The Annals of Statistics
2017, Vol. 45, No. 5, 1835–1862
DOI: 10.1214/16-AOS1510

"LOCAL" VS. "GLOBAL" PARAMETERS—BREAKING THE GAUSSIAN COMPLEXITY BARRIER

BY SHAHAR MENDELSON[1]

Technion—Israel Institute of Technology

We show that if F is a convex class of functions that is L-sub-Gaussian, the error rate of learning problems generated by independent noise is equivalent to a fixed point determined by "local" covering estimates of the class (i.e., the covering number at a specific level), rather than by the Gaussian average, which takes into account the structure of F at an arbitrarily small scale. To that end, we establish new sharp upper and lower estimates on the error rate in such learning problems.

1. Introduction. The focus of this article is on the question of *prediction*. Given a class of functions F defined on a probability space (Ω, μ) and an unknown target random variable Y, one would like to identify an element of F whose "predictive capabilities" are (almost) the best possible in the class. The notion of "best" is measured via the pointwise cost of predicting $f(x)$ instead of y, and the best function in the class is the one that minimizes the average cost. Here, we will consider the squared loss: the cost of predicting $f(x)$ rather than y is $(f(x) - y)^2$, and if X is distributed according to μ, the goal is to identify

$$f^* = \operatorname*{argmin}_{f \in F} \mathbb{E}\big(f(X) - Y\big)^2 = \operatorname*{argmin}_{f \in F} \big\| f(X) - Y \big\|_{L_2}^2,$$

where the expectation is taken with respect to the joint distribution of X and Y on the product space $\Omega \times \mathbb{R}$.

The information at one's disposal is rather limited: a random sample $(X_i, Y_i)_{i=1}^N$, selected according to the N-product of the joint distribution of X and Y. And, using this data, one must produce some (random) $f \in F$.

DEFINITION 1.1. Given a sample size N and a class F defined on (Ω, μ), a learning procedure is a map $\Psi : (\Omega \times \mathbb{R})^N \to F$. For a set $\mathcal{Y}$ of admissible targets, Ψ performs with confidence $1 - \delta$ and accuracy $\mathcal{E}_p$ if for every $Y \in \mathcal{Y}$, and setting $\tilde{f} = \Psi((X_i, Y_i)_{i=1}^N)$,

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

Received October 2015; revised August 2016.

[1]Supported in part by the Mathematical Sciences Institute, The Australian National University, Canberra, ACT 2601, Australia. Additional support was given by an Israel Science Foundation grant.

MSC2010 subject classifications. 62G08, 62C20, 60G15.

Key words and phrases. Error rates, Gaussian averages, covering numbers.

with probability at least $1-\delta$ relative to the N-product of the joint distribution of X and Y.

The accuracy (or *prediction error*) $\mathcal{E}_p$ is a function of F, N and δ, and may depend on certain properties of the target Y as well, for example, its norm in some L_q space.

A fundamental problem in learning theory is to identify the features of the underlying class F and of the set of admissible targets $\mathcal{Y}$ that govern $\mathcal{E}_p$. Moreover, a question of particular significance is the way in which $\mathcal{E}_p$ scales with the sample size N (the so-called *error rate*). This question has been studied extensively, and we refer the reader to the manuscripts [2, 3, 5, 8, 10, 11, 15] for more information on its history and on some more recent progress.

The aim of this article is to obtain *matching* upper and lower bounds on $\mathcal{E}_p$, at least, under some additional assumptions.

1.1. *$\mathcal{E}_p$ and the structure of F.* Before diving into an accurate (and somewhat technical) description of our results, let us present a brief overview, trying to put the question at hand in some perspective.

It is well understood that some notion that captures the "statistical size" of F must play a dominant role in the characterization of $\mathcal{E}_p$. However, and regardless of the notion of size one uses, there are situations in which geometric obstructions distort the effect the size of F has on $\mathcal{E}_p$.

A simple, yet in some sense generic example of such a distorted behaviour is when $F=\{f_1, f_2\}$ and Y is a $1/\sqrt{N}$-perturbation of the midpoint $(f_1+f_2)/2$. Although F is clearly a small class, one may show that no learning procedure can perform with an error rate that is better than $c/\sqrt{N}$, having been given a sample of cardinality N (see, e.g., [1]). On the other hand, and based solely on the size of F, one would expect a much faster error rate for problems involving such a class—exhibiting the distortion in $\mathcal{E}_p$.

This type of distortion may be avoided by imposing additional geometric conditions on F and $\mathcal{Y}$, which ensure that all the admissible targets in $\mathcal{Y}$ are located in a favourable position relative of F (see [9] for an accurate definition of "a favourable position"). For instance, one may show that if $F \subset L_2(\mu)$ is compact and convex, any target $Y \in L_2$ is in a favourable position relative to F; hence, regardless of the target Y, there is no distortion in $\mathcal{E}_p$ for convex classes.

To avoid potential geometric obstructions, we will focus on the study of the error rate only when F is a convex class.

Intuitively, once the geometric obstructions have been removed, $\mathcal{E}_p$ should depend on two key features of F:

- The right notion of the class' intrinsic complexity, which captures the difficulty of prediction problems for targets of the form $Y=f(X)$, for some $f \in F$.

- The way class members interact with the "noise" $Y - f^*(X)$ [and the reason for calling $Y - f^*(X)$ "the noise" originates from the case $Y = f(X) + W$, where $f \in F$ and W is independent of X].

Recent results [6, 10, 11] show that this intuitive description is indeed true. Moreover, and as will be explained in greater detail in what follows, the "intrinsic complexity" of F and the way class members interact with $Y - f^*(X)$ may be upper bounded by controlling the suprema of two empirical processes. Thus, for a convex class, the question of an upper bound on $\mathcal{E}_p$ is reduced to obtaining upper estimates on the suprema of those empirical processes. Moreover, it is well understood (see, e.g., [16]) that one may obtain such upper bounds using *random* L_2 *covering numbers*; that is, the number of balls of a given radius, and with respect to the random distance $d^2(f,h) = N^{-1}\sum_{i=1}^N (f-h)^2(X_i)$ endowed by the sample $(X_1, \ldots, X_N)$, that are needed to cover the indexing class. What is the key point in the context of this article, is that the upper bounds are given by an aggregate functional—a so-called *entropy integral*—which takes into account the random covering numbers at an arbitrary small scale, rather than the covering numbers at a single, well-chosen scale.

Unfortunately, random covering numbers are a notoriously difficult object to handle. In the past, it has been standard practice to control them using other complexity parameters of the indexing class, like the *combinatorial dimension* (e.g., the VC-dimension or its scale-sensitive versions). Another alternative it to assume that F is a *sub-Gaussian class* (see Definition 2.1), in which case the empirical entropy integral may be replaced by the expectation of the supremum of the Gaussian process indexed by certain subsets of F, or by an entropy integral generated by the $L_2(\mu)$ covering numbers of F. Again, all these parameters take into account the structure of the class at an arbitrary small scale, and as such, are of global nature.

In contrast, known lower bounds on $\mathcal{E}_p$ are based on a different approach and on "local" parameters—specifically, on the $L_2(\mu)$ covering numbers of F at a *single level*. Thus, and somewhat roughly put, the question we would like to study is whether the true behaviour of $\mathcal{E}_p$ is determined by parameters of a global nature (an entropy integral that takes into account the covering numbers at an arbitrarily small level, Gaussian averages associated with the class, etc.), or of a local nature (covering numbers at one level).

It should be noted that prior to this work, there were no known results bridging the gap between the global upper bounds and the local lower ones—except when the two happen to be equivalent. However, in general, there is a gap between the global and the local (see more on that in Section 2.4 and in the supplementary material to this article [13]).

The next section is devoted to a more accurate description of the problem and the parameters involved.

2. The "global" and the "local". From here on, let $F \subset L_2(\mu)$ be a convex class of functions (thus preventing geometric obstructions that distort $\mathcal{E}_p$).

Recall that the centred, canonical Gaussian process indexed by F is a random process that assigns to each $f \in F$ a centred Gaussian variable G_f. The covariance structure of the process is given by $\mathbb{E}G_f G_h = \mathbb{E}fh$, that is, it is endowed by the inner product in $L_2(\mu)$.

A Gaussian process is bounded if $\sup_{f\in F} G_f$ is bounded almost surely, and here we will ignore the question of the measurability of $\sup_{f\in F} G_f$. For the same reason, we set

$$\mathbb{E}\|G\|_F = \sup\Big\{\mathbb{E}\sup_{f\in F'} G_f : F' \subset F, F' \text{ is finite}\Big\}.$$

We refer the reader to the books [4, 7, 16] for more information on Gaussian processes.

DEFINITION 2.1. A class $F \subset L_2(\mu)$ is L-sub-Gaussian with respect to the measure μ if for every $p \geq 2$ and every $f, h \in F \cup \{0\}$,

$$\|f - h\|_{L_p(\mu)} \leq L\sqrt{p}\|f - h\|_{L_2(\mu)},$$

and if the centred canonical Gaussian process $\{G_f : f \in F\}$ is bounded.

A survey on the properties of sub-Gaussian classes may be found in [4, 6, 7, 14, 16].

EXAMPLE. Let $T \subset \mathbb{R}^n$ be convex. For every $t \in T$, set $f_t = \langle t, \cdot\rangle$ and put $F_T = \{f_t : t \in T\}$. Hence, F_T is the class of linear functionals generated by T, and since T is convex, so is F_T.

Denote by $\|\ \|_{\ell_2^n}$ the standard Euclidean norm on $\mathbb{R}^n$. Recall that a probability measure μ on $\mathbb{R}^n$ is *isotropic* if it is symmetric and satisfies

$$\int_{\mathbb{R}^n} \langle t, x\rangle^2 \, d\mu(x) = \|t\|_{\ell_2^n}^2 \qquad \text{for every } t \in \mathbb{R}^n.$$

If X is distributed according to an isotropic measure μ, and if for every $z \in \mathbb{R}^n$ and $p \geq 2$, $(\mathbb{E}|\langle z, X\rangle|^p)^{1/p} \leq L\sqrt{p}\|z\|_{\ell_2^n}$, then F_T is an L-sub-Gaussian class. Indeed, for every $f_t, f_s \in F_T$ and $p \geq 2$,

$$\|f_s - f_t\|_{L_p(\mu)} = \big(\mathbb{E}|\langle X, s - t\rangle|^p\big)^{1/p} \leq L\sqrt{p}\|s - t\|_{\ell_2^n} = L\sqrt{p}\|f_s - f_t\|_{L_2(\mu)}.$$

There are numerous examples of isotropic, L-sub-Gaussian measures on $\mathbb{R}^n$ for L that is an absolute constant, independent of the dimension or of any other parameter. To name a few, the standard Gaussian measure on $\mathbb{R}^n$; the uniform measure on $\{-1, 1\}^n$; any n-product measure given by $X = (x_1, \ldots, x_n)$ that is endowed by n independent copies of a mean-zero, variance 1, random variable x

that satisfies $\|x\|_{L_p} \leq c\sqrt{p}$ for every $p \geq 2$; and the normalized volume measure on the set $n^{1/p}B_p^n$, where B_p^n is the unit ball of normed space $\ell_p^n = (\mathbb{R}^n, \|\ \|_{\ell_p})$ and $p \geq 2$.

