 4.2** Fix $r_U > 0$, $u_0 \in U$, $1 \leq \ell \leq N$, $0 < \alpha \leq 1 \leq \beta$, and set $\rho = (\alpha/20\beta)^2 \leq 1/400$.

Let $\mathcal{A}_{u_0}$ be the set of $N$-samples $(X_i, Y_i)_{i=1}^N$ for which the following holds:

- For every $u \in U$
  
  \[
  \left| \frac{1}{N} \sum_{i=1}^{N} (u_0(X_i) - Y_i)(u - u_0)(X_i) - \mathbb{E}(u_0(X) - Y)(u - u_0)(X) \right| 
  \leq \rho \max \left\{ r_U^2, \|u - u_0\|^2_{L^2} \right\};
  \]

- If $u_1, u_2 \in U$ and $\|u_1 - u_2\|_{L^2} \geq r_U$, then
  
  \[
  \left( \frac{1}{N} \sum_{i=1}^{N} (u_1 - u_2)^2(X_i) \right) \geq (1 - \rho)\|u_1 - u_2\|^2_{L^2};
  \]

- If $u_1, u_2 \in U$ and $\|u_1 - u_2\|_{L^2} \geq r_U$, then
  
  \[
  \alpha\|u_1 - u_2\|_{L^2} \leq \text{Med}_\ell(|u_1 - u_2|(X_i))_{i=1}^N \leq \beta\|u_1 - u_2\|_{L^2},
  \]

  and if $\|u_1 - u_2\|_{L^2} < r_U$ then
  
  \[
  \text{Med}_\ell(|u_1 - u_2|(X_i))_{i=1}^N \leq \beta r_U.
  \]

At this point, the right choice of $r_U$, $\ell$, $\alpha$ and $\beta$ is not clear, nor that $\mathcal{A}_{u_0}$ is nonempty, for that matter.

At last, we are ready to define the aggregation procedure:

**Definition 4.3** Let $(X_i, Y_i)_{i=1}^{2N}$ be a $2N$-sample and set $\mathcal{D}_1 = (X_i, Y_i)_{i=1}^N$ and $\mathcal{D}_2 = (X_i, Y_i)_{i=N+1}^{2N}$. Recall that

\[
\hat{f} = \arg\min_{f \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} (f(X_i) - Y_i)^2,
\]
and let

\[ V(D_1) = \left\{ f \in \mathcal{F} : \frac{1}{N} \sum_{i=1}^{N} (f(X_i) - Y_i)^2 \leq \frac{1}{N} \sum_{i=1}^{N} (\hat{f}(X_i) - Y_i)^2 + 3 \max \left\{ r_U^2, \rho \alpha^{-2} \text{Med}^2 \left( |f - \hat{f}(X_i)| i \right) \right\} \right\}. \]  

(4.3)

Set

\[ W(D_1) = \left\{ \frac{v_1 + v_2}{2} : v_1, v_2 \in V(D_1) \right\} \]

and define the aggregation procedure by

\[ \tilde{w} = \arg\min_{w \in W(D_1)} \frac{1}{N} \sum_{i=N+1}^{2N} (w(X_i) - Y_i)^2. \]  

(4.4)

**Theorem 4.4** Let \( w^* = \arg\min_{w \in V(D_1)} \mathbb{E}(w(X) - Y)^2 \). If \( D_1 \in \mathcal{A}_{f^*} \) and \( D_2 \in \mathcal{A}_{w^*} \), then \( V(D_1) \) and \( W(D_1) \) satisfy the conditions of Lemma 4.1 for \( \theta < 1/32 \), \( C = 6 \) and \( r = r_U \). In particular, for such a 2N-sample, and if \( \tilde{w} \) is the empirical minimizer selected in \( W(D_1) \) using the sample \( D_2 \), one has

\[ \mathbb{E}\left((\tilde{w}(X) - Y)^2\right) \leq \mathbb{E}\left((f^*(X) - Y)^2\right) + 6r_U^2. \]

The rest of this section is devoted to the proof of Theorem 4.4.

**Remark 4.5** Note that if \( \mathcal{F} \) is finite, so is \( W(D_1) \)—which may be much smaller than \( \mathcal{F} \). Thus, if the original dictionary is finite then unlike the empirical star algorithm, the second step in the aggregation procedure is carried out in a finite set.

