Improving the Sample Complexity Using Global Data

Shahar Mendelson

***Abstract*—We study the sample complexity of proper and improper learning problems with respect to different q-loss functions. We improve the known estimates for classes which have relatively small covering numbers in empirical L_2 spaces (e.g., log-covering numbers which are polynomial with exponent $p < 2$). We present several examples of relevant classes which have a "small" fat-shattering dimension, hence fit our setup, the most important of which are kernel machines.**

***Index Terms*—Fat-shattering dimension, Glivenko–Cantelli classes, kernel machines, learning sample complexity, uniform convexity.**

I. Introduction

In this paper, we present sample complexity estimates for various learning problems with respect to different norms, under the assumption that the classes are not "too large."

The question we explore is the following: let G be a class of functions defined on a probability space (Ω, μ) where μ is an unknown probability measure and each $g \in G$ maps Ω into $[0, 1]$. Set T to be an unknown function, which is not necessarily a member of G. Let (X_i) be independent random variables distributed according to μ. Recall that a learning rule L is a map which assigns to each sample $S_n = (X_1, \ldots, X_n)$ a function $L_{S_n} \in G$. The *learning sample complexity* associated with a q-loss function, accuracy ε, and confidence δ is the first integer n_0 such that the following holds: there exists a learning rule L such that for every $n \geq n_0$ and every probability measure μ

$$\Pr\left\{\mathbb{E}_\mu|L_{S_n} - T|^q \geq \inf_{g\in G}\mathbb{E}_\mu|g - T|^q + \varepsilon\right\} \leq \delta$$

where $\mathbb{E}_\mu$ is the expectation with respect to μ.

In other words, the objective of the learning rule is to find a function in the class G which is "almost" the best approximation to T in G with respect to the $L_q(\mu)$ norm.

Estimating the sample complexity is closely related to the notion of the rate of convergence of regression problems [20]. For example, assume that $q = 2$ and that Y is an unknown random variable. It follows that solving the regression problem for Y with respect to the class G is equivalent to solving the learning problem for the function $T = \mathbb{E}(Y|X)$. In addition, the translation from the notion of sample complexity to that of the rate of convergence used in statistics is relatively standard. Our presentation will be given from the "learning-theoretic" point of view rather than from the statistical one.

Unlike previous results, originating from the work of Vapnik and Chervonenkis, in which complexity estimates were based on the covering numbers at a scale which is roughly the desired accuracy, we use global data regarding the "size" of the class to obtain complexity estimates at every scale. One example we focus on is when the log-covering numbers of the class in question are polynomial in ε^{-1} with exponent $p < 2$.

We were motivated by two methods previously used in the investigation of sample complexity [2]. The first is the standard approach which uses the Glivenko–Cantelli (GC) condition to estimate the sample complexity. By this we mean the following: let G be a class of functions defined on Ω, let T be a fixed function, set $1 \leq q < \infty$, and put $F = \{|g - T|^q | g \in G\}$. The GC sample complexity of the class F with respect to accuracy ε and confidence δ is the smallest integer n_0 such that for every $n \geq n_0$

$$\sup_\mu \mu\left\{\sup_{g\in G}|\mathbb{E}_\mu|g - T|^q - \mathbb{E}_{\mu_n}|g - T|^q| \geq \varepsilon\right\} \leq \delta$$

where μ_n is the empirical measure supported on $(X_1, \ldots, X_n)$.

Hence, if g is an "almost" minimizer of the empirical loss functional $n^{-1}\sum_{i=1}^n |g(x_i) - T(x_i)|^q$ and if the sample is "large enough" then g is an "almost" minimizer of the average distance to T with respect to the $L_q(\mu)$ norm. One can show that the learning sample complexity is bounded by the supremum of the GC sample complexities, where the supremum is taken over all possible targets T, bounded by 1. This is true even in the regression scenario, simply be setting $T = \mathbb{E}(Y|X)$ (for further details, see [2]).

Recently, it was shown [15] that if the log-covering numbers (resp., the fat-shattering dimension) of G are of the order of ε^{-p} then the GC sample complexity of F is $\Theta(\varepsilon^{-\max\{2,p\}})$ up to logarithmic factors in ε^{-1}, δ^{-1}. This implied that if $p \geq 2$, and in the case of the squared loss $\mathbb{E}|g - T|^2$, the learning sample complexity has the same rate as the GC sample complexity. Indeed, in this case, the learning sample complexity is $\Omega(\mathrm{fat}_{4\varepsilon}(G))$ [2].

It is important to emphasize that the learning sample complexity may be established by other means rather than via the GC condition. Therefore, it comes with no surprise that there are certain cases in which it is possible to improve this bound on the learning sample complexities. In [11], [12], the following case was examined; let F be the class given by

$$\{|g - T|^2 - |T - P_G T|^2 | g \in G\}$$

where $P_G T$ is a nearest point to T in G with respect to the $L_2(\mu)$ norm. In other words, F is a class which consists of "shifted"

Manuscript received March 13, 2001; revised December 14, 2001. The material in this paper was presented in part at COLT 01, Amsterdam, The Netherlands, July 2001.

The author is with the Computer Sciences Laboratory, RSISE, The Australian National University, Canberra 0200, Australia (e-mail: shahar@csl.anu.edu.au).