For a reason that will become clear later, we will not study a general class of admissible targets $\mathcal{Y}$, but rather consider targets of the form $Y = f(X) + W$ for some $f \in F$ and W that is orthogonal to span(F)—the linear span of F [and of course, if $W \notin L_2(\mu)$ one considers orthogonality relative to the natural L_2 space to which both F and W belong].

One significant example that fits the setup we shall study is when W is a mean-zero random variable that is independent of X. Also, observe that for such targets the minimizer in F of $h \to \mathbb{E}(h(X) - Y)^2$ is f, and thus $\mathcal{E}_p = \|\tilde{f} - f\|_{L_2(\mu)}^2$.

With that in mind, let us formulate the question we would like to explore.

QUESTION 2.2. Let $F \subset L_2(\mu)$ be a compact, convex class that is L-sub-Gaussian with respect to μ. Given targets of the form $Y = f(X) + W$ as above, find matching upper and lower bounds (up to constants) on $\mathcal{E}_p$.

2.1. *Global parameters and upper bounds.* A relatively standard way of establishing upper bounds $\mathcal{E}_p$ is the following decomposition of the squared excess loss: let Y be the unknown target and recall that $f^* = \operatorname{argmin}_{f \in F} \|f(X) - Y\|_{L_2}$. For every $f \in F$, let $\ell_f(X, Y) = (f(X) - Y)^2$ and set

$$\begin{aligned} \mathcal{L}_f^F(X, Y) &= (\ell_f - \ell_{f^*})(X, Y) = \big(f(X) - Y\big)^2 - \big(f^*(X) - Y\big)^2 \\ &= 2\big(f^*(X) - Y\big)\big(f - f^*\big)(X) + \big(f - f^*\big)^2(X). \end{aligned} \tag{2.1}$$

For a sample $(X_i, Y_i)_{i=1}^N$ consisting of N independent copies of (X, Y), let $P_N h = \frac{1}{N}\sum_{i=1}^N h(X_i, Y_i)$ and set

$$\hat{f} = \operatorname*{argmin}_{f \in F} P_N \ell_f = \operatorname*{argmin}_{f \in F} P_N \mathcal{L}_f^F,$$

where the second equality holds because $\mathcal{L}_f^F$ is a shift of each ℓ_f by the same function ℓ_{f^*}.

The learning procedure that assigns to every sample $(X_i, Y_i)_{i=1}^N$ the function $\hat{f} \in F$ is called *Empirical Risk Minimization* (ERM), and $\hat{f}$ is the empirical risk minimizer.

Clearly, $\mathcal{L}_{f^*}^F = 0$; thus, for every sample $(X_i, Y_i)_{i=1}^N$, $P_N \mathcal{L}_{\hat{f}}^F \leq 0$, implying that members of the random set $\{f \in F : P_N \mathcal{L}_f^F > 0\}$ cannot be empirical risk minimizers. Moreover, the decomposition (2.1) provides one with a way of identifying the random set in question. Indeed, assume that $(X_i, Y_i)_{i=1}^N$ is a sample for which, if $\|f - f^*\|_{L_2} \geq r$, one has

$$\frac{1}{N}\sum_{i=1}^N \big(f - f^*\big)^2(X_i) \geq \kappa \big\|f - f^*\big\|_{L_2}^2, \tag{2.2}$$

and setting $\xi = f^*(X) - Y$ and $\xi_i = f^*(X_i) - Y_i$,

$$\left| \frac{1}{N} \sum_{i=1}^N \xi_i (f - f^*)(X_i) - \mathbb{E}\xi (f - f^*)(X) \right| \leq \frac{\kappa}{4} \|f - f^*\|_{L_2}^2. \tag{2.3}$$

Observe that by the characterization of the metric projection onto a closed convex set in an inner product space, one has

$$\mathbb{E}\xi (f - f^*)(X) \geq 0 \qquad \text{for every } f \in F. \tag{2.4}$$

Hence, if (2.2) and (2.3) hold for the sample $(X_i, Y_i)_{i=1}^N$, then for every $f \in F$ that satisfies $\|f - f^*\|_{L_2} \geq r$,

$$\begin{aligned} P_N \mathcal{L}_f^F &\geq \frac{1}{N} \sum_{i=1}^N (f - f^*)^2 (X_i) - 2 \left| \frac{1}{N} \sum_{i=1}^N \xi_i (f - f^*)(X_i) - \mathbb{E}\xi (f - f^*)(X) \right| \\ &\quad + \mathbb{E}\xi (f - f^*)(X) \geq (\kappa - 2(\kappa/4)) r^2 > 0; \end{aligned}$$

in particular, $\{f \in F : \|f - f^*\|_{L_2} \geq r\} \subset \{f \in F : P_N \mathcal{L}_f^F > 0\}$, and thus $\|\hat{f} - f^*\|_{L_2} < r$.

REMARK 2.3. This argument had been used in [11] and was extended further in [10], showing that

$$\mathbb{E}(\mathcal{L}_{\hat{f}}^F | (X_i, Y_i)_{i=1}^N) \leq r^2$$

for general convex losses ℓ and not just for the squared loss.

In what follows, we shall refer to the process $f \to \frac{1}{N} \sum_{i=1}^N (f - f^*)^2 (X_i)$ as the *quadratic component* and to $f \to \frac{1}{N} \sum_{i=1}^N \xi_i (f - f^*)(X_i) - \mathbb{E}\xi (f - f^*)(X)$ as the *multiplier component* [the name of the latter originating from the multipliers $(\xi_i)_{i=1}^N$ that do not depend on $f \in F$]. Observe that if $Y = f(X)$ for some $f \in F$ then $\xi = 0$ and the multiplier component is trivial.

With the decomposition of the squared loss at our disposal, let us define two parameters that can be used to measure the complexity of F. It turns out that one captures the class' intrinsic complexity via the quadratic component, while the other governs the interaction class members have with the noise ξ, via the multiplier component.

Let $F \subset L_2(\mu)$. Set $F - h = \{f - h : f \in F\}$ and $F - F = \{f - h : f, h \in F\}$, and put D to be the unit ball in $L_2(\mu)$.

DEFINITION 2.4. For $\kappa_1, \kappa_2 > 0$ let

$$r_M(\kappa_1, f) = \inf\{r > 0 : \mathbb{E}\|G\|_{(F-f) \cap rD} \leq \kappa_1 r^2 \sqrt{N}\} \tag{2.5}$$

and

$$r_Q(\kappa_2, f) = \inf\{r > 0 : \mathbb{E}\|G\|_{(F-f) \cap rD} \leq \kappa_2 r \sqrt{N}\}. \tag{2.6}$$

Put

$$r_M(\kappa_1) = \sup_{f \in F} r_M(\kappa_1, f) \quad \text{and} \quad r_Q(\kappa_2) = \sup_{f \in F} r_Q(\kappa_2, f).$$

As may be indicated by their names, r_M is used to govern the multiplier component and r_Q controls the quadratic component.

It is important to stress the global nature of r_Q and r_M. The two depend on Gaussian oscillations of the form $\mathbb{E}\|G\|_{(F-f)\cap rD}$. And although $(F-f)\cap rD$ is a localized set—obtained by intersecting $F-f$ with a ball of radius r, $\mathbb{E}\|G\|_{(F-f)\cap rD}$ is not determined solely by the structure of $F-f$ at a scale that is proportional to r. In fact, it is straightforward to construct examples in which $\mathbb{E}\|G\|_{(F-f)\cap rD}$ is dictated by a subset of $F-f$ consisting of functions whose $L_2(\mu)$ norms are well below r. Moreover, this global behaviour cannot be avoided: $\mathbb{E}\|G\|_{(F-f)\cap rD}$ often captures the true nature of the quadratic and multiplier components, and the two are highly affected by the "richness" of F around f at every level—even by functions that are very close to the centre f.

An upper estimate on $\mathcal{E}_p$ using the complexity parameters r_M and r_Q has been established in [6], and here it is formulated only in the context we are interested in—for targets of the form $Y = f(X) + W$ for W that is orthogonal to $\operatorname{span}(F)$.

THEOREM 2.5. *For every $L \geq 1$, there exist constants c_1, c_2, c_3 and c_4 that depend only on L for which the following holds. Let $F \subset L_2(\mu)$ be a compact, convex, L-sub-Gaussian class of functions, set $Y = f(X) + W$ for $f \in F$ and W that is orthogonal to* $\operatorname{span}(F)$. *Assume further that for every $p \geq 2$, $\|W\|_{L_p} \leq L\sqrt{p}\|W\|_{L_2}$.*

There is a learning procedure (empirical risk minimization performed in F) for which, if

$$r \geq 2\max\{r_M(c_0/\|W\|_{L_2}), r_Q(c_1)\} \equiv r^*,$$

then with probability at least $1 - 2\exp(-c_2 N \min\{1, (r^*)^2/\|W\|_{L_2}^2\})$, the error of the procedure is at most $\mathcal{E}_p \leq r^2$.

As mentioned previously, one may also provide upper bounds on $\mathcal{E}_p$ using the notion of covering numbers, and thanks to the sub-Gaussian assumption, the covering numbers in question are with respect to the $L_2(\mu)$ norm.

DEFINITION 2.6. Let B be a unit ball of a normed space. Set $\mathcal{N}(A, B)$ to be the minimal number of centres $a_1, \ldots, a_n \in A$ for which $A \subset \bigcup_{i=1}^n (a_i + B)$. $(a_i)_{i=1}^n$ is called a cover of A with respect to B.

An r-cover is a cover with respect to $rB = \{rb : b \in B\}$—the ball of radius r—rather than with respect to B.

$\mathcal{M}(A, rB)$ is the cardinality of a maximal r-separated subset of A with respect to the given norm, that is, the cardinality of the largest subset $(a_i)_{i=1}^m \subset A$ for which $\|a_i - a_j\| \geq r$ for every $i \neq j$.

It is standard to verify that $\mathcal{M}(A, 2B) \leq \mathcal{N}(A, B) \leq \mathcal{M}(A, B)$ (see, e.g., Theorem 1.2.1 in [4]). Moreover, a fundamental fact in the theory of Gaussian processes is that there is an absolute constant[2] c for which

$$\mathbb{E}\|G\|_F \leq c \int_0^{\operatorname{diam}(F, L_2(\mu))} \sqrt{\log \mathcal{N}(F, \varepsilon D)}\, d\varepsilon.$$

This entropy integral bound, due to Dudley (see, e.g., [4, 16]), serves as a further indication of the global nature of $\mathbb{E}\|G\|_{(F-f)\cap rD}$, as the estimate requires information on the $L_2(\mu)$ covering numbers of F at an arbitrarily small scale.

2.2. *Local parameters and lower bounds.* Unlike the global nature of the upper bounds described above, the known lower bounds on $\mathcal{E}_p$ are based on the cardinality of a well-separated subset of F at a single level. We will outline the reason why that is natural in what follows, but first let us introduce the two "local" counterparts of r_M and r_Q.