### 4.2 Proof of Theorem 4.4

Given a set \( \mathcal{K} \subset U \), put \( h^* = \arg\min_{h \in \mathcal{K}} \mathbb{E}(h(X) - Y)^2 \) and set

\[ \mathcal{L}_h^K(X, Y) = (h(X) - Y)^2 - (h^*(X) - Y)^2 \]

to be the excess square loss functional associated with \( \mathcal{K} \).

**Lemma 4.6** If \( (X_i, Y_i)_{i=1}^{N} \in \mathcal{A}_{h^*} \) then for every \( h \in \mathcal{K} \),

\[ \mathbb{E}\mathcal{L}_h^K \leq P_N \mathcal{L}_h^K + 3 \max\{r_U^2, \rho \|h - h^*\|_{L_2}^2\}. \]  

(4.5)

\( \copyright \) Springer
Proof Set \( \xi(X, Y) = h^*(X) - Y \), let \((X_i, Y_i)_{i=1}^N \in \mathcal{A}_{h^*}\) and put \( \xi_i = h^*(X_i) - Y_i \). Note that for every \( h \in \mathcal{K} \),

\[
P_N \mathcal{L}_h^K = \frac{2}{N} \sum_{i=1}^N \xi_i(h - h^*)(X_i) + \frac{1}{N} \sum_{i=1}^N (h - h^*)^2(X_i)
\]

\[
\geq \mathbb{E} \mathcal{L}_h^K - 2 \left| \frac{1}{N} \sum_{i=1}^N \xi_i(h - h^*)(X_i) - \mathbb{E} \xi(h - h^*)(X) \right|
\]

\[
+ \frac{1}{N} \sum_{i=1}^N (h - h^*)^2(X_i) - \|h - h^*\|_{L_2}^2.
\]

Recall that if \((X_i, Y_i)_{i=1}^N \in \mathcal{A}_{h^*}\) and \( \|h - h^*\|_{L_2} \geq r_U \), one has

\[
\frac{1}{N} \sum_{i=1}^N (h - h^*)^2(X_i) \geq (1 - \rho)\|h - h^*\|_{L_2}^2,
\]

and

\[
\left| \frac{1}{N} \sum_{i=1}^N \xi_i(h - h^*)(X_i) - \mathbb{E} \xi(h - h^*)(X) \right| \leq \rho\|h - h^*\|_{L_2}^2;
\]

thus

\[
P_N \mathcal{L}_h^K \geq \mathbb{E} \mathcal{L}_h^K - 3\rho\|h - h^*\|_{L_2}^2.
\]

Otherwise, if \( \|h - h^*\|_{L_2} \leq r_U \),

\[
P_N \mathcal{L}_h^K \geq \mathbb{E} \mathcal{L}_h^K - 2 \left| \frac{1}{N} \sum_{i=1}^N \xi_i(h - h^*)(X_i) - \mathbb{E} \xi(h - h^*)(X) \right| - \|h - h^*\|_{L_2}^2
\]

\[
\geq \mathbb{E} \mathcal{L}_h^K - 3r_U^2,
\]

as claimed. \( \square \)

Lemma 4.7 For a sample \( \mathcal{D} = (X_i, Y_i)_{i=1}^N \), let \( d = \text{diam}(V(\mathcal{D}), L_2) \) and recall that \( f^* = \text{argmin}_{f \in \mathcal{F}} \mathbb{E}(f(X) - Y)^2 \). If \( \mathcal{D} \in \mathcal{A}_{f^*} \) then

1. \( f^* \in V(\mathcal{D}) \), and
2. For every \( v \in V(\mathcal{D}) \),

\[\|v - Y\|_{L_2}^2 \leq \|f^* - Y\|_{L_2}^2 + 6 \max \{r_U^2, d^2/400\}\]

Proof Fix \( \mathcal{D} = (X_i, Y_i)_{i=1}^N \in \mathcal{A}_{f^*} \). Recall that if \( f_1, f_2 \in \mathcal{F} \subset U \) and \( \|f_1 - f_2\|_{L_2} \geq r_U \), one has

\[
\alpha \|f_1 - f_2\|_{L_2} \leq \text{Med}_\ell(\|f_1 - f_2\|(X_i))_{i=1}^N \leq \beta \|f_1 - f_2\|_{L_2}.
\]
In addition, applying Lemma 4.6 for $\mathcal{K} = \mathcal{F}$, it follows that for every $f \in \mathcal{F}$,

$$0 \leq \mathbb{E} \mathcal{L}_f^\mathcal{F} \leq P_N \mathcal{L}_f^\mathcal{F} + 3 \max\{r_U^2, \rho \|f - f^*\|_{L_2}^2\}. \quad (4.6)$$