Communicated by G. Lugosi, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Publisher Item Identifier S 0018-9448(02)05152-0.

elements in such a way that the "best" distance with respect to $L_2(\mu)$ is now attained at 0. Assume that there is an absolute constant C such that for every $f \in F$, $\mathbb{E}_\mu f^2 \leq C\mathbb{E}_\mu f$. This simply means that it is possible to control the variance of each function in F using its expectation. In this case, the learning sample complexity with accuracy ε and confidence δ can be bounded by

$$O\left(\frac{1}{\varepsilon}\left(\mathrm{fat}_\varepsilon(G)\log^2\frac{\mathrm{fat}_\varepsilon(G)}{\varepsilon}+\log\frac{1}{\delta}\right)\right).$$

Therefore, if $\mathrm{fat}_\varepsilon(G) = O(\varepsilon^{-p})$, the learning sample complexity is bounded by (up to logarithmic factors) $O(\varepsilon^{-(1+p)})$. If $p < 1$, this estimate is better than the one obtained using the GC sample complexities.

As it turns out, the preceding assumption is not so far fetched; it is possible to show [11], [12] that there are two generic cases in which $\mathbb{E}_\mu f^2 \leq C\mathbb{E}_\mu f$. The first case is when $T \in G$, because it implies that each $f \in F$ is nonnegative. The other case is when G is convex and $q = 2$, in which case, every function in F is given by $|g-T|^2 - |T-P_GT|^2$, where P_GT is the nearest point to T in G with respect to the $L_2(\mu)$ norm. Thus, one immediate question which comes to mind is whether the same kind of a result holds in other L_q spaces. The reason for this interest in various L_q norms is not just for the sake of generalizing known results. As q increase, one obtains an approximation of the unknown function with respect to a "stronger" norm. Therefore, it is only natural to investigate the price one has to pay (in the complexity sense) for this finer approximation.

Here, we combine the ideas used in [12] and in [15] to improve the learning complexity estimates. We show that if G maps Ω into $[0, 1]$ such that

$$\sup_n \sup_{\mu_n} \log N(\varepsilon, G, L_2(\mu_n)) = O\left(\varepsilon^{-p}\right)$$

for $p < 2$, and if either $T \in G$ or if $q \geq 2$ and G is a compact and convex subset of $L_q(\mu)$, then the learning sample complexity with respect to the q-loss is $O(\varepsilon^{-(1+p/2)})$ up to logarithmic factors in ε^{-1}, δ^{-1}. Recently, it was shown in [15] that there is an absolute constant C such that

$$\sup_n \sup_{\mu_n} \log N(\varepsilon, G, L_2(\mu_n)) \leq C\mathrm{fat}_{\frac{\varepsilon}{8}}(G)\log^2\left(\frac{2\mathrm{fat}_{\frac{\varepsilon}{8}}(G)}{\varepsilon}\right) \tag{1.1}$$

therefore, if G has a polynomial fat-shattering dimension with exponent $p < 2$ one can bound the uniform covering numbers and obtain a bound which improves the one established in [12].

The idea behind our analysis is that the sample complexity of an arbitrary class F is bounded by the GC sample complexity of two classes associated with F, where the deviation in the GC condition is roughly the same as the largest variance of a class member.

Formally, if G is a class of functions, T is the unknown function (which will be referred to as the "target concept") and $1 \leq q < \infty$, then for every $g \in G$ let $\ell_q(g)$ be its q-loss function. Thus,

$$\ell_q(g) = |g-T|^q - |T-P_GT|^q$$

where P_GT is a nearest element to T in G with respect to the L_q norm. We denote by F the set of loss functions $\ell_q(g)$.

Let G be a GC class. For every $0 < \varepsilon, \delta < 1$, denote by $S_G(\varepsilon, \delta)$ the GC sample complexity of the class G associated with accuracy ε and confidence δ. Let $\mathcal{C}^q_{G,T}(\varepsilon, \delta)$ be the learning sample complexity of the class G with respect to the target T and the q-loss, for accuracy ε and confidence δ.

The assumption we have to make is that it is possible to bound the variance of every class member by an appropriate power of its expectation, which is the idea behind the following lemma. Its proof will be presented in Section IV.

Lemma 1.1: Let G be a class of functions which map Ω into $[0, 1]$, set $q \geq 2$, and let F be the q-loss class associated with G and the target concept T, which also maps Ω into $[0, 1]$. Assume that there is some constant B such that for any $f \in F$, $\mathbb{E}_\mu f^2 \leq B(\mathbb{E}_\mu f)^{2/q}$. Let $\varepsilon > 0$, $\alpha = 2 - 2/q$, set

$$H = \left\{\frac{\varepsilon^\alpha f}{\mathbb{E}_\mu f}\middle|\, f \in F, \mathbb{E}_\mu f \geq \varepsilon, \mathbb{E}_\mu f^2 \geq \varepsilon\right\} \tag{1.2}$$

and put

$$F_\varepsilon = \left\{f \in F \,|\mathbb{E}_\mu f^2 < \varepsilon\right\}$$

and

$$H_\varepsilon = \left\{h \in H \,|\mathbb{E}_\mu h^2 < B\varepsilon^\alpha\right\}.$$