DEFINITION 2.7. For $\eta_1, \eta_2 > 0$, set

$$\lambda_M(\eta_1, f) = \inf\{r > 0 : \log \mathcal{M}\big((F - f) \cap 4rD, (r/2)D\big) \leq \eta_1^2 r^2 N\}$$

and

$$\lambda_Q(\eta_2, f) = \inf\{r > 0 : \log \mathcal{M}\big((F - f) \cap 4rD, (r/2)D\big) \leq \eta_2^2 N\}.$$

Put $d = \operatorname{diam}(F, L_2(\mu))$ and set

$$\lambda_M(\eta_1) = \min\Big\{\sup_{f \in F} \lambda_M(\eta_1, f), d/4\Big\},$$

and

$$\lambda_Q(\eta_2) = \min\Big\{\sup_{f \in F} \lambda_Q(\eta_2, f), d/4\Big\}.$$

The connection between the global and local parameters is another fundamental fact in the theory of Gaussian processes—Sudakov's inequality (see, e.g., [7]): there is an absolute constant c for which, for every $H \subset L_2(\mu)$,

$$c \sup_{\varepsilon > 0} \varepsilon \log^{1/2} \mathcal{M}(H, \varepsilon D) \leq \mathbb{E}\|G\|_H.$$

Note that $r_M(\kappa_1) \leq 4r$ if for every $f \in F$, $\mathbb{E}\|G\|_{(F-f)\cap 4rD} \leq \kappa_1 (4r)^2 \sqrt{N}$. Applying Sudakov's inequality to $H = (F - f) \cap 4rD$ and for the choice of $\varepsilon = r/2$, one has

$$c(r/2) \log^{1/2} \mathcal{M}\big((F - f) \cap 4rD, (r/2)D\big) \leq \mathbb{E}\|G\|_{(F-f)\cap 4rD} \leq 16\kappa_1 r^2 \sqrt{N},$$

[2]Here and throughout the article, absolute constants are simply fixed positive numbers that are independent of any other parameter associated with the problem.

which shows that $\lambda_M(c_1\kappa_1) \leq r$. A similar observation is true for r_Q and λ_Q. Hence, λ_M and λ_Q, which are not affected by the structure of F at a level below $r/2$, are indeed smaller than r_Q and r_M, respectively.

The following is an example of a lower bound on $\mathcal{E}_p$ in terms of λ_M.

THEOREM 2.8 ([6]). *There exist absolute constants c_1 and c_2 for which the following holds. Let F be a class of functions, set W to be a centred normal random variable that is independent of X, and for every $f \in F$ put $Y^f = f(X) + W$. If Ψ is a learning procedure that performs for every target Y^f with confidence at least $3/4$, then there is some Y^f for which $\mathcal{E}_p \geq c_1\lambda_M^2(c_2/\|W\|_{L_2})$.*

Let us emphasize once again that while there are cases in which the estimates in Theorem 2.5 and Theorem 2.8 coincide, it is not the generic situation, and typically there is a gap between the two.

2.3. *The main results.* Given that in the generic problem there is a gap between the two sets of parameters, one must face the obvious question: which of the two captures the correct behaviour of $\mathcal{E}_p$? Is it the "global" pair, r_Q and r_M, or the "local" one of λ_Q and λ_M?

Our main result is that the "local" parameters are the right answer—at least in the setup outlined above. To that end, we shall improve the upper bound in Theorem 2.5 from dependence on the global parameters to a dependence on the local ones. We will also add the component missing from Theorem 2.8, namely, a lower bound in term of λ_Q.

THEOREM 2.9. *For every $L > 1$ and $q > 2$, there are constants $c_0, \ldots, c_5$ that depend only of q and L for which the following holds. Let $F \subset L_2(\mu)$ be a compact, convex, L-sub-Gaussian class of functions with respect to μ. There is a learning procedure $\Psi : (\Omega \times \mathbb{R})^N \to F$, for which, if $Y = f(X) + W$ for $f \in F$ and $W \in L_q$ that is orthogonal to* $\operatorname{span}(F)$, *then with probability at least*

$$1 - 2\exp\bigl(-c_0 N \min\bigl\{1, \lambda_M^2\bigl(c_1/\|W\|_{L_q}\bigr)\bigr\}\bigr) - c_2\frac{\log^q N}{N^{(q/2)-1}},$$

$$\mathcal{E}_p \leq c_3 \max\left\{\lambda_M^2\left(\frac{c_1}{\|W\|_{L_q}}\right), \lambda_Q^2(c_4)\right\} + r_Q^2(c_4)\exp\bigl(-c_5\exp(N)\bigr).$$

Note that W need not be sub-Gaussian—Theorem 2.9 is valid even for a heavy-tailed noise. Also, the term $r_Q^2(c_4)\exp(-c_5\exp(N))$ is almost certainly an artifact of the proof, but in any case, it is significantly smaller than the dominating term in any reasonable example.

The proof of Theorem 2.9 is based on a rather obvious idea: "erasing" all the fine structure of F, by replacing the class with an appropriate separated subset.

The difficultly in such an approach is that by changing the geometry of F, one re-introduces the geometric obstructions that distort $\mathcal{E}_p$ and which have been mentioned previously.

To complement Theorem 2.9, we also obtain the following lower bound.

THEOREM 2.10. *There exist absolute constants c_0 and c_1 for which the following holds. Let $F \subset L_2(\mu)$ be a class of functions and let Ψ be any learning procedure that performs with confidence $7/8$ for any target of the form $Y^f = f(X) + W$ for $f \in F$ and $W \in L_2$ that is orthogonal to* $\operatorname{span}(F)$.

(a) *If F is convex and centrally-symmetric,*[3] *then for any $W \in L_2$ that is orthogonal to* $\operatorname{span}(F)$, *there is some $f \in F$, for which, for Y^f as above,*

$$\mathcal{E}_p \geq c_0 \lambda_Q^2(c_1).$$

(b) *Let W be a centred random variable with density* $\exp(-\phi)$ *for an even function ϕ that satisfies* $\sup_{t\in\mathbb{R}} |\phi''(t)| \leq \kappa$. *If W is independent of X, there is some $f \in F$ for which, for Y^f as above,*

$$\mathcal{E}_p \geq c_0 \lambda_M^2(c_1\kappa).$$

In particular, if W is a centred, normal random variable, there is some $f \in F$ for which, for Y^f as above,

$$\mathcal{E}_p \geq c_0 \lambda_M^2\biggl(\frac{c_1}{\|W\|_{L_2}}\biggr).$$

The obvious outcome of Theorem 2.9 and Theorem 2.10 is that if W is a centred Gaussian random variable that is independent of X, then for any convex, centrally-symmetric, L-sub-Gaussian class F, the upper and lower estimates match [up to the parasitic and negligible term $r_Q^2(c_4)\exp(-c_5\exp(N))$ in the upper bound]: when considering targets of the form $Y = f(X) + W$ for $f \in F$,

$$\mathcal{E}_p \sim \max\bigl\{\lambda_Q^2(c_1), \lambda_M^2\bigl(c_2/\|W\|_{L_2}\bigr)\bigr\}.$$

REMARK 2.11. One should note that there were no known lower bounds based on λ_Q prior to this work. Also, the Gaussian version of part (b) of Theorem 2.10 is well known (see, e.g., [15]), but existing proofs seem to rely heavily on the fact that W is Gaussian. The proof we shall present holds in a more general situation than part (b) (see Theorem 5.6, below). It is based on a new volumetric argument which we believe to be of independent interest and which we will now outline.

[3] F is centrally-symmetric if the fact that $f \in F$ implies that $-f \in F$.

To explain why a lower bound based on the cardinality of a separated subset of F at a specific level is possible, assume that a procedure $\tilde{f}$ is accurate up to an error of ε for every target of the form $Y = f(X) + W$, for $f \in F$ and W that is mean-zero and is independent of X. Thus, a procedure performs well on the sample $(X_i, f(X_i) + W_i)_{i=1}^N$ if the function $\tilde{f}$ it generates given the sample satisfies $\|\tilde{f} - f\|_{L_2}^2 \leq \varepsilon$.

Let $\|f_1 - f_2\|_{L_2}^2 \geq 2\varepsilon$, fix a sample $\mathbb{X} = (X_i)_{i=1}^N$ and set

$$z_1 = \big(f_1(X_i)\big)_{i=1}^N \quad \text{and} \quad z_2 = \big(f_2(X_i)\big)_{i=1}^N.$$

Consider the sets $A_j(\mathbb{X}) \subset \mathbb{R}^N$, $j = 1, 2$, each consisting of all $(w_1, \ldots, w_N) \in \mathbb{R}^N$ on which $\tilde{f}$ performs well after being given $(X_i, Y_i)_{i=1}^N = (X_i, f_j(X_i) + w_i)_{i=1}^N$ as data. Note that the sets $z_1 + A_1(\mathbb{X})$ and $z_2 + A_2(\mathbb{X})$ must be disjoint, because for each $t \in z_j + A_j(\mathbb{X})$, the pair $(\mathbb{X}, t)$ is mapped by the procedure to a ball of radius $\sqrt{\varepsilon}$ around f_j—but the two balls do not intersect.

This simple argument will be used in Section 5 to show that a $2\sqrt{\varepsilon}$-separated subset of F endows a collection of disjoint subsets of $\mathbb{R}^N$ [that depends on $(X_i)_{i=1}^N$]; the high confidence of $\tilde{f}$ implies that for most samples $(X_i)_{i=1}^N$, each one of the disjoint subsets must have a nonnegligible probability with respect to the measure endowed on $\mathbb{R}^N$ by $(W_1, \ldots, W_N)$. Therefore, the number of subsets in the collection, which is the cardinality of the separated set, cannot be too big.

REMARK 2.12. Let us mention that if F happens to be convex and centrally symmetric, what is essentially the "richest" shift of F is the 0-shift. Indeed, since $F - F = 2F = \{2f : f \in F\}$, it is evident that for every $f \in F$

$$(F - f) \cap 4rD \subset (F - F) \cap 4rD = 2(F \cap 2rD).$$

This fact makes one's life much simpler when studying lower bounds, as it gives an obvious choice of where to look. Indeed, the "richest" part of F is the hardest part for a learning procedure to deal with—and that part is a neighbourhood of 0.

2.4. *The Yang–Barron theorem.* One result that seems similar to ours may be found in the celebrated work of Yang and Barron [17].

Yang and Barron study various prediction problems and obtain upper and lower bounds on the error rate that are of the order ε_N^2, for ε_N that satisfies $\log \mathcal{M}(F, \varepsilon D) = N\varepsilon^2$; as such, ε_N is closely related to λ_M. However, a closer inspection of [17] shows that there are substantial differences between those bounds and ours.

To begin with, the setup in [17] is different: a function class consisting of uniformly bounded functions and the noise is independent Gaussian noise, both of which are crucial to the proof (see Section 3.2 in [17]). Moreover, the upper estimate is an existence result of a "good" procedure—rather than a specific choice of

a procedure; the estimates hold in expectation and not with high probability; and they do not tend to zero with the "noise level" of the problem (i.e., the variance of the Gaussian noise).

All these differences are significant, but still are not a conclusive indication that the results in [17] are of a different nature to ours. The key point is an assumption that is at the heart of [17]: that the underlying class is "large"—in the sense that

$$\liminf_{\varepsilon\to 0} \frac{\log \mathcal{M}(F,(\varepsilon/2)D)}{\log \mathcal{M}(F,\varepsilon D)} = \kappa > 1. \tag{2.7}$$

It turns out that this assumption excludes all linear regression problems involving classes of linear functionals that are indexed by subsets of $\mathbb{R}^n$; in particular, all the modern "high-dimensional" learning problems—as we now show.