Let $\hat{f}$ be the empirical minimizer in $\mathcal{F}$ and consider the following two cases: if $\|\hat{f} - f^*\|_{L_2} \geq r_U$ then $\text{Med}_\ell(|\hat{f} - f^*|)_{i=1}^N \geq \alpha \|\hat{f} - f^*\|_{L_2}$; alternatively, $\|\hat{f} - f^*\|_{L_2} \leq r_U$. Therefore, by (4.6)

$$P_N \mathcal{L}_{\hat{f}}^\mathcal{F} = P_N (\hat{f} - Y)^2 - P_N (f^* - Y)^2 \geq -3 \max\{r_U^2, \rho \|\hat{f} - f^*\|_{L_2}^2\}$$

$$\geq -3 \max\left\{r_U^2, \rho \alpha^{-2} \text{Med}_\ell \left(\|f - f^*\|_{L_2}\right)_{i=1}^N\right\},$$

implying that $f^* \in V(\mathcal{D})$.

Turing to the second part, note that $\hat{f} \in V(\mathcal{D})$, $P_N \mathcal{L}_{\hat{f}}^\mathcal{F} \leq 0$ and that for every $u \in U$,

$$\text{Med}_\ell(|\hat{f} - u|)_{i=1}^N \leq \beta \max\{r_U, \|\hat{f} - u\|_{L_2}\} \leq \beta \max\{r_U, d\}. $$

Hence, it follows from the definition of $V(\mathcal{D})$ that for every $v \in V(\mathcal{D})$,

$$P_N \mathcal{L}_v^\mathcal{F} \leq P_N \mathcal{L}_{\hat{f}}^\mathcal{F} + 3 \max\left\{r_U^2, \rho \alpha^{-2} \text{Med}_\ell \left(\|f - v\|_{L_2}\right)_{i=1}^N\right\}$$

$$\leq 3 \max\left\{r_U^2, \rho \left(\frac{\beta}{\alpha}\right)^2 d^2\right\}. \quad (4.7)$$

Combining (4.7) with (4.6), for every $v \in V(\mathcal{D})$,\n
$$\|v - Y\|_{L_2}^2 - \|f^* - Y\|_{L_2}^2 = \mathbb{E} \mathcal{L}_v^\mathcal{F} \leq 6 \max\left\{r_U^2, d^2/400\right\},$$

by the choice of $\rho$. \hfill \Box

**Proof of Theorem 4.4.** One has to show that the assumptions of Lemma 4.1 hold for $V(\mathcal{D}_1)$ and for $W(\mathcal{D}_1)$ for the sample $\mathcal{D}_2$. By Lemma 4.7, $f^* \in V(\mathcal{D}_1)$ and thus, for every $v \in V(\mathcal{D}_1)$,

$$\|v - Y\|_{L_2}^2 - \|f^* - Y\|_{L_2}^2 \leq \max\left\{6r_U^2, d^2/50\right\}.$$

Also, applying Lemma 4.6 for $W(\mathcal{D}_1) \equiv W \subset U$, it follows that if $\mathcal{D}_2 \in A_{w^*}$ then for every $w \in W$,

$$\mathbb{E} \mathcal{L}_w^W \leq P_N \mathcal{L}_w^W + 3 \max\{r_U^2, \rho \|w - w^*\|_{L_2}^2\}$$

$$\leq P_N \mathcal{L}_w^W + \max\{3r_U^2, \|w - w^*\|_{L_2}^2/50\}.$$
where $P_N$ is the empirical mean relative to $D_2$. Thus, the assumptions of Lemma 4.1 are verified, completing the proof of Theorem 4.4.

\[ \square \]

5 Proof of Theorem 1.10

As noted earlier, Theorem 1.10 follows from Theorem 4.4, once one shows that $r_U^2$ can be selected to be $r_{\text{opt}}$ for the right choice of constants, and that the probability of the events $A_{u_0}$ is sufficiently high. The choice of $u_0 \in U$ one is interested in is $u_0 = f^*$ and $u_0 = w^*$. In particular, for every target $Y$, $\|u_0 - Y\|_{L_2} \leq \|f^* - Y\|_{L_2}$, and thus it suffices to consider $u_0 \in U_Y$ as in (1.6).