Then, for every $0 < \varepsilon, \delta < 1$

$$\mathcal{C}^q_{G,T}(\varepsilon, \delta) \leq \max\left\{S_{F_\varepsilon}\left(\frac{\varepsilon}{2}, \frac{\delta}{2}\right), S_{H_\varepsilon}\left(\frac{\varepsilon^\alpha}{2}, \frac{\delta}{2}\right)\right\}.$$

Thus, the learning sample complexity of G at scale ε may be determined by the GC sample complexity of the classes F_ε and H_ε, at a scale which is proportional to the largest variance of a member of F_ε (resp., H_ε), and this holds provided that F consists of functions for which $\mathbb{E}_\mu f^2$ may be bounded by $B(\mathbb{E}_\mu f)^{2/q}$ for some constant B.

This key lemma dictates the structure of this paper. In the second section, we investigate the GC condition for classes F which have "small" log-covering numbers, and we focus on the case where the deviation in the GC condition is of the same order of magnitude as $\sup_{f\in F}\mathbb{E}_\mu f^2$. The proof is based on estimates on the Rademacher averages (defined later) associated with the class. Next, we explore sufficient conditions which imply that if F is the q-loss function associated with a convex class G, then $\mathbb{E}_\mu f^2$ may be controlled by powers of $\mathbb{E}_\mu f$. We use a geometric approach to prove that if $q \geq 2$, there indeed is some constant B, such that for every q-loss function f, $\mathbb{E}_\mu f^2 \leq B(\mathbb{E}_\mu f)^{2/q}$. Then, we present results regarding the learning sample complexity, which is investigated in the fourth section. The final sections are devoted to examples of interesting classes for which our results apply. Among the examples we present are estimates on the learning sample complexity of convex hulls of Vapnik–Chervonenkis (VC) classes, classes of sufficiently smooth functions, and kernel machines. In fact, we present new bounds on the fat-shattering dimension of the latter. We demonstrate that in some sense, the fat-shattering dimension can be controlled by the rate of decay of the eigenvalues of an

integral operator associated with kernel, and improve the covering numbers estimates established in [21].

It is important to mention that throughout this paper we are only interested in the rates by which the sample complexity changes and its relations to the covering numbers. Though it is also possible to derive bounds on the constants which appear in the estimates, we have made no such attempt, nor do we claim that the constants could not be improved by some other method of proof. We do believe, however, that rate-wise, our results are optimal, though this is something we leave for future research.

Next, we turn to some definitions, notation, and basic observations we shall use throughout this paper.

Given a real Banach space X, let $B(X)$ be the unit ball of X. If $B \subset X$ is a ball, set $\operatorname{int}(B)$ to be the interior of B and ∂B is the boundary of B. The *dual* of X, denoted by X^*, consists of all the bounded linear functionals on X, endowed with the norm $\|x^*\| = \sup_{\|x\|=1} |x^*(x)|$. ℓ_2^n is a real n-dimensional inner product space, which will always be identified with $\mathbb{R}^n$ with respect to the Euclidean norm. ℓ_2 is the space of all the real sequences $(x_i)_{i=1}^{\infty}$ such that $\sum_{i=1}^{\infty} x_i^2 < \infty$, endowed with the inner product

$$\langle x, y\rangle = \sum_{i=1}^{\infty} x_i y_i.$$

For any $x, y \in X$, the interval $[x, y]$ is defined by

$$[x, y] = \{tx + (1-t)y | 0 \le t \le 1\}.$$

If μ is a probability measure on a measurable space (Ω, Σ), let $\mathbb{E}_\mu$ be the expectation with respect to μ. $L_q(\mu)$ is the set of functions which satisfy $\mathbb{E}_\mu |f|^q < \infty$ and set $\|f\|_q = (\mathbb{E}|f|^q)^{1/q}$. $L_\infty(\Omega)$ is the space of bounded functions on Ω, with respect to the norm $\|f\|_\infty = \sup_{\omega\in\Omega} |f(\omega)|$. We denote by μ_n an empirical measure supported on a set of n points, hence, $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{\omega_i}$, where δ_{ω_i} is the point evaluation functional on $\{\omega_i\}$. If F is a class of functions and g is any function, let $|F - g|^q = \{|f - g|^q | f \in F\}$. In general, given sets A and B, let $A + B = \{a + b | a \in A, b \in B\}$.

If (X, d) is a metric space, $Y \subset X$ and $x \in X$, the distance of x to Y is defined as $d(x, Y) = \inf_{y\in Y} d(x, y)$. A set A is called symmetric if the fact that $x \in A$ implies that $-x \in A$. The symmetric convex hull of A, denoted by $\operatorname{absconv}(A)$, is the convex hull of $A \cup -A$.

If (X, d) is a metric space, set $B(x, r)$ to be the closed ball centered at x with radius r. Recall that if $F \subset X$, the ε-covering number of F, denoted by $N(\varepsilon, F, d)$, is the minimal number of open balls with radius $\varepsilon > 0$ (with respect to the metric d) needed to cover F. A set $A \subset X$ is said to be an ε-cover of F if the union of open balls $\bigcup_{a\in A} B(a, \varepsilon)$ contains F. In cases where the metric d is clear, we denote the covering numbers of F by $N(\varepsilon, F)$. The logarithm of the covering numbers of a set is sometimes referred to as the *metric entropy* of the set.