EXAMPLE 2.13. Fix $T \subset \mathbb{R}^n$ and let $F_T = \{\langle t, \cdot\rangle : t \in T\}$ be the class of linear functionals indexed by T. For the sake of simplicity, first assume that the underlying measure μ is an isotropic measure on $\mathbb{R}^n$; thus, the $L_2(\mu)$ unit ball endowed on $\mathbb{R}^n$ coincides with B_2^n, the standard Euclidean unit ball in $\mathbb{R}^n$.

If $T \subset RB_2^n$ and has a nonempty interior, it also contains a Euclidean ball, say of radius $\rho > 0$. Recall that by a straightforward volumetric estimate, if $0 < \varepsilon \le \rho/2$ then $(c_1\rho/\varepsilon)^n \le \mathcal{M}(\rho B_2^n, \varepsilon B_2^n) \le (c_2\rho/\varepsilon)^n$ for suitable absolute constants c_1 and c_2. Thus, for ε small enough,

$$\left(\frac{c_1\rho}{\varepsilon}\right)^n \le \mathcal{M}(T,\varepsilon D) \le \left(\frac{c_2 R}{\varepsilon}\right)^n, \tag{2.8}$$

and κ [the liminf in (2.7)] is 1.

Also, $\kappa = 1$ even if μ is not isotropic, as long as it is not supported on a hyperplane. For a nontrivial μ, the $L_2(\mu)$ unit ball endowed on $\mathbb{R}^n$ is an ellipsoid with a nonempty interior; at very small scales ε, covering by such an ellipsoid is equivalent to covering by εB_2^n.

Condition (2.7) has other implications whose proof may be found in the supplementary material to this article [13]:

- If $\kappa > 1$, then the $r/2$ log-covering numbers of F and of $F \cap rD$ are, in some sense, equivalent, which means that F is very rich locally.
- If $\kappa > 1$, then the local and global parameters, λ_M/λ_Q and r_M/r_Q respectively, are equivalent. Indeed, for the sake of brevity let us ignore cases in which

$$\limsup_{\varepsilon\to 0} \frac{\log \mathcal{M}(F,(\varepsilon/2)D)}{\log \mathcal{M}(F,\varepsilon D)} = \ell \ge 4$$

(if $\ell > 4$ then the centred, canonical Gaussian process $\{G_f : f \in F\}$ is not bounded, while if $\ell = 4$, covering estimates are not enough to determine whether the Gaussian process is bounded).

Hence, the case that is comparable to ours is when $\ell < 4$. In that case, one may show that for $R \le R_0$, and for the "worst" $f \in F$,

$$\mathbb{E}\|G\|_{(F-f)\cap RD} \sim_{\kappa,\ell} R\log^{1/2}\mathcal{M}\big((F-f)\cap RD, (R/2)D\big);$$

therefore, if $\kappa > 1$ and $\ell < 4$, the global parameters r_M and r_Q are equivalent to the local ones λ_M and λ_Q, and there is nothing to be gained by the "local analysis" we present in what follows.

3. Preliminaries. Let us begin with some natation. Throughout, absolute constants are denoted by $c, c_1, \ldots$ etc. Their value may change from line to line. $c(\alpha)$ is a constant that depends only on the parameter α, and $A \sim_p B$ means that $c_1(p)A \le B \le c_2(p)A$. We use $\kappa_1, \kappa_2, \eta_1, \eta_2$, etc. to denote fixed constants whose value remains unchanged throughout the article.

In what follows, we will, at times, abuse notation and not specify the probability space on which each random variable is defined. For example, $\|f - Y\|_{L_2}^2 = \mathbb{E}(f(X) - Y)^2$ and integration is with respect to the joint distribution of X and Y, while $\|f - h\|_{L_2}^2 = \mathbb{E}(f-h)^2(X)$, in which case integration is with respect to μ. The same goes for the notion of orthogonality—where the underlying assumption is that the functions involved belong to a single Hilbert space that contains $L_2(\mu)$.

Recall that D the unit ball in $L_2(\mu)$ and set $\operatorname{star}(F) = \{\lambda f : f \in F\ 0 \le \lambda \le 1\}$; $\operatorname{star}(F)$ is the star-shaped hull of F with 0. We say that F is star-shaped around 0 if $\operatorname{star}(F) = F$.

The following lemma is straightforward but still plays a crucial part in the proof of Theorem 2.9.

LEMMA 3.1. *Let $T \subset W \subset L_2(\mu)$. For $s > r > 0$, set*

$$\phi(s,r) = \sup_{w\in W} \mathcal{N}\big(T\cap(w+sD), rD\big).$$

Then:

1. $\phi(s,r) \le \phi(s,s/2)\cdot\phi(s/2,r)$.
2. *If T and W are star-shaped around 0 then*

$$\log\phi(s,r) \le c_0\log(2s/r)\cdot\log\phi(4r,r)$$

for a suitable absolute constant c_0.

PROOF. Fix $w \in W$ and let $t_1, \ldots, t_N \in T\cap(w+sD)$ be centres of a minimal $s/2$-cover of that set. Note that for every $1 \le i \le N$,

$$T\cap(w+sD)\cap\big(t_i+(s/2)D\big) \subset T\cap\big(t_i+(s/2)D\big),$$

and $\mathcal{N}(T\cap(t_i+(s/2)D), rD) \le \phi(s/2,r)$, because $t_i \in T \subset W$. Therefore,

$$\sup_{w\in W}\mathcal{N}\big(T\cap(w+sD), rD\big) \le \sup_{w\in W}\mathcal{N}\big(T\cap(w+sD), (s/2)D\big)\cdot\phi(s/2,r)$$

and the first part follows.

Turning to the second part of the claim, assume that T and W are star-shaped around 0. Let $w \in W$, set $t_1, \ldots, t_m$ to be a maximal $s/2$-separated subset of $T \cap (w + sD)$ with respect to the $L_2(\mu)$ norm and put $y_i = (r/s)t_i$. Since T is star-shaped around 0, $y_i \in T$ and $(y_i)_{i=1}^m$ is an $r/2$-separated subset of $(r/s)w + rD$. For the same reason, $(r/s)w \in W$ and

$$\mathcal{M}\big(T \cap (w + sD), (s/2)D\big) \leq \sup_{v \in W} \mathcal{M}\big(T \cap (v + rD), (r/2)D\big).$$

Using the standard connection between packing numbers and covering numbers and taking the supremum over w,

$$\begin{aligned}\phi(s, s/2) &= \sup_{w \in W} \mathcal{N}\big(T \cap (w + sD), (s/2)D\big) \\ &\leq \sup_{w \in W} \mathcal{M}\big(T \cap (w + sD), (s/2)D\big) \\ &\leq \sup_{w \in W} \mathcal{M}\big(T \cap (w + rD), (r/2)D\big).\end{aligned}$$

Iterating the first part of the lemma,

$$\begin{aligned}\log \phi(s, r) &\leq c_0 \log_2(2s/r) \cdot \sup_{w \in W} \log \mathcal{M}\big(T \cap (w + 4rD), 2rD\big) \\ &\leq c_0 \log_2(2s/r) \cdot \sup_{w \in W} \log \mathcal{N}\big(T \cap (w + 4rD), rD\big) \\ &= c_0 \log_2(2s/r) \cdot \log \phi(4r, r),\end{aligned}$$

as claimed. □

Before we turn to the proof of the upper bound, let us revisit the complexity parameters in question. Since F is a convex class, $F - f$ is star-shaped around 0; hence, if $s > r$

$$\mathcal{M}\big((F - f) \cap 4sD, (s/2)D\big) \leq \mathcal{M}\big((F - f) \cap 4rD, (r/2)D\big).$$

In particular, if $\lambda_M(\eta_1, f) < r$ then

$$\log \mathcal{M}\big((F - f) \cap 4sD, (s/2)D\big) \leq \eta_1^2 N r^2 \leq \eta_1^2 N s^2.$$

Hence, if $r < \lambda_M(\eta_1, f)$ then $\log \mathcal{M}((F - f) \cap 4rD, (r/2)D) \geq \eta_1^2 N r^2$, while if $r > \lambda_M(\eta_1, f)$, the reverse inequality holds.

A similar assertion holds for λ_Q, r_M and r_Q. The rather standard proof of those facts, which is almost identical to the argument outlined above, is omitted.

4. The upper bound. The path we will take in proving the upper bound is as follows:

- Choose a "correct" level r using the parameters λ_M and λ_Q for well-chosen constants η_1 and η_2.
- Replace F by V, a maximal r-separated subset of F with respect to the $L_2(\mu)$ norm, and study ERM in V. To that end, recall that $f^* = \operatorname{argmin}_{f\in F} \|f - Y\|_{L_2}$, set $v_0 = \operatorname{argmin}_{v\in V} \|v - Y\|_{L_2}$ and observe that by the orthogonality of W to $\operatorname{span}(F)$, v_0 is the nearest point in V to f^*; thus, $\|v_0 - f^*\|_{L_2} \le r$ and for every $v \in V$,
$$\big|\mathbb{E}\big(v_0(X) - Y\big)(v - v_0)(X)\big| = \big|\mathbb{E}\big(v_0 - f^*\big)(v - v_0)(X)\big| \le r\|v - v_0\|_{L_2}.$$
It follows that the empirical excess risk relative to V satisfies
$$\begin{aligned} P_N\mathcal{L}_v^V \ge{}& \frac{1}{N}\sum_{i=1}^N (v - v_0)^2(X_i) \\ &- 2\left|\frac{1}{N}\sum_{i=1}^N \big(v_0(X_i) - Y_i\big)(v - v_0)(X_i) - \mathbb{E}\big(v_0(X) - Y\big)(v - v_0)(X)\right| \\ &- 2r\|v - v_0\|_{L_2}. \end{aligned}$$
- Next, one may study the corresponding quadratic and multiplier processes indexed by appropriate subsets of V and show that with high probability, if $\|v - v_0\|_{L_2} \ge c_1 r$ then $P_N\mathcal{L}_v^V > 0$. Thus, ERM performed in V produces $\hat{v}$ for which $\|\hat{v} - v_0\|_{L_2} \le c_1 r$.
- Since $\|v_0 - f^*\|_{L_2} \le r$ one has that on the same event $\|\hat{v} - f^*\|_{L_2} \le c_2 r$, and using the orthogonality of W to $\operatorname{span}(F)$ once again,
$$\mathbb{E}\big(\mathcal{L}_{\hat{v}}^F \,|\, (X_i, Y_i)_{i=1}^N\big) \le c_3 r^2,$$
as required.