Observe that the second the third conditions in the definition of $A_{u_0}$ are independent of $u_0$, and we shall begin by verifying those.

Given the base class $\mathcal{F}$, recall that $U = \{(f_1 + f_2)/2 : f_1, f_2 \in \mathcal{F}\}$ and that $H = \text{star}(U - U)$. Thus, for every $h \in H$, $\|h\|_{L_q} \leq L \|h\|_{L_2}$. Let $r_0$ be the infimum of the set of all $r > 0$ for which

\[
\mathbb{E}\|G\|_{H \cap rD} \leq \zeta_1(q, L)\sqrt{Nr} \quad \text{and} \quad \mathbb{E}\sup_{h \in (H - H) \cap rD} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i h(X_i) \right| \leq \zeta_2(q, L)\sqrt{Nr}
\]

(5.1)

for constants $\zeta_1$ and $\zeta_2$ as in the proof of Theorem 3.3.

Since $H$ and $H - H$ are star-shaped around 0, (5.1) holds for every $r \geq r_0$. Invoking Theorem 3.3 for the class $U - U$ and $r = 2r_0$, there are constants $\alpha \leq 1 \leq \beta$ and $\ell$ that depend only on $q$ and $L$ for which, with probability at least $1 - 2 \exp(-c_0(q, L)N)$, for every $h \in H$,

- If $\|h\|_{L_2} \geq 2r_0$ then $\alpha \|h\|_{L_2} \leq \text{Med}_\ell(|h(X_i)|)^N_{i=1} \leq \beta \|h\|_{L_2}$;
- If $\|h\|_{L_2} \leq 2r_0$ then $\text{Med}_\ell(|h(X_i)|)^N_{i=1} \leq \beta \cdot 2r_0$.

In particular, for any $r_U \geq 2r_0$, the third condition in the definition of $A_{u_0}$ is verified.

Next, let $\alpha$ and $\beta$ be as above and set $\rho = (\alpha/20\beta)^2$. Consider Theorem 2.1 for the class $U - U$ (and in which case, $\mathcal{K} = \text{star}(U - U) = H$). Recall that $\gamma_1 = q/2(q - 1)$ and $\gamma_2 = (q - 2)/2(q - 1)$, set $\zeta_3$ by $\rho \sim \zeta_3^{\gamma_2}$, and in particular, $\zeta_3$ depends only on $q$ and $L$. Consider $r_1$ for which

\[
\mathbb{E}\|G\|_{(H - H) \cap r_1D} \leq \zeta_3\sqrt{Nr_1}, \quad \text{and} \quad \mathbb{E}\sup_{h \in (H - H) \cap r_1D} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i h(X_i) \right| \leq \zeta_3\sqrt{Nr_1}.
\]

It follows that with probability at least

\[
1 - 2N \exp(-c_1\zeta_3^{\gamma_1}N) \geq 1 - 2 \exp(-c_2(q, L)\rho^{\gamma_1/\gamma_2}N),
\]
if \( h \in H - H \) and satisfies \( \|h\|_{L^2} \geq r_1 \) then

\[
\frac{1}{N} \sum_{i=1}^{N} h^2(X_i) \geq (1 - \rho)\|h\|_{L^2}^2.
\]

Moreover, since \( 0 \in H \), the same is true for every difference \( h = u_1 - u_2 \in U - U \subset H - H \) provided that \( \|u_1 - u_2\|_{L^2} \geq r_1 \). Thus, the second part in the definition of \( A_{u_0} \) holds for \( r_U \geq r_1 \).

Turning to the first part of the definition of \( A_{u_0} \) (which does depend on \( u_0 \)), one may apply Lemma 1.7 to the set \( U \) and for \( r_2 \geq r_M(\mathcal{F}, \rho/4, \delta/2, u_0) \).

Setting \( \xi = u_0(X) - Y \) and \( \xi_i = u_0(X_i) - Y_i \), it follows that with probability at least \( 1 - \delta \), for every \( u \in U \),

\[
\left| \frac{1}{N} \sum_{i=1}^{N} \xi_i (u - u_0)(X_i) - \mathbb{E}\xi (u - u_0) \right| \leq \rho \max \left\{ \|u - u_0\|_{L^2}^2, r_2^2 \right\},
\]

as required.