A set is called ε-separated if the distance between any two elements of the set is larger than ε. Let $D(\varepsilon, F)$ be the maximal cardinality of an ε-separated set in F. $D(\varepsilon, F)$ are called the packing numbers of F (with respect to the fixed metric d). The packing numbers are closely related to the covering numbers, since $N(\varepsilon, F) \le D(\varepsilon, F) \le N(\varepsilon/2, F)$.

It is possible to show that the covering numbers of the q-loss class F are essentially the same as those of G.

Lemma 1.2: Let $G \subset B(L_\infty(\Omega))$ and set F to be the q-loss class associated with G. Then, for any probability measure μ and every $\varepsilon > 0$

$$\log N(\varepsilon, F, L_2(\mu)) \le \log N(\varepsilon/q, G, L_2(\mu)).$$

Proof: For every target function T, $|T - P_G T|^q$ is a fixed function, thus, the covering numbers of F are determined by the covering numbers of $H = \{|g - T|^q | g \in G\}$.

First, assume that $q > 1$. By Lagrange's theorem for $v(x) = |x|^q$ and $x_1, x_2 \in [-1, 1]$, it follows that

$$||x_1|^q - |x_2|^q| \le q|x_1 - x_2|.$$

Hence, for every $g' : \Omega \to [0, 1]$ and any $\omega \in \Omega$

$$||g(\omega) - T(\omega)|^q - |g'(\omega) - T(\omega)|^q| \le q|g(\omega) - g'(\omega)|. \quad (1.3)$$

Let G' be an ε-cover of G with respect to the $L_2(\mu)$ norm. Clearly, we may assume that every $g' \in G'$ maps Ω into $[0, 1]$, which, combined with (1.3), implies that $|G' - T|^q$ is an $q\varepsilon$-cover of H with respect to the $L_2(\mu)$ norm, as claimed.

The case $q = 1$ may be derived using a similar argument, but the triangle inequality will replace Lagrange's theorem. □

Two combinatorial parameters used in Learning Theory are the VC dimension and the fat-shattering dimension [2].

Definition 1.3: Let F be a class of $\{0, 1\}$-valued functions on a space Ω. We say that F shatters $\{\omega_1, \ldots, \omega_n\} \subset \Omega$, if for every $I \subset \{1, \ldots, n\}$ there is a function $f_I \in F$ for which $f_I(\omega_i) = 1$ if $i \in I$ and $f_I(\omega_i) = 0$ if $i \notin I$. Let

$$\mathrm{VC}(F, \Omega) = \sup\{|A| \,|\, A \subset \Omega,\ A \text{ is shattered by } F\}.$$

$\mathrm{VC}(F, \Omega)$ is called the VC dimension of F, and we shall sometimes denote it by $\mathrm{VC}(F)$.

It is possible to use a parametric version of the VC dimension, called the fat-shattering dimension.

Definition 1.4: For every $\varepsilon > 0$, a set $A = \{\omega_1, \ldots, \omega_n\} \subset \Omega$ is said to be ε-shattered by F if there is some function $s : A \to \mathbb{R}$, such that for every $I \subset \{1, \ldots, n\}$ there is some $f_I \in F$ for which $f_I(\omega_i) \ge s(\omega_i) + \varepsilon$ if $i \in I$, and $f_I(\omega_i) \le s(\omega_i) - \varepsilon$ if $i \notin I$. Let

$$\mathrm{fat}_\varepsilon(F, \Omega) = \sup\{|A| \,|\, A \subset \Omega,\ A \text{ is } \varepsilon\text{-shattered by } F\}.$$

f_I is called the shattering function of the set I and the set $(s_i) = \{s(\omega_i) | \omega_i \in A\}$ is called a witness to the ε-shattering. In cases where the set Ω is clear, we shall denote the fat-shattering dimension by $\mathrm{fat}_\varepsilon(F)$.

The important property of the VC dimension and the fat-shattering dimension is that given any probability measure, it is possible to estimate the $L_2(\mu)$ covering numbers of a given class using those parameters, as presented in the next result. The first part of the result is due to Haussler [20], while the second was established in [15].

Theorem 1.5: Let $F \subset B(L_\infty(\Omega))$.

1) If F is $\{0, 1\}$-valued and $\mathrm{VC}(F) = d$, then there is an absolute constant C such that for every probability measure μ on Ω and every $\varepsilon > 0$

$$N(\varepsilon, F, L_2(\mu)) \leq C\, d(4e)^d \left(\frac{1}{\varepsilon}\right)^{2d}.$$

2) If for every $\varepsilon > 0$ F has a finite fat-shattering dimension, then there is some absolute constant C such that for every probability measure μ

$$\log N(\varepsilon, F, L_2(\mu)) \leq C\mathrm{fat}_{\frac{\varepsilon}{32}}(F) \log^2\left(\frac{2\mathrm{fat}_{\frac{\varepsilon}{32}}(F)}{\varepsilon}\right).$$

Next, we define the Rademacher averages of a given class of functions, which is the main tool we use in the analysis of GC classes.