Let $F \subset L_2(\mu)$ be a compact, convex class of functions. Fix $r > 0$ that will be named later and let V to be a maximal r-separated subset of F. Note that for every $v_0 \in V$, $F_{v_0} = F - v_0$ is star-shaped around 0, and $\operatorname{star}(V - v_0) \subset F - v_0$. Using the notation of Lemma 3.1, let $T = W = F_{v_0}$, and for $s > 2r > 0$,
$$\begin{aligned} &\log \mathcal{N}\big(\big(\operatorname{star}(V - v_0)\big) \cap sD, rD\big) \\ &\quad\le \log \mathcal{N}(F_{v_0} \cap sD, rD) \\ &\quad\le \sup_{x\in F} \log \mathcal{N}\big(F_{v_0} \cap (x - v_0 + sD), rD\big) \\ &\quad\le c_0 \log(s/r) \sup_{x\in F} \log \mathcal{N}\big(F_{v_0} \cap (x - v_0 + 4rD), rD\big) \\ &\quad= c_0 \log(s/r) \sup_{x\in F} \log \mathcal{N}\big(F \cap (x + 4rD), rD\big). \end{aligned}$$

Observe that $F \cap (x + 4rD) \subset ((F - x) \cap 4rD) + x$, implying that

$$\begin{aligned}(4.1)\qquad &\log \mathcal{N}\big((\operatorname{star}(V - v_0)) \cap sD, rD\big) \\ &\quad \le c_0 \log(s/r) \cdot \sup_{x \in F} \log \mathcal{N}\big((F - x) \cap 4rD, rD\big).\end{aligned}$$

Clearly, the same estimate holds for $(V - v_0) \cap sD$, and since $V - v_0$ is r-separated,

$$\begin{aligned}(4.2)\qquad \log\big|(V - v_0) \cap sD\big| &= \log \mathcal{M}\big((V - v_0) \cap sD, rD\big) \\ &\le \log \mathcal{N}\big((V - v_0) \cap sD, (r/2)D\big) \\ &\le \log \mathcal{N}\big(F_{v_0} \cap sD, (r/2)D\big) \\ &\le c_0 \log(s/r) \cdot \sup_{x \in F} \log \mathcal{N}\big((F - x) \cap 4rD, (r/2)D\big) \\ &\le c_0 \log(s/r) \cdot \sup_{x \in F} \log \mathcal{M}\big((F - x) \cap 4rD, (r/2)D\big).\end{aligned}$$

With that in mind, fix constants η_1, η_2, κ_2 and κ_3 that will be specified later, and for that choice of constants, let $r > 0$ for which

$$\sup_{x \in F} \log \mathcal{M}\big((F - x) \cap 4rD, (r/2)D\big) \le \max\big\{\eta_1^2 N r^2, \eta_2^2 N\big\},$$

and $r \ge r_Q(\kappa_2) \exp(-\kappa_3 \exp(N))$; that is,

$$(4.3)\qquad r \ge \max\big\{\lambda_M(\eta_1), \lambda_Q(\eta_2), r_Q(\kappa_2) \exp\big(-\kappa_3 \exp(N)\big)\big\}.$$

Following the path outlined earlier, the idea is to study ERM in V, given the data $(X_i, Y_i)_{i=1}^N$ generated by $Y = f(X) + W$ for W that is orthogonal to $\operatorname{span}(F)$. To that end, one must control the multiplier and quadratic components in the decomposition of the squared loss relative to V.

Let us begin with the analysis of the multiplier component.

LEMMA 4.1. *Fix $0 < \theta < 1$, $L > 1$ and $q > 2$. There exist constants c_0, c_1 and c_2 that depend only on L and q, and for which the following holds. Let F be a convex, L-sub-Gaussian class, set $\xi \in L_q$ for some $q > 2$ and put $\eta_1 = c_0\theta/\|\xi\|_{L_q}$. If r is as in (4.3), then for every $v_0 \in V$, with probability at least*

$$1 - c_1 \frac{\log^q N}{N^{(q/2)-1}} - 2\exp\big(-c_2 \eta_1^2 r^2 N\big),$$

$$\sup_{\{v \in V : \|v - v_0\|_{L_2} \ge 2r\}} \left| \frac{1}{N} \sum_{i=1}^N \xi_i \frac{v - v_0}{\|v - v_0\|_{L_2}^2}(X_i) - \mathbb{E}\xi \frac{v - v_0}{\|v - v_0\|_{L_2}^2} \right| \le \theta.$$

The proof of Lemma 4.1 is based on the following fact from [12].

THEOREM 4.2. *For $L > 1$ and $q > 2$, there exist constants c_0, c_1 and c_2 that depend only on L and q for which the following holds. Let $\xi \in L_q$, set H to be an L-sub-Gaussian class and denote by $d_H = \sup_{h\in H} \|h\|_{L_2}$. For $w, u \geq 8$, with probability at least*

$$1 - c_0 w^{-q} \frac{\log^q N}{N^{(q/2)-1}} - 2\exp\left(-c_1 u^2 \left(\frac{\mathbb{E}\|G\|_H}{L d_H}\right)^2\right),$$

$$\sup_{h\in H}\left|\frac{1}{N}\sum_{i=1}^N \xi_i h(X_i) - \mathbb{E}\xi h\right| \leq c_2 L w u \|\xi\|_{L_q} \frac{\mathbb{E}\|G\|_H}{\sqrt{N}}.$$

PROOF OF LEMMA 4.1. The proof consists of two parts: first, controlling the process indexed by $\{f \in F : \|f - v_0\|_{L_2} \geq s\}$ for $s = (3/2) r_M(\eta_1, v_0)$, and then treating the process indexed by $\{v \in V : r \leq \|v - v_0\|_{L_2} \leq s\}$. Clearly, without loss of generality one may assume that $r \leq r_M(\eta_1, v_0)$. By the regularity of r_M and since $s > r_M(\eta_1, v_0)$, it is evident that $\mathbb{E}\|G\|_{(F-v_0)\cap sD} \leq \eta_1 \sqrt{N} s^2$. Also, $(F - v_0) \cap (s/4)D \subset (F - v_0) \cap sD$, and since $s/4 \leq r_M(\eta_1, v_0)$, one has $\mathbb{E}\|G\|_{(F-v_0)\cap sD} \geq \eta_1 \sqrt{N} s^2/16$.

Therefore, applying Theorem 4.2 to the set $H = (F - v_0) \cap sD$, there are constants c_1, c_2 and c_3 that depend only on q and L for which, with probability at least $1 - c_1 N^{-((q/2)-1)} \log^q N - 2\exp(-c_2 \eta_1^2 s^2 N)$, if $f \in F$ and $\|f - v_0\|_{L_2} \leq s$,

$$\left|\frac{1}{N}\sum_{i=1}^N \xi_i (f - v_0)(X_i) - \mathbb{E}\xi(f - v_0)\right| \leq c_3 L \|\xi\|_{L_q} \eta_1 s^2 = (*).$$

Observe that $(*) \leq \theta s^2$ if $\eta_1 \leq \theta / c_3 L \|\xi\|_{L_q}$. For such a choice of η_1, if $\|f - v_0\|_{L_2} = s$ then

$$\left|\frac{1}{N}\sum_{i=1}^N \xi_i (f - v_0)(X_i) - \mathbb{E}\xi(f - v_0)\right| \leq \theta \|f - v_0\|_{L_2}^2, \tag{4.4}$$

and since $F - v_0$ is star-shaped around 0, (4.4) holds on the same event for every $f \in F$ for which $\|f - v_0\| \geq s$.

Next, one has to control the process indexed by $\{v \in V : r \leq \|v - v_0\|_{L_2} < s\}$. Set $j_0 = \lceil s/r \rceil$, fix $s_j = 2^j r$ for $0 \leq j \leq j_0$ and let $V_j = \mathrm{star}((V - v_0) \cap s_j D)$. By Theorem 4.2, on an event $\mathcal{A}_j$, for every $h \in V_j$,

$$\left|\frac{1}{N}\sum_{i=1}^N \xi_i h(X_i) - \mathbb{E}\xi h\right| \leq c_4(L, q) w_j u_j \|\xi\|_{L_q} \frac{\mathbb{E}\|G\|_{V_j}}{\sqrt{N}} = (**)_j.$$

The aim it to ensure that $(**)_j \leq \theta s_j^2/4$ and that $\mathcal{A}_j$ is of high enough probability. Indeed, on $\mathcal{A}_j$, if $v \in V$ and $s_j/2 \leq \|v - v_0\|_{L_2} \leq s_j$,

$$\left|\frac{1}{N}\sum_{i=1}^N \xi_i h(X_i) - \mathbb{E}\xi h\right| \leq \theta \|v - v_0\|_{L_2}^2.$$

To that end, let $w_j = \sqrt{j}$, recall that $d_V = \sup_{v\in V} \|v\|_{L_2}$, and thus $d_{V_j} = s_j = r2^j$. Put

$$u_j = \max\left\{8, \frac{\theta\sqrt{N}r}{4c_4\|\xi\|_{L_q}} \cdot \frac{2^j}{\sqrt{j}} \cdot \frac{d_{V_j}}{\mathbb{E}\|G\|_{V_j}}\right\}$$

and consider two cases: first, if $u_j > 8$ then $(**)_j \le \theta s_j^2/4$ and

$$\Pr(\mathcal{A}_j) \ge 1 - c_5\frac{\log^q N}{j^{q/2}N^{(q/2)-1}} - 2\exp\left(-c_6(q,L)\frac{\theta^2}{\|\xi\|_{L_q}^2} \cdot Nr^2 \cdot \frac{2^{2j}}{j}\right).$$

Alternatively, observe that if $u_j = 8$, then

$$u_j^2\left(\frac{\mathbb{E}\|G\|_{V_j}}{d_{V_j}}\right)^2 \ge c_7(q,L) \cdot \frac{\theta^2}{\|\xi\|_{L_q}^2} \cdot Nr^2 \cdot \frac{2^{2j}}{j}.$$

By (4.2), V_j has at most $|(V - v_0) \cap s_j D|$ extreme points, and as noted previously,

$$\begin{aligned}\log\left|(V - v_0) \cap s_j D\right| &\le c_8 \log(s_j/r) \cdot \log \mathcal{M}\left(F_{v_0} \cap 4rD, (r/2)D\right)\\ &\le c_8 \log(s_j/r) \cdot (\eta_1\sqrt{N}r)^2.\end{aligned}$$

Thus, applying standard properties of Gaussian processes

$$\begin{aligned}\mathbb{E}\|G\|_{V_j} &\le c_9 d_{V_j} \cdot \log^{1/2}\left|(V - v_0) \cap s_j D\right| \le c_{10}s_j \log^{1/2}\left(\frac{2s_j}{r}\right) \cdot \eta_1\sqrt{N}r\\ &= c_{10}\eta_1\sqrt{N}\sqrt{\frac{j}{2^j}}s_j^2,\end{aligned}$$

and in particular there are constants c_{11} and c_{12} that depend only on q and L for which

$$\sup_{h\in V_j}\left|\frac{1}{N}\sum_{i=1}^N \xi_i h(X_i) - \mathbb{E}\xi h\right| \le c_{11}\frac{\|\xi\|_{L_q}}{\sqrt{j}} \cdot \eta_1\sqrt{\frac{j}{2^j}}s_j^2 \le \theta s_j^2/4$$

provided that $\eta_1 \le c_{12}\theta/\|\xi\|_{L_q}$.

It follows that in both cases, there are constants c_{13} and c_{14} that depend only on q and L, and with probability at least

$$1 - c_{13}\frac{\log^q N}{j^{q/2}N^{(q/2)-1}} - 2\exp(-c_{14}\eta_1^2 \cdot Nr^2 \cdot 2^{2j}/j),$$

$$\sup_{h\in V_j}\left|\frac{1}{N}\sum_{i=1}^N \xi_i h(X_i) - \mathbb{E}\xi h\right| \le \theta s_j^2/4.$$

One may conclude the proof by applying the union bound to this estimate for $0 \le j \le j_0$. $\square$

Next, let us turn to the infimum of the quadratic process

$$\inf_{\{v\in V:\|v-v_0\|_{L_2}\geq cr\}} \frac{1}{N}\sum_{i=1}^{N}\Big(\frac{(v-v_0)}{\|v-v_0\|_{L_2}}\Big)^2(X_i), \tag{4.5}$$

where r was selected in (4.3) for a well-chosen η_2 and a suitable constant c.