Finally, one may combine all the above conditions, by noting that for \( c_3 = \rho/4 \), \( c_4 = \min\{\zeta_2, \zeta_3\} \) and \( c_5 = \min\{\zeta_1, \zeta_3\} \), the choice of \( r^2 = r_{opt}(\mathcal{F}, \delta, c_3, c_4, c_5) \) is a valid choice in all of the above. Hence, for every \( u_0 \in U \),

\[
\Pr(A_{u_0}) \geq 1 - \delta - 2\exp(-c_6(q, L)N),
\]

and Theorem 1.10 follows from Theorem 4.4.

\[\square\]

5.1 Proof of Corollary 1.12

Let \( \mathcal{F} \) be a finite dictionary. While a learning procedure can only guarantee an error rate of the order of \( \sqrt{N^{-1} \log M} \), one may show that the aggregation procedure suggested above leads to a much better estimate.

Let us begin by reformulating Corollary 1.12:

**Theorem 5.1** For every \( L \geq 1 \) and \( q > 2 \) there exists a constant \( c_1 \) that depends only on \( L \) and \( q \) for which the following holds. Let \( \mathcal{F} = \{f_1, \ldots, f_M\} \) and assume that for \( w \in \text{span}(\mathcal{F}) \) and every \( p \geq 2 \), \( \|w\|_{L^p} \leq L \sqrt{p} \|w\|_{L^2} \). Assume further that \( Y \in L_q \) for some \( q > 2 \). Then for every \( 0 < \delta < 1 \), with probability at least \( 1 - \delta \),

\[
\mathbb{E}\left( (\tilde{f}(X) - Y)^2 | (X_i, Y_i)_{i=1}^{N} \right) \leq \mathbb{E}(f^*(X) - Y)^2 + c_1 \delta^{-2/q} \log(2/\delta) \|f^* - Y\|_{L_q}^2 \frac{\log M}{N}.
\]

As all the assumptions of Theorem 1.10 are satisfied here, what is left is to identify \( r_{opt} \). To that end, note that \( |U - U| \leq M^4 \). Thus, for every \( r > 0 \) there is \( k_r \leq M^8 \) and functions \( (w_{i,r})_{i=1}^{k_r} \) that satisfy \( \|w_{i,r}\|_{L^2} \leq r \), and

\[\copyright\ Springer\]
star\((H - H) \cap rD \subset \{ \lambda w_{i,r} : 1 \leq i \leq k_r, 0 \leq \lambda \leq 1 \}\) \equiv W_r.

By the moment equivalence in \(\text{span}(\mathcal{F})\), a straightforward chaining argument and the Majorizing Measures Theorem (see, e.g., [16] for similar arguments) it follows that

\[
E \sup_{w \in W_r} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i w(X_i) \right| \leq c_1 L \mathbb{E} \| G \|_{w_r}.
\]

And, it is standard to verify that

\[
E \| G \|_{w_r} = E \sup_{1 \leq i \leq kr} G_{w_i} \leq c_2 r \sqrt{\log M}.
\]

Therefore, if \(N \geq c_3 (L, \zeta) \log M\), then

\[
r_{Q,1}(\mathcal{F}, \zeta) = r_{Q,2}(\mathcal{F}, \zeta) = 0.
\]

Turning our attention to \(r_M\), one may invoke the following fact from [20]:

**Theorem 5.2** Let \(\xi \in L_q\) for some \(q > 2\) and assume that for every \(f, h \in \mathcal{F} \cup \{0\}\) and every \(p \geq 2\), \(\| f - h \|_{L_p} \leq L \sqrt{p} \| f - h \|_{L_2}\). Then, for every \(u, w > 1\), with probability at least

\[
1 - c_0(q) w^{-q} \frac{\log^q N}{N^{q/2 - 1}} - 2 \exp \left( -c_1(L) u^2 \left( \frac{\mathbb{E} \| G \|_{\mathcal{F}}}{\sup_{f \in \mathcal{F}} \| f \|_{L_2}} \right)^2 \right),
\]

one has

\[
\sup_{f \in \mathcal{F}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i \xi_i f(X_i) \right| \leq c_2(q) L w u \| \xi \|_{L_q} \mathbb{E} \| G \|_{\mathcal{F}}.
\]