Definition 1.6: Let F be a class of functions and let μ be a probability measure on Ω. Set

$$\overline{R}_{n,\mu} = \frac{1}{\sqrt{n}}\, \mathbb{E}_\mu \mathbb{E}_\varepsilon \sup_{f\in F}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|$$

where ε_i are independent Rademacher random variables (that is, symmetric, $\{-1, 1\}$ valued) and (X_i) are independent, distributed according to μ.

The Rademacher averages play an important role in the theory of empirical processes because they can be used to control the deviation of the empirical means from the actual ones. As an example, we will mention the following symmetrization result, due to Giné and Zinn [20], who proved that

$$\mathbb{E}_\mu\left[\sup_{f\in F}\sum_{i=1}^n (f(X_i) - \mathbb{E}_\mu f)\right] \leq 2\mathbb{E}_\mu \mathbb{E}_\varepsilon\left[\sup_{f\in F}\sum_{i=1}^n \varepsilon_i f(X_i)\right].$$

It is possible to estimate $\overline{R}_{n,\mu}$ of a given class using its $L_2(\mu_n)$ covering numbers. The fundamental result behind this estimate is the following theorem which is due to Dudley for Gaussian processes. This was extended to the more general setting of subgaussian processes [20]. We shall formulate it only for Rademacher processes.

Theorem 1.7: There is an absolute constant C such that for every sample $(X_1, \ldots, X_n)$

$$\frac{1}{\sqrt{n}}\, \mathbb{E}_\varepsilon \sup_{f\in F}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right| \leq C\int_0^\delta \log^{\frac{1}{2}} N(\varepsilon, F, L_2(\mu_n))\, d\varepsilon$$

where

$$\delta = \sup_{f\in F}\left(\frac{1}{n}\sum_{i=1}^n f^2(X_i)\right)^{\frac{1}{2}}$$

and μ_n is the empirical measure supported on the given sample $(X_1, \ldots, X_n)$.

Finally, throughout this paper, all absolute constants are assumed to be positive and are denoted by C or c. C_p denotes a constant which depends only on p. The values of constants may change from line to line or even within the same line.

II. Glivenko–Cantelli (GC) Estimates

The main tool we use in the analysis of the GC sample complexity is an exponential inequality which is due to Talagrand [19]. This result enables one to control the GC sample complexity using two parameters. The first one is the nth Rademacher average associated with the class. The second is the "largest" variance of a member of the class. As explained in the Introduction, this is very significant from our point of view, as in the sequel we will show that the learning sample complexity is governed by GC deviation estimates of certain classes associated with F, where the deviation is of the same order of magnitude as the largest variance of a member of those classes.

Theorem 2.1: Assume that F is a class of functions into $[0, 1]$. Let

$$\sigma^2 = \sup_{f\in F}\mathbb{E}_\mu(f - \mathbb{E}_\mu f)^2, \qquad S_n = n\sigma^2 + \sqrt{n}\,\overline{R}_{n,\mu}.$$

For every L, $S > 0$ and $t > 0$ define

$$\phi_{L,S}(t) = \begin{cases} \frac{t^2}{L^2 S}, & \text{if } t \leq LS \\ \frac{t}{L}\left(\log\frac{et}{LS}\right)^{1/2}, & \text{if } t > LS. \end{cases}$$

There is an absolute constant C such that if $t \geq C\sqrt{n}\,\overline{R}_{n,\mu}$, then

$$\Pr\left\{\sup_{f\in F}\left|\sum_{i=1}^n f(X_i) - n\mathbb{E}_\mu f\right| \geq t\right\} \leq \exp(-\phi_{C,S_n}(t)).$$

This result was improved by Massart [13] by providing an estimate on the constants appearing above.

The strength of Talagrand's inequality is that, unlike the usual results in Machine Learning literature, it does use the union bound to estimate the deviation of the empirical means from the actual ones. This result may be viewed as a "functional" Bernstein inequality, and it is evident that the performance of a class will depend on the behavior of the Rademacher averages associated with it.

In the following subsection we present a bound on the Rademacher averages using a "global" estimate on the covering numbers—the growth rates of the covering numbers.

A. Estimating $\overline{R}_{n,\mu}$

As a starting point, the classes we are interested in are relatively small. This may be seen by the fact that $\overline{R}_{n,\mu}$ are uniformly bounded as a function of n [15]. Our objective here is to estimate $\overline{R}_{n,\mu}$ as a function of $\sup_{f\in F}\mathbb{E}_\mu f^2$.

An important part of our analysis is the following result, again, due to Talagrand [19], on the expectation of the diameter of F when considered as a subset of $L_2(\mu_n)$.

Lemma 2.2: Let $F \subset B(L_\infty(\Omega))$ and set $\tau^2 = \sup_{f\in F}\mathbb{E}_\mu f^2$. Then

$$\mathbb{E}_\mu \sup_{f\in F}\sum_{i=1}^n f^2(X_i) \leq n\tau^2 + 8\sqrt{n}\,\overline{R}_{n,\mu}.$$

Now, we are ready to present the estimates on $\overline{R}_{n,\mu}(F)$ using data on τ^2 and on the covering numbers of F in empirical L_2 spaces. We use global data, namely, the growth rates of the covering numbers, and not the covering numbers at a specific scale. This enables one to use a "chaining procedure" which leads to considerably better bounds on $\overline{R}_{n,\mu}(F)$, and, thus, to sharper generalization bounds. The chaining argument is hidden because we use Dudley's entropy integral which is based on that idea.