LEMMA 4.3. *For every $L>1$ there exist constants c_0, c_1 and c_2 that depend only on L for which the following holds. For every $v_0\in V$, with probability at least $1-2\exp(-c_0N)$, if $v\in V$ and $\|v-v_0\|_{L_2}\geq c_1r$ then*

$$\frac{1}{N}\sum_{i=1}^{N}(v-v_0)^2(X_i)\geq c_2\|v-v_0\|_{L_2}^2.$$

The proof of Lemma 4.3 is similar to the one used in the analysis of the multiplier component: controlling relatively "large distances" in F, that is, when $f\in F$ for which $\|f-v_0\|_{L_2}\geq (3/2)r_Q(\eta_2)\equiv s$; and then "small distances" in V, that is, $v\in V$ for which $r\leq\|v-v_0\|_{L_2}\leq s$ [again, one may assume that $r<r_Q(\eta_2)$].

For the constant η_2 (yet to be specified), one has:

- $\mathbb{E}\|G\|_{(F-v_0)\cap sD}\leq\eta_2\sqrt{N}s$,
- for every $2r<t<s$,

$$\log\mathcal{N}\big(\big(\mathrm{star}(V-v_0)\big)\cap sD, tD\big)\leq c_0\log(2s/t)\cdot\eta_2^2N$$

and

$$\log\big|\big(\mathrm{star}(V-v_0)\big)\cap sD\big|\leq c_0\log(2s/r)\cdot\eta_2^2N.$$

The required lower bound on the infimum of the quadratic process (4.5) is based on estimates from [11] and [10], which will be formulated under the sub-Gaussian assumption, rather than using the original (and much weaker) small-ball condition.

THEOREM 4.4. *For every $L>1$, there are constants κ_4,κ_5 and κ_6 that depend only on L for which the following holds. Let H be an L-sub-Gaussian class that is star-shaped around zero. Set $H_\rho=H\cap\rho D$ and fix ρ for which*

$$\mathbb{E}\|G\|_{H_\rho}\leq\kappa_4\sqrt{N}\rho.$$

Then, with probability at least $1-2\exp(-\kappa_5N)$,

$$\inf_{\{h\in H:\|h\|_{L_2}\geq\rho\}}\frac{1}{N}\sum_{i=1}^{N}\Big(\frac{h(X_i)}{\|h\|_{L_2}}\Big)^2\geq\kappa_6.$$

We will apply Theorem 4.4 to the class $H=(F-v_0)\cap sD$ (large distances) and then to $V_j=\mathrm{star}((V-v_0)\cap s_jD)$ for $s_j=2^jr$ (small distances).

LEMMA 4.5. *There exist absolute constants c_0 and c_1 for which the following holds. For every $s > \rho \geq c_0 r$,*

$$\mathbb{E}\|G\|_{V_j \cap \rho D} \leq c_1 \eta_2 \sqrt{N}\big(\rho \log^{3/2}(2s_j/\rho) + r \log^{3/2}(2s/r)\big).$$

In particular, setting $\rho = s_j/2$ for $\eta_2 = c_2\kappa_4$, one has

$$\mathbb{E}\|G\|_{V_j \cap (s_j/2)D} \leq \kappa_4 \sqrt{N}(s_j/2).$$

PROOF. Fix $\rho < s_j$ and note that by Dudley's entropy integral bound (see, e.g., [7, 16]),

$$\begin{aligned}\mathbb{E}\|G\|_{V_j \cap \rho D} &\leq c_1 \int_0^{\rho} \log^{1/2} \mathcal{N}(V_j \cap \rho D, t D)\, dt \\ &= c_1 \int_0^{r} \log^{1/2} \mathcal{N}(V_j \cap \rho D, t D)\, dt \\ &\quad + c_1 \int_r^{\rho} \log^{1/2} \mathcal{N}(V_j \cap \rho D, t D)\, dt.\end{aligned}$$

By (4.1), and since

$$V_j = \operatorname{star}\big((V - v_0) \cap s_j D\big) \subset \big(\operatorname{star}(V - v_0)\big) \cap s_j D,$$

it follows that for $r < t < \rho$,

$$\begin{aligned}\log \mathcal{N}(V_j \cap \rho D, r D) &\leq \log \mathcal{N}\big(\big(\operatorname{star}(V - v_0)\big) \cap \rho D, r D\big) \\ &\leq c_2 \log(2\rho/r) \cdot \sup_{x \in F} \log \mathcal{N}\big((F - x) \cap 4r D, r D\big) \\ &\leq c_2 \log(2\rho/r) \cdot \eta_2^2 N.\end{aligned}$$

Turning to the case of $0 < t < r$, observe that by (4.2),

$$\log\big|(V - v_0) \cap s_j D\big| \leq c_2 \log(2s_j/r) \cdot \eta_2^2 N = (*)$$

and V_j is the union of at most $\exp(*)$ "intervals" of the from $[0, v - v_0]$. Hence, for $0 < t < r$,

$$\log \mathcal{N}(V_j \cap \rho D, t D) \leq c_2\big(\eta_2^2 N \log(2s_j/r) + \log(2\rho/t)\big).$$

Now the first part of the claim follows from integration, and the second part is an immediate outcome of the first. □

PROOF OF LEMMA 4.3. Combining Theorem 4.4 and Lemma 4.5 for $\eta_2 = c_0\kappa_4$, it follows that with probability at least $1 - 2\exp(-\kappa_5 N)$, if $v \in V$ and $s_j/2 \leq \|v - v_0\|_{L_2} \leq s_j$,

$$\frac{1}{N}\sum_{i=1}^{N} (v - v_0)^2(X_i) \geq \kappa_6 \|v - v_0\|_{L_2}^2. \tag{4.6}$$

Repeating this argument for $s_j = 2^j r$ and then applying it to the set $F_{v_0} \cap sD$ for $s = (3/2)r_Q(\eta_2)$, it is evident that if $\log_2(s/r) \le \exp(\kappa_5 N/2)$ then with probability at least $1 - 2\exp(-\kappa_5 N/2)$, (4.6) holds for every $v \in V$ that satisfies $\|v - v_0\|_{L_2} \ge c_1 r$. $\square$

With all the ingredients in place, we may now conclude the proof of the upper estimate.

Fix $f \in F$ and set $Y = f(X) + W$ for $W \in L_q$ that is orthogonal to $\operatorname{span}(F)$. Let r, V and v_0 as above, recall that for every $v \in V$,

$$\|v - Y\|_{L_2}^2 = \|W\|_{L_2}^2 + \|v - f\|_{L_2}^2, \tag{4.7}$$

and thus $\|v_0 - f\|_{L_2} \le r$. Moreover, for every $v \in V$, $\mathbb{E}(W \cdot (v - v_0)(X)) = 0$ and

$$\begin{aligned}\big|\mathbb{E}\big(v_0(X) - Y\big)(v - v_0)(X)\big| &= \big|\mathbb{E}(v_0 - f)(X) \cdot (v - v_0)(X)\big| \\ &\le \|v_0 - f\|_{L_2} \cdot \|v - v_0\|_{L_2} \le r\|v - v_0\|_{L_2}.\end{aligned}$$

By Lemma 4.3, with probability at least $1 - 2\exp(-\kappa_5 N/2)$, if $v \in V$ and $\|v - v_0\|_{L_2} \ge c(L)r$, then

$$\frac{1}{N}\sum_{i=1}^N (v - v_0)^2(X_i) \ge \kappa_6 \|v - v_0\|_{L_2}^2.$$

Using the notation of Lemma 4.1, set $\theta = \kappa_6/4$ and $\eta_1 = c_0(q, L)\theta/\|W\|_{L_q}$. Hence, there are constants c_1 and c_2 that depend only on q and L, for which, with probability at least

$$1 - c_1 \frac{\log^q N}{N^{(q/2)-1}} - 2\exp\big(-c_2 \eta_1^2 r^2 N\big),$$

for every $v \in V$, $\|v - v_0\|_{L_2} \ge 2r$,

$$\begin{aligned}&\left|\frac{1}{N}\sum_{i=1}^N \big(v_0(X_i) - Y_i\big)(v - v_0)(X_i) - \mathbb{E}\big(v_0(X) - Y\big)(v - v_0)(X)\right| \\ &\qquad \le \frac{\kappa_6}{4}\|v - v_0\|_{L_2}^2.\end{aligned}$$

On the intersection of the two events and for a constant $c_3 = c_3(q, L)$, if $\|v - v_0\|_{L_2} \ge c_3 r$ then

$$\begin{aligned}P_N \mathcal{L}_v^V &= \frac{1}{N}\sum_{i=1}^N (v - v_0)^2(X_i) + \frac{2}{N}\sum_{i=1}^N \big(v_0(X_i) - Y_i\big)(v - v_0)(X_i) \\ &\ge \frac{1}{N}\sum_{i=1}^N (v - v_0)^2(X_i) - 2\big|\mathbb{E}\big(v_0(X) - Y\big)(v - v_0)(X)\big|\end{aligned}$$

$$
\begin{aligned}
&- 2\left|\frac{1}{N}\sum_{i=1}^{N}\big(v_0(X_i)-Y_i\big)(v-v_0)(X_i) - \mathbb{E}\big(v_0(X)-Y\big)(v-v_0)(X)\right| \\
&\geq \kappa_6\|v-v_0\|_{L_2}^2 - 2r\|v-v_0\|_{L_2} - (\kappa_6/2)\|v-v_0\|_{L_2}^2 \\
&\geq (\kappa_6/4)\|v-v_0\|_{L_2}^2.
\end{aligned}
$$

Thus, for every such sample, the empirical risk minimizer $\hat{v} \in V$ satisfies that

$$\|\hat{v}-v_0\|_{L_2} \leq c_4 r.$$

And, since W is orthogonal to $\operatorname{span}(F)$,

$$
\begin{aligned}
\mathbb{E}\big(\mathcal{L}_{\hat{v}}^F \,|\, (X_i,Y_i)_{i=1}^N\big) &= \|\hat{v}-Y\|_{L_2}^2 - \|f-Y\|_{L_2}^2 \\
&= \|\hat{v}-f-W\|_{L_2}^2 - \|W\|_{L_2}^2 \\
&= \|\hat{v}-f\|_{L_2}^2 - 2\mathbb{E}\big(W\cdot(\hat{v}-f)(X)\big) \\
&\leq \big(\|\hat{v}-v_0\|_{L_2} + \|v_0-f\|_{L_2}\big)^2 \leq (1+c_4)^2 r^2.
\end{aligned}
$$

5. The lower bound. The lower estimates we present are based on a volumetric argument. The idea is that if a learning procedure is "too successful", a well-separated subset of F endows a collection of disjoint sets in $\mathbb{R}^N$ (each collection depends on $X_1,\dots,X_N$). However, because of some volumetric constraint, there is not "enough room" for such a collection to exist, leading to a contradiction.

The notions of volume are different in the two estimates: one is based on the Lebesgue measure while the other is determined by the choice of the "noise" W.