For every \(u_0 \in U_Y\) let \(\mathcal{F} = \text{star}(U - u_0) \cap rD\). Since \(|U - u_0| = |U| \leq M^2\), it follows from Theorem 5.2 that with probability as in (5.2),

\[
\sup_{f \in \mathcal{F}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \varepsilon_i \xi_i f(X_i) \right| \leq c(q) L w u \| \xi \|_{L_q} r \sqrt{\log M}.
\]

Therefore, if

\[
r \geq c(q) \frac{L w u}{\xi} \cdot \| \xi \|_{L_q} \frac{\sqrt{\log M}}{N},
\]

@ Springer
one has
\[
\sup_{f \in \mathcal{F}} \left| \frac{1}{\sqrt{N}} \sum_{i=1}^{N} \epsilon_i \xi_i f(X_i) \right| \leq \zeta \sqrt{Nr^2}.
\]

Clearly, for any nontrivial class \( \mathcal{F} \), \( \mathbb{E}\|G\|_{\mathcal{F}} \gtrsim \sup_{f \in \mathcal{F}} \|f\|_{L_2} \); thus, setting \( w \sim (1/\delta)^{1/q} \) and \( u \sim \sqrt{\log(2/\delta)} \), with probability at least \( 1 - \delta \)
\[
r_M \leq c_1(q) \frac{L}{\zeta} \cdot \delta^{-1/q} \log^{1/2}(2/\delta) \|\xi\|_{L_q} \sqrt{\frac{\log M}{N}}.
\]

Finally, by the \( L_q - L_2 \) norm equivalence and the fact that \( \|u_0 - Y\|_{L_2} \leq \|f^* - Y\|_{L_2} \), it is straightforward to verify that \( \|\xi\|_{L_q} \leq 3L \|f^* - Y\|_{L_q} \), which completed the proof of Theorem 5.1.

5.2 A remark on the bounded case

Let us briefly mention a way in which one may obtain a version of Theorem 1.10 when both the dictionary and the target are assumed to be bounded in \( L_\infty \), but \( \mathcal{F} \) may be infinite.

As noted in [13], an \( L_\infty \) type of assumption is of a very different nature than an assumption on norm equivalence: the former does not lead to a useful small-ball estimate on class members, and in particular, the proofs presented here do not hold in that case.

Fortunately, there are highly potent tools at one’s disposal when bounded classes are concerned, namely, Talagrand’s concentration inequality for bounded empirical processes and the contraction principle for empirical and Bernoulli processes indexed by bounded classes (see, e.g., [4, 15, 26]). Using that well established machinery, one may show that \( A_{u_0} \) is a high probability event. In fact, thanks to the two-sided concentration estimates, the argument is much simpler.

For example, assuming that the functions involved are bounded by 1 almost surely and applying a contraction argument, it follows that with high probability and in expectation,
\[
\sup_{u \in U} \left| \frac{1}{N} \sum_{i=1}^{N} (u_0(X_i) - Y_i)(u - u_0)(X_i) - \mathbb{E}(u_0(X) - Y)(u - u_0)(X) \right| \lesssim \sup_{u \in U} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i (u - u_0)(X_i) \right|
\]
and
\[
\sup_{u \in U} \left| \frac{1}{N} \sum_{i=1}^{N} (u - u_0)^2(X_i) - \mathbb{E}(u - u_0)^2 \right| \lesssim \sup_{u \in U} \left| \frac{1}{N} \sum_{i=1}^{N} \epsilon_i (u - u_0)(X_i) \right|,
\]
where $(\varepsilon_i)_{i=1}^{N}$ are independent, symmetric $\{-1, 1\}$-valued random variables that are independent of $(X_i, Y_i)_{i=1}^{N}$.

Moreover, the multiplier and quadratic processes exhibit a strong concentration phenomenon, leading to a natural complexity parameter that is rather similar to $r_{\text{opt}}$, and to an exponential probability estimate.

The obvious downside in this concentration-contraction based argument is that it totally eliminates the dependence on the ‘noise-level’ as captures by $\|\xi\|_{L^q}$ (see the discussion in [18, 19] for more details). As an outcome, the estimate in the bounded case does not improve when the problem becomes more ‘realizable’.