We present our estimates regarding the Rademacher averages in several parts, according to the different growth rates of the covering numbers which are of interest to us.

Lemma 2.3: Let F be a class of functions into $[0, 1]$ and set $\tau^2 = \sup_{f\in F} \mathbb{E}_\mu f^2$. Assume that there are $\gamma > 1$, $d \geq 1$, and $p \geq 1$ such that for every empirical measure μ_n

$$\log N(\varepsilon, F, L_2(\mu_n)) \leq d \log^p \left(\frac{\gamma}{\varepsilon}\right).$$

Then, there is a constant $C_{p,\gamma}$ such that

$$\overline{R}_{n,\mu} \leq C_{p,\gamma} \max\left\{\frac{d}{\sqrt{n}} \log^p \frac{1}{\tau}, \sqrt{d}\tau \log^{\frac{p}{2}} \frac{1}{\tau}\right\}.$$

Before proving the lemma, we require the next result.

Lemma 2.4: For every $0 \leq p < \infty$ and $\gamma > 1$, there is some constant $c_{p,\gamma}$ such that for every $0 < x < 1$

$$\int_0^x \log^p \frac{\gamma}{\varepsilon}\, d\varepsilon \leq 2x \log^p \frac{c_{p,\gamma}}{x}$$

and $x^{1/2} \log^p \frac{c_{p,\gamma}}{x}$ is increasing and concave in $(0, 10)$.

The first part of the proof follows from the fact that both terms are equal at $x = 0$, but for an appropriate constant $c_{p,\gamma}$, the derivative of the function on left-hand side is smaller than that of the function on the right-hand side. The second part is evident by differentiation.

Proof of Lemma 2.3: Set $Y = \frac{1}{n}\sup_{f\in F}\sum_{i=1}^n f^2(X_i)$. By Theorem 1.7, there is an absolute constant C such that

$$\frac{1}{\sqrt{n}}\,\mathbb{E}_\varepsilon \sup_{f\in F}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right| \leq C\int_0^{\sqrt{Y}} \log^{\frac{1}{2}} N(\varepsilon, F, L_2(\mu_n))\, d\varepsilon = C\sqrt{d}\int_0^{\sqrt{Y}} \log^{\frac{p}{2}} \frac{\gamma}{\varepsilon}\, d\varepsilon.$$

By Lemma 2.4, there is a constant $c_{p,\gamma}$ such that for every $0 < x \leq 1$

$$\int_0^x \log^{\frac{p}{2}} \frac{\gamma}{\varepsilon}\, d\varepsilon \leq 2x \log^{\frac{p}{2}} \frac{c_{p,\gamma}}{x}$$

where $v(x) = \sqrt{x}\log^{p/2}(c_{p,\gamma}/x)$ is increasing and concave in $(0, 10)$.

Since $Y \leq 1$

$$\frac{1}{\sqrt{n}}\,\mathbb{E}_\varepsilon \sup_{f\in F}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right| \leq C_p\sqrt{dY}\log^{\frac{p}{2}} \frac{c_{p,\gamma}}{Y}$$

and since $\tau^2 + 8\overline{R}_{n,\mu}/\sqrt{n} \leq 9$, then by Jensen's inequality, Lemma 2.2 and the fact that v is increasing in $(0, 10)$

$$\begin{aligned}
&\mathbb{E}_\mu\left(Y^{\frac{1}{2}} \log^{\frac{p}{2}} \frac{c_{p,\gamma}}{Y}\right) \\
&\quad\leq (\mathbb{E}_\mu Y)^{\frac{1}{2}} \log^{\frac{p}{2}} \frac{c_{p,\gamma}}{\mathbb{E}_\mu Y} \\
&\quad\leq c_{p,\gamma}\left(\tau^2 + 8\frac{\overline{R}_{n,\mu}}{\sqrt{n}}\right)^{\frac{1}{2}} \log^{\frac{p}{2}} \frac{1}{\tau^2 + \frac{8\overline{R}_{n,\mu}}{\sqrt{n}}} \\
&\quad\leq c_{p,\gamma}\left(\tau^2 + \frac{8\overline{R}_{n,\mu}}{\sqrt{n}}\right)^{\frac{1}{2}} \log^{\frac{p}{2}} \frac{1}{\tau}.
\end{aligned}$$

Therefore,

$$\overline{R}_{n,\mu} \leq C_{p,\gamma}\sqrt{d}\left(\tau^2 + \frac{\overline{R}_{n,\mu}}{\sqrt{n}}\right)^{\frac{1}{2}} \log^{\frac{p}{2}} \frac{1}{\tau}$$

and our claim follows from a straightforward computation. □

Now, we turn to the case where the log-covering numbers are polynomial with exponent $p < 2$.