DEFINITION 5.1. Let F be a class of functions and assume that $\mathbb{X} = (x_1,\dots,x_N) \in \Omega^N$. For every $f \in F$, set

$$\mathcal{K}(f,\mathbb{X}) = \big\{h \in F : h(x_i) = f(x_i) \text{ for every } 1 \leq i \leq N\big\}.$$

The set $\mathcal{K}(f,\mathbb{X})$ is called the *version space* of F associated with f and $\mathbb{X}$.

In other words, $\mathcal{K}(f,\mathbb{X})$ consists of all the functions in F that agree with f on $\mathbb{X}$. Naturally, in the context of statistical learning theory, $\mathbb{X}$ will be a random sample $(X_i)_{i=1}^N$, selected according to the underlying measure μ.

The diameter of the version space is a reasonable candidate for a lower bound on the performance of any learning procedure: if W is mean-zero and independent of X and $(Y_i)_{i=1}^N = (f(X_i)+W_i)_{i=1}^N$, a learning procedure cannot distinguish between f and any other function in the version space associated with f and $(X_i)_{i=1}^N$. Hence, the largest typical diameter of a version space should be a lower estimate on the performance of any learning procedure, as the following well-known fact shows (see, e.g., [6]).

THEOREM 5.2. *Given a mean-zero random variable W that is independent of X, for every $f \in F$ set $Y^f = f(X) + W$. If Ψ is a learning procedure, then*

$$\sup_{f\in F} \Pr\Big(\big\| \Psi\big((Y_i^f, X_i)_{i=1}^N\big) - f \big\|_{L_2(\mu)} \geq \frac{1}{4} \operatorname{diam}\big(\mathcal{K}(f, \mathbb{X}), L_2(\mu)\big) \Big) \geq 1/2,$$

where the probability is relative to the product measure endowed on $(\Omega \times \mathbb{R})^N$ by the N-product of the joint distribution of X and W.

As noted earlier, if W is orthogonal to $\operatorname{span}(F)$, then for every $h \in F$ and every target Y^f, $\mathbb{E}\mathcal{L}_h^f = \|h - f\|_{L_2}^2$. Thus, the largest typical diameter of a version space $\mathcal{K}(f, \mathbb{X})$ is a lower bound on $\mathcal{E}_p$ for the set of admissible targets $\mathcal{Y} = \{f(X) + W : f \in F\}$.

This leads to the following question.

QUESTION 5.3. Given a class F defined on a probability space (Ω, μ), $f \in F$ and $\mathbb{X} = (x_1, \ldots, x_N) \subset \Omega^N$, find a lower estimate on $\operatorname{diam}(\mathcal{K}(f, \mathbb{X}), L_2(\mu))$.

The first result of this section deals with Question 5.3.

THEOREM 5.4. *There exists an absolute constant c for which the following holds. Let $F \subset L_2(\mu)$ be a convex and centrally-symmetric set. If*

$$\log \mathcal{M}\big(F \cap 2rD, (r/4)D\big) \geq cN,$$

then for every $\mathbb{X} = (x_1, \ldots, x_N)$, $\operatorname{diam}(\mathcal{K}(0, \mathbb{X}), L_2(\mu)) \geq r/8$. In particular, for any $f \in F$,

$$\operatorname{diam}\big(\mathcal{K}(f/2, \mathbb{X}), L_2(\mu)\big) \geq r/16.$$

Since F is star-shaped around 0, it follows that $\mathcal{M}(F \cap 4rD, (r/2)D) \leq \mathcal{M}(F \cap 2rD, (r/4)D)$. Therefore, Theorem 5.4 implies that if $\lambda_Q(c, 0) > r$ then for every $\mathbb{X} = (x_1, \ldots, x_N)$, $\operatorname{diam}(\mathcal{K}(0, \mathbb{X}), L_2(\mu)) \geq r/8$. In particular, for every $W \in L_2$ that is orthogonal to $\operatorname{span}(F)$, the best possible error rate in F that holds with probability $1/2$ and for every target $Y^f = f(X) + W$, is at least of the order of $\lambda_Q^2(c, 0) \geq c_1 \lambda_Q^2(c)$.

PROOF. Let $f_1, \ldots, f_m$ be $r/4$-separated in $F \cap 2rD$. Set

$$A_i = \frac{f_i}{2} + \frac{1}{32}(F \cap 2rD),$$

and observe that $A_i \subset F \cap 2rD$. Also, for every $h \in A_i$, $\|(f_i/2) - h\|_{L_2} \leq r/16$. Therefore, if $h_i \in A_i$ and $h_\ell \in A_\ell$, then $\|h_i - h_\ell\|_{L_2} \geq r/8$.

Fix $\mathbb{X} = (x_1, \ldots, x_N)$ and for $A \subset F$ set

$$P_{\mathbb{X}}(A) = \big\{\big(h(X_i)\big)_{i=1}^N : h \in A\big\} \subset \mathbb{R}^N,$$

the coordinate projection of A associated with $\mathbb{X}$. Note that for every $1 \le i \le m$

$$P_{\mathbb{X}}(A_i) = \frac{1}{2}\big(f_i(x_j)\big)_{j=1}^N + \frac{1}{32}P_{\mathbb{X}}(F \cap 2rD) \tag{5.1}$$

and consider two possibilities.

First, if there are $i \neq \ell$ for which $P_{\mathbb{X}}(A_i) \cap P_{\mathbb{X}}(A_\ell) \neq \varnothing$, there are $h_i \in A_i$ and $h_\ell \in A_\ell$ that satisfy $h_i - h_\ell \in \mathcal{K}(0, \mathbb{X})$, thus showing that $\operatorname{diam}(\mathcal{K}(0, \mathbb{X}), L_2(\mu)) \ge r/8$.

Otherwise, the sets $P_{\mathbb{X}}(A_i)$ are disjoint subsets of $P_{\mathbb{X}}(F \cap 2rD)$. And, setting $T = P_{\mathbb{X}}(F \cap 2rD)$, (5.1) implies that $\mathcal{M}(T, T/32) \ge m$. Since T is a convex, centrally symmetric subset of $\mathbb{R}^N$, a standard volumetric argument shows that $\mathcal{M}(T, T/32) \le \exp(cN)$ for a suitable absolute constant c. Hence, if $m > \exp(cN)$, $\operatorname{diam}(\mathcal{K}(0, \mathbb{X}), L_2(\mu)) \ge r/8$, as claimed.

The second claim follows immediately from the first: if h is of (almost) maximal $L_2(\mu)$ distance from 0 in $\mathcal{K}(0, \mathbb{X})$ then for any $f \in F$, $(h + f)/2 \in \mathcal{K}(f/2, \mathbb{X})$; thus $\operatorname{diam}(\mathcal{K}(f/2, \mathbb{X}), L_2(\mu)) \ge r/16$. □

Our final result is the "noise-dependent" lower bound. The result we shall prove holds in a slightly more general situation than was formulated: we will assume that the noise vector $U = (W_1, \dots, W_N)$ is symmetric, independent of $X_1, \dots, X_N$, but its coordinates need not be independent. For such a noise vector, a procedure Ψ performs with accuracy $\mathcal{E}$ and confidence $1-\delta$, if for every $f \in F$, with probability at least $1 - \delta$, upon receiving the data $(X_i, f(X_i) + W_i)_{i=1}^N$,

$$\big\|\Psi\big((X_i, f(X_i) + W_i)_{i=1}^N\big) - f\big\|_{L_2}^2 \le \mathcal{E}.$$

Naturally, if $(W_1, \dots, W_N)$ has i.i.d. coordinates then $\mathcal{E}$ coincides with $\mathcal{E}_p$ for the set of targets $Y = f(X) + W$.

The main assumption that will be needed in the proof of the lower bound is the following.

ASSUMPTION 5.1. Assume that there is an nonincreasing, nonnegative function $\rho : \mathbb{R}_+ \to \mathbb{R}_+$ for which, for every centrally symmetric set $A \subset \mathbb{R}^N$ and every $z \in \mathbb{R}^N$,

$$\Pr(U \in A + z) \ge \rho\big(\|z\|_{\ell_2^N}\big)\Pr(U \in A).$$

The most important example of a random vector that satisfies Assumption 5.1 is a Gaussian vector on $\mathbb{R}^N$ with covariance σI_N, and in which case, $\rho(\|z\|_{\ell_2^N}) = \exp(-\|z\|_{\ell_2^N}^2/2\sigma^2)$. But as the next lemma shows, a Gaussian vector is just one in a rather large family of measures that satisfy Assumption 5.1.

LEMMA 5.5. *Let $A \subset \mathbb{R}^N$ be centrally symmetric and set $z \in \mathbb{R}^N$. Let W be a symmetric random variable with a density $\exp(-\phi)$ and assume that $\sup_{t\in\mathbb{R}} |\phi''(t)| \le \kappa$. If ν is the corresponding product measure on $\mathbb{R}^N$, then*

$$\nu(z + A) \ge \exp\bigl(-\kappa \|z\|_{\ell_2^N}^2/2\bigr) \cdot \nu(A).$$

PROOF. Observe that ϕ is an even function and that the density of ν is $\exp(-\sum_{i=1}^N \phi(x_i))$. Therefore, if $z = (z_i)_{i=1}^N$ then

$$\nu(z + A) = \int_{z+A} \exp\left(-\sum_{i=1}^N \phi(x_i)\right) dx = \int_A \exp\left(-\sum_{i=1}^N \phi(t_i + z_i)\right) dt = (*).$$

Because ϕ has a bounded second derivative, it is evident that, for $1 \le i \le N$,

$$\phi(t_i + z_i) \le \phi(t_i) + \phi'(t_i)z_i + \kappa z_i^2/2.$$

Setting $D\Phi_t = (\phi'(t_i))_{i=1}^N$,

$$\begin{aligned}(*) &\ge \exp\bigl(-\kappa \|z\|_{\ell_2^N}^2/2\bigr) \int_A \exp\bigl(-\langle D\Phi_t, z\rangle\bigr) \exp\left(-\sum_{i=1}^N \phi(t_i)\right) dt \\ &= \exp\bigl(-\kappa \|z\|_{\ell_2^N}^2/2\bigr)\nu(A) \cdot \mathbb{E}_{\nu|A} \exp\bigl(-\langle D\Phi_t, z\rangle\bigr),\end{aligned}$$

where $\mathbb{E}_{\nu|A}$ is expectation with respect to the measure ν conditioned on A.

Since ν is a symmetric measure and A is centrally symmetric, $\nu|A$ is a symmetric measure as well, and because ϕ is an even function, $\mathbb{E}_{\nu|A}\langle D\Phi_t, z\rangle = 0$. Therefore, by Jensen's inequality,

$$\mathbb{E}_{\nu|A} \exp\bigl(-\langle D\Phi_t, z\rangle\bigr) \ge \exp(-\mathbb{E}_{\nu|A}\langle D\Phi_t, z\rangle) = 1,$$

and the claim follows. □

THEOREM 5.6. *There exists an absolute constant c for which the following holds. Let $F \subset L_2(\mu)$ be a class of functions, set U to be a symmetric random vector that satisfies Assumption 5.1 and which is independent of $(X_i)_{i=1}^N$. If the procedure Ψ performs with accuracy $\mathcal{E} \equiv r^2/36$ and confidence $7/8$, then*

$$\log \mathcal{M}\bigl(F \cap 4rD, (r/2)D\bigr) \le 2 + \log\bigl(\rho^{-1}(c\sqrt{N}r)\bigr).$$

In particular, if $(W_1, \dots, W_N)$ has independent coordinates, distributed according to a symmetric random variable W for which $\rho(\|z\|_{\ell_2^N}) = \exp(-\|z\|_{\ell_2^N}^2/2\|W\|_{L_2}^2)$, then

$$\mathcal{E}_p \ge c_1 \lambda_M^2 \left(\frac{c_2}{\|W\|_{L_2}}\right).$$

REMARK 5.7. Versions of Theorem 5.6 for U that has i.i.d. Gaussian coordinates can be established in several different ways: for example, using information theoretic tools (see Theorem 2.5 in [15], which is based on estimates on the KL-divergence), or, alternatively, by applying the Gaussian isoperimetric inequality as in [6]. The obvious downside is that these arguments rely on rather special properties of the noise, and thus do not seem to extend to the setup of Theorem 5.6.