References

1. Anthony, M., Bartlett, P.L.: Neural network learning: theoretical foundations. Cambridge University Press, Cambridge (1999)
2. Audibert, J.Y.: Proof of the optimality of the empirical star algorithm, unpublished note. Available at http://certis.enpc.fr/~audibert/Mes%20articles/NIPS07supplem2. Accessed 21 June 2016
3. Audibert, J.Y.: Fast learning rates in statistical inference through aggregation. Ann. Stat. 37(4), 1591–1646 (2009)
4. Boucheron, S., Lugosi, G., Massart, P.: Concentration inequalities: a nonasymptotic theory of independence. Oxford University Press, Oxford (2013)
5. Catoni, O.: Statistical learning theory and stochastic optimization, vol. 1851 of Lecture Notes in Mathematics, Springer, Berlin (2004)
6. de la Peña, V., Giné, E.: Decoupling: from dependence to independence. Springer-Verlag, Berlin (1999)
7. Giné, E., Zinn, J.: Some limit theorems for empirical processes. Ann. Probab. 12(4), 929–989 (1984)
8. Juditsky, A., Rigollet, P., Tsybakov, A.B.: Learning by mirror averaging. Ann. Stat. 36(5), 2183–2206 (2008)
9. Koltchinskii, V., Mendelson, S.: Bounding the smallest singular value of a random matrix without concentration. http://arxiv.org/abs/1312.3580 (preprint)
10. Lecué, G.: HDR Thesis. Available at http://www.cmap.polytechnique.fr/~lecue/HDR. Accessed 21 June 2016
11. Lecué, G., Mendelson, S.: Aggregation via Empirical risk minimization. Probab. Theor. Relat. Fields 145, 591–613 (2009)
12. Lecué, G., Mendelson, S.: Learning subgaussian classes: upper and minimax bounds. Available at http://arxiv.org/abs/1305.4825 (preprint)
13. Lecué, G., Mendelson, S.: Performance of empirical risk minimization in linear aggregation. Bernoulli 22(3), 1520–1534 (2016). doi:10.3150/15-BEJ701
14. Ledoux, M.: The concentration of measure phenomenon. In: Mathematical Surveys and Monographs, vol. 89, p. x+181. American Mathematical Society, Providence, RI (2001)
15. Ledoux, M., Talagrand, M.: Probability in banach spaces. Isoperimetry and processes, Ergebnisse der Mathematik und ihrer Grenzgebiete (3), vol. 23. Springer-Verlag, Berlin (1991)
16. Mendelson, S., Pajor, A., Tomczak-Jaegermann, N.: Reconstruction and subgaussian operators. Geom. Funct. Anal. 17(4), 1248–1282 (2007)
17. Mendelson, S.: A remark on the diameter of random sections of convex bodies, geometric aspects of functional analysis (GAFA Seminar Notes), In: Klartag, B., Milman, E. (eds) Lecture notes in Mathematics 2116, pp. 395–404 (2014)
18. Mendelson, S.: Learning without concentration. J ACM 62(3)(21), 1–25 (2015)
19. Mendelson, S.: Learning without concentration for a general loss function. http://arxiv.org/abs/1410.3192 (preprint)
20. Mendelson, S.: Upper bounds on product and multiplier empirical processes. Available at http://arxiv.org/abs/1410.8003 (preprint)
21. Mossel, E., O’Donnell, R., Oleszkiewicz, K.: Noise stability of functions with low influences: invariance and optimality. Ann. Math. 171(1), 295–341 (2010)
22. Nemirovski, A.: Topics in non-parametric statistics. In: Lectures on probability theory and statistics (Saint-Flour, 1998), vol 1738 of Lecture Notes in Math., pp. 85–277. Springer, Berlin (2000)
23. Pisier, G.: The volume of convex bodies and Banach space geometry. In: Cambridge Tracts in Mathematics, vol. 94, p. xvi+250. Cambridge University Press, Cambridge (1989). doi: 10.1017/CBO9780511666245

24. Talagrand, M.: Sharper bounds for Gaussian and empirical processes. Ann. Probab. 22(1), 28–76 (1994)

25. Tsybakov, A.B.: Introduction to nonparametric estimation. Springer, New York (2009)

26. Van der Vaart, A.W., Wellner, J.A.: Weak convergence and empirical processes. Springer Verlag, Berlin (1996)