Lemma 2.5: Let F be a class of functions into $[0, 1]$ and set $\tau^2 = \sup_{f\in F} \mathbb{E}_\mu f^2$. Assume that there are $\gamma \geq 2$ and $p < 2$ such that for every empirical measure μ_n

$$\log N(\varepsilon, F, L_2(\mu_n)) \leq \frac{\gamma}{\varepsilon^p}.$$

Then, there is a constant $C_{p,\gamma}$ such that

$$\overline{R}_{n,\mu} \leq C_{p,\gamma} \max\left\{n^{-\frac{1}{2}\frac{2-p}{2+p}}, \tau^{1-\frac{p}{2}}\right\}.$$

Proof: Again, let

$$Y = \frac{1}{n}\sup_{f\in F}\sum_{i=1}^n f^2(X_i)$$

and given $X_1, \ldots, X_n$, set μ_n to be the empirical measure supported on $X_1, \ldots, X_n$. By Theorem 1.7, for every fixed sample, there is an absolute constant C such that

$$\begin{aligned}
&\frac{1}{\sqrt{n}}\,\mathbb{E}_\varepsilon \sup_{f\in F}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right| \\
&\quad\leq C\int_0^{\sqrt{Y}} \log^{\frac{1}{2}} N(\varepsilon, F, L_2(\mu_n))\, d\varepsilon \\
&\quad\leq \frac{C\gamma^{\frac{1}{2}}}{1 - p/2}\left(\frac{1}{n}\sup_{f\in F}\sum_{i=1}^n f^2(X_i)\right)^{\frac{1}{2}(1-\frac{p}{2})}.
\end{aligned}$$

Taking the expectation with respect to μ and applying Jensen's inequality and Lemma 2.2

$$\begin{aligned}
\overline{R}_{n,\mu} &\leq C_{p,\gamma}\left(\frac{1}{n}\,\mathbb{E}_\mu \sup_{f\in F}\sum_{i=1}^n f^2(X_i)\right)^{\frac{1}{2}(1-\frac{p}{2})} \\
&\leq C_{p,\gamma}\left(\tau^2 + \frac{\overline{R}_{n,\mu}}{\sqrt{n}}\right)^{\frac{1}{2}(1-\frac{p}{2})}.
\end{aligned}$$

Therefore,

$$\overline{R}_{n,\mu} \leq C_{p,\gamma}\left(\tau^2 + \frac{\overline{R}_{n,\mu}}{\sqrt{n}}\right)^{\frac{1}{2}(1-\frac{p}{2})}$$

from which the claim easily follows. □

In a similar fashion to the proofs in Lemmas 2.3 and 2.5, one can obtain the following.

Lemma 2.6: Let F be a class of functions into $[0, 1]$ and set $\tau^2 = \sup_{f\in F} \mathbb{E}_\mu f^2$. Assume that there are $\gamma \geq 2$ and $p < 2$ such that for every empirical measure μ_n and every $\varepsilon < 1$

$$\log N(\varepsilon, F, L_2(\mu_n)) \leq \frac{\gamma}{\varepsilon^p} \log^2 \frac{2}{\varepsilon}.$$

Then, there is a constant $C_{p,\gamma}$ such that

$$\overline{R}_{n,\mu} \leq C_{p,\gamma} \max\left\{ n^{-\frac{1}{2}\frac{2-p}{2+p}} \log^\beta \frac{2}{\tau}, \tau^{1-\frac{p}{2}} \log \frac{2}{\tau} \right\}$$

where $\beta = 4/(2+p)$.

B. Deviation Estimates

After bounding $\overline{R}_{n,\mu}$ using the growth rates of the covering numbers, it is possible to obtain the deviation results we require by applying Theorem 2.1.

Theorem 2.7: Let F be a class of functions whose range is contained in $[0, 1]$ and set $\tau^2 = \sup_{f\in F} \mathbb{E}_\mu f^2$.

1) If there are $\gamma \geq 2$, $d \geq 1$, and $p > 1$ such that for every empirical measure μ_n
$$\log N(\varepsilon, F, L_2(\mu_n)) \leq d \log^p \frac{\gamma}{\varepsilon}$$
then there is a constant $C_{p,\gamma}$ which satisfies that for every $k > 0$
$$S_F(k\tau^2, \delta) \leq C_{p,\gamma} d \max\{k^{-1}, k^{-2}\} \left(\frac{1}{\tau^2} \log^p \frac{\gamma}{\tau}\right) \log \frac{1}{\delta}.$$

2) If there are $\gamma \geq 2$ and $p < 2$ such that for any empirical measure
$$\log N(\varepsilon, F, L_2(\mu_n)) \leq \frac{\gamma}{\varepsilon^p}$$
then there is a constant $C_{p,\gamma}$ which satisfies that for every $k > 0$
$$S_F(k\tau^2, \delta) \leq C_{p,\gamma} \max\{k^{-1}, k^{-2}\} \left(\frac{1}{\tau^2}\right)^{1+\frac{p}{2}} \left(1 + \log \frac{1}{\delta}\right).$$

3) If there are $\gamma \geq 2$ and $p < 2$ such that for any empirical measure μ_n
$$\log N(\varepsilon, F, L_2(\mu_n)) \leq \frac{\gamma}{\varepsilon^p} \log^2 \frac{1}{\varepsilon}$$
there is a constant $C_{p,\gamma}$ for which
$$S_F(k\tau^2, \delta) \leq C_{p,\gamma} \max\{k^{-1}, k^{-2}\} \left(\frac{1}{\tau^2}\right)^{1+\frac{p}{2}} \left(\log^2 \frac{1}{\tau}\right) \left(1 + \log \frac{1}{\delta}\right)$$
for every $k > 0$.