PROOF. Recall that if $\tau = (x_i, f(x_i) + w_i)_{i=1}^N \in (\Omega \times \mathbb{R})^N$ is a sample on which Ψ performs with accuracy $\mathcal{E}$ then $\|\Psi(\tau) - f\|_{L_2}^2 \leq \mathcal{E}$.

Choose r that satisfies $(r/2)^2 = 9\mathcal{E}$ and let $(f_j)_{j=1}^m$ be a subset of $F \cap 4rD$ that is $r/2$ separated in $L_2(\mu)$. Fix $\mathbb{X} = (x_1, \ldots, x_N) \in \Omega^N$ and set $z_j = (f_j(x_i))_{i=1}^N \in \mathbb{R}^N$.

For every $1 \leq j \leq m$, put

$$A_j(\mathbb{X}) = \big\{(w_i)_{i=1}^N : \Psi\big((x_i, f_j(x_i) + w_i)_{i=1}^N\big) \in f_j + \sqrt{\mathcal{E}}D\big\} \subset \mathbb{R}^N.$$

Thus, $A_j(\mathbb{X})$ consists of all the vectors $(w_i)_{i=1}^N \in \mathbb{R}^N$, for which, upon receiving the data $(x_i, f_j(x_i) + w_i)_{i=1}^N$, Ψ selects a point whose $L_2(\mu)$ distance to f_j is at most $r/6 = \sqrt{\mathcal{E}}$.

Let ν be measure on $\mathbb{R}^N$ according to which $U = (W_1, \ldots, W_N)$ is distributed. Recall that U is independent of $X_1, \ldots, X_N$; hence, if Ψ performs with accuracy $\mathcal{E}$ and with probability at least $7/8$, then

$$\begin{aligned}\mu^N \otimes \nu\big(\big\{(x_i, w_i)_{i=1}^N : \Psi\big((x_i, f_j(x_i) + w_i)_{i=1}^N\big) \in f_j + \sqrt{\mathcal{E}}D\big\}\big)\\ = \mu^N \otimes \nu\big(\big\{(x_i, w_i)_{i=1}^N : (w_i)_{i=1}^N \in A_j(\mathbb{X})\big\}\big) \geq 7/8.\end{aligned}$$

A standard Fubini argument shows that there is an event $\mathcal{C}_j \subset \Omega^N$ of μ^N probability at least $1/2$, and for every $\mathbb{X} = (x_i)_{i=1}^N \in \mathcal{C}_j$, $\nu(A_j(\mathbb{X})) \geq 3/4$. Moreover, if $\mathbb{X} \in \mathcal{C}_j$ then by the symmetry of ν, $\nu(-A_j(\mathbb{X})) \geq 3/4$, and the centrally-symmetric set $A_j(\mathbb{X}) \cap -A_j(\mathbb{X}) \subset A_j(\mathbb{X})$ satisfies that

$$\nu\big(A_j(\mathbb{X}) \cap -A_j(\mathbb{X})\big) \geq 1/2.$$

Observe that if $\mathbb{X} \in \mathcal{C}_j \cap \mathcal{C}_\ell$, the sets $z_j + A_j(\mathbb{X})$ and $z_\ell + A_\ell(\mathbb{X})$ are disjoint, because Ψ maps $z_j + A_j(\mathbb{X})$ to an $r/6$-neighbourhood of f_j and $z_\ell + A_\ell(\mathbb{X})$ to an $r/6$-neighbourhood of f_ℓ—but $\|f_j - f_\ell\|_{L_2} \geq r/2$. Therefore,

$$\sum_{j=1}^m \mathbb{1}_{\mathcal{C}_j}(\mathbb{X})\nu\big(z_j + \big(A_j(\mathbb{X}) \cap -A_j(\mathbb{X})\big)\big) \leq 1.$$

Integrating with respect to μ^N,

$$\sum_{i=1}^m \mathbb{E}_{\mathbb{X}}\mathbb{1}_{\mathcal{C}_j}(\mathbb{X})\nu\big(z_j + \big(A_j(\mathbb{X}) \cap -A_j(\mathbb{X})\big)\big) \leq 1,$$

and all that remains is to control $\mathbb{E}_{\mathbb{X}}\mathbb{1}_{\mathcal{C}_j}(\mathbb{X})\nu(z_j+(A_j(\mathbb{X})\cap -A_j(\mathbb{X})))$ from below.

By Assumption 5.1,

$$\nu\big(z_j+\big(A_j(\mathbb{X})\cap -A_j(\mathbb{X})\big)\big)\ge \rho\big(\|z_j\|_{\ell_2^N}\big)\cdot \nu\big(A_j(\mathbb{X})\cap -A_j(\mathbb{X})\big),$$

and by Chebychev's inequality, recalling that $\|f_j\|_{L_2}\le 4r$,

$$\mu^N\big(\|z_j\|_{\ell_2^N}^2\le c_0Nr^2\big)=\mu^N\left(\sum_{i=1}^N f_j^2(X_i)\le c_0Nr^2\right)\ge 3/4$$

for an appropriate choice of an absolute constant c_0. Thus, for every $1\le j\le m$, there is an event of μ^N measure at least $1/4$ on which $\mathbb{X}\in\mathcal{C}_j$, $\nu(A_j(\mathbb{X})\cap -A_j(\mathbb{X}))\ge 3/4$, and, since ρ is nonincreasing, $\rho(\|z\|_{\ell_2^N})\ge\rho(c_0\sqrt{N}r)$. Therefore,

$$1\ge\sum_{i=1}^m \mathbb{E}_{\mathbb{X}}\mathbb{1}_{\mathcal{C}_j}(\mathbb{X})\nu\big(z_j+\big(A_j(\mathbb{X})\cap -A_j(\mathbb{X})\big)\big)\ge(3m/16)\rho(c_0\sqrt{N}r),$$

implying that

$$\log m=\log\mathcal{M}\big(F\cap 4rD,(r/2)D\big)\le\log 6+\log\big(1/\rho(c_0\sqrt{N}r)\big),$$

as claimed.

Turning to the second part of the theorem, let Ψ be a learning procedure that performs with accuracy $\mathcal{E}_p$ for every target $Y^f=f(X)+W$ for $f\in F$ and W that is independent of X. Therefore, $\mathcal{E}_p=\mathcal{E}$ and by the first part of the theorem for $\rho(t)=\exp(-t^2/2\|W\|_{L_2}^2)$ and $\sqrt{\mathcal{E}}=r/6$, one has

$$\log\mathcal{M}\big(F\cap 4rD,(r/2)D\big)\le\log 6+c_1Nr^2/\|W\|_{L_2}^2.$$

Since F is convex and centrally symmetric, if $4r\le\operatorname{diam}(F,L_2(\mu))$ then $F\cap 4D$ contains an interval of length $8r$; thus, $\log\mathcal{M}(F\cap 4rD,(r/2)D)\ge 2\log 6$, implying that

$$\log\mathcal{M}\big(F\cap 4rD,(r/2)D\big)\le 2c_1Nr^2/\|W\|_{L_2}^2.$$

Therefore, by the definition of λ_M, $\mathcal{E}_p\ge c_2\lambda_M^2(c_3/\|W\|_{L_2})$. □

The combination of Theorem 5.6 and Lemma 5.5 concludes the proof of Theorem 2.10.

SUPPLEMENTARY MATERIAL

Supplement to "'Local' vs. 'global' parameters—breaking the Gaussian complexity barrier" (DOI: 10.1214/16-AOS1510SUPP; .pdf). We prove two observations: the first shows that the setup of the Young–Barron theorem is different from the one we study here, and the other is that for $p>1$ there is a true gap between the "local" and "global" complexities of B_p^n.

REFERENCES

[1] ANTHONY, M. and BARTLETT, P. L. (1999). *Neural Network Learning*: *Theoretical Foundations*. Cambridge Univ. Press, Cambridge. MR1741038

[2] BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields* **97** 113–150. MR1240719

[3] BÜHLMANN, P. and VAN DE GEER, S. (2011). *Statistics for High-Dimensional Data*: *Methods, Theory and Applications*. Springer, Heidelberg. MR2807761

[4] DUDLEY, R. M. (1999). *Uniform Central Limit Theorems*. *Cambridge Studies in Advanced Mathematics* **63**. Cambridge Univ. Press, Cambridge. MR1720712

[5] KOLTCHINSKII, V. (2011). *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems*. *Lecture Notes in Math*. **2033**. Springer, Heidelberg. MR2829871

[6] LECUÉ, G. and MENDELSON, S. (2013). Learning subgaussian classes: Upper and minimax bounds. Technical report, CNRS, Ecole polytechnique and Technion.

[7] LEDOUX, M. and TALAGRAND, M. (1991). *Probability in Banach Spaces*: *Isoperimetry and Processes*. *Ergebnisse der Mathematik und Ihrer Grenzgebiete* (3) [*Results in Mathematics and Related Areas* (3)] **23**. Springer, Berlin. MR1102015

[8] MASSART, P. (2007). *Concentration Inequalities and Model Selection*. *Lecture Notes in Math*. **1896**. Springer, Berlin. MR2319879

[9] MENDELSON, S. (2008). Obtaining fast error rates in nonconvex situations. *J. Complexity* **24** 380–397. MR2426759

[10] MENDELSON, S. (2014). Learning without concentration for general loss functions. Preprint. Available at arXiv:1410.3192.

[11] MENDELSON, S. (2015). Learning without concentration. *J. ACM* **62** Art. 21, 25. MR3367000

[12] MENDELSON, S. (2016). Upper bounds on product and multiplier empirical processes. *Stochastic Process. Appl.* **126** 3652–3680. MR3565471

[13] MENDELSON, S. (2017). Supplement to "'Local' vs. 'global' parameters—breaking the Gaussian complexity barrier." DOI:10.1214/16-AOS1510SUPP.

[14] MENDELSON, S., PAJOR, A. and TOMCZAK-JAEGERMANN, N. (2007). Reconstruction and subgaussian operators in asymptotic geometric analysis. *Geom. Funct. Anal.* **17** 1248–1282. MR2373017

[15] TSYBAKOV, A. B. (2009). *Introduction to Nonparametric Estimation*. Springer, New York. MR2724359

[16] VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer, New York. MR1385671

[17] YANG, Y. and BARRON, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.* **27** 1564–1599. MR1742500

DEPARTMENT OF MATHEMATICS
TECHNION—ISRAEL INSTITUTE OF TECHNOLOGY
HAIFA
ISRAEL
E-MAIL: shahar@tx.technion.ac.il