Since the proof is a straightforward (but tedious) calculation and follows from Theorem 2.1, we omit the details.

III. Dominating the Variance

The main assumption used in the proof of learning sample complexity estimate established in [12] was that there is some $B > 0$ such that for every loss function f, $\mathbb{E}_\mu f^2 \leq B\mathbb{E}_\mu f$. Though this is easily satisfied in proper learning (that is, when the target function belongs to the class G) because each f is nonnegative, it is far from obvious whether the same holds for improper learning. In [12], it was observed that if G is convex and F is the squared-loss class then $\mathbb{E}_\mu f^2 \leq B\mathbb{E}_\mu f$, and B depends on the L_∞ bound on the members of G and the target. The question we study is whether the same kind of bound can be established with respect to other L_q norms. We will show that if $q \geq 2$ and if F is the q-loss function associated with G, there is some B such that for every $f \in F$, $\mathbb{E}_\mu f^2 \leq B(\mathbb{E}_\mu f)^{\frac{2}{q}}$. Our proof is based on a geometric characterization of the nearest point map onto a convex subset of L_q. This fact was used in [12] for $q = 2$, but no emphasis was put on the geometric idea behind it. Our methods enable us to obtain the bound in L_q for $q \geq 2$.

Formally, let $2 \leq q < \infty$, set G to be a compact, convex subset of $L_q(\mu)$ which is contained in $B(L_\infty(\Omega))$, and let F be the q-loss class associated with G and T. Hence, each $f \in F$ is given by $f = |T - g|^q - |P_G T - T|^q$, where T is the target concept and $P_G T$ is the nearest point to T in G with respect to the $L_q(\mu)$ norm.

It is possible to show (see the Appendix) that if $1 < q < \infty$ and if $G \subset L_q$ is convex and compact, the nearest point map onto G is a well-defined map, in the sense that each $T \in L_q$ has a unique best approximation in G.

We start our discussion by proving an upper bound on $\mathbb{E}_\mu f^2$.

Lemma 3.1: Let $g \in G$, $1 < q < \infty$, and set $f = \ell_q(g)$. Then

$$\mathbb{E}_\mu f^2 \leq q^2 \mathbb{E}_\mu |g - P_G T|^2.$$

Proof: Given any $\omega \in \Omega$, apply Lagrange's theorem to the function $y = |x|^q$ for $x_1 = g(\omega) - T(\omega)$ and $x_2 = P_G T(\omega) - T(\omega)$. The result follows by taking the expectation and since $|x_1|, |x_2| \leq 1$. □

The next step, which is to bound $\mathbb{E}_\mu |g - P_G T|^2$ from above using $\mathbb{E}_\mu f$ is considerably more difficult. To that end, we require the following definitions which are standard in Banach spaces theory [3], [10].

Definition 3.2: A Banach space is called strictly convex if every $x, y \in X$ such that $x \neq y$ and $||x||, ||y|| = 1$, satisfy that $||x + y|| < 2$. X is called uniformly convex if there is a positive function $\delta(\varepsilon)$ which satisfies that for every $0 < \varepsilon < 2$ and every $x, y \in X$ for which $||x||, ||y|| \leq 1$ and $||x - y|| \geq \varepsilon$, $||x + y|| \leq 2 - 2\delta(\varepsilon)$. Thus,

$$\delta(\varepsilon) = \inf\left\{ 1 - \tfrac{1}{2}||x + y|| \,\middle|\, ||x||, ||y|| \leq 1, ||x - y|| \geq \varepsilon \right\}.$$

The function $\delta(\varepsilon)$ is called the modulus of convexity of X.

It is easy to see that X is strictly convex if and only if its unit sphere does not contain intervals. Indeed, if the unit sphere contains an interval then it is clearly not strictly convex. On the other hand, let $x \neq y$ be such that $||x|| = ||y|| = 1$ and $||(x + y)/2|| = 1$. If there is some $0 < t < 1$ for which $z = tx + (1 - t)y$ satisfies that $||z|| < 1$, then $(x + y)/2$ is a convex combination of z and either x or y. Therefore, $||(x + y)/2|| < 1$—which is impossible, implying that the interval $[x, y]$ is on the sphere of X.

Clearly, if X is uniformly convex then it is strictly convex. Using the modulus of convexity one can provide a lower bound on the distance of an average of elements on the unit sphere of X and the sphere.

From the quantitative point of view, it was shown in [9] that if $2 \leq q < \infty$, the modulus of convexity of L_q is given by

$$\delta_q(\varepsilon) = 1 - (1 - (\varepsilon/2)^q)^{1/q}$$

while for $1 < q < 2$

$$\delta_q(\varepsilon) = (q-1)\varepsilon^2/8 + o(\varepsilon^2).$$

The next lemma enables one to prove the desired bound on $\mathbb{E}_\mu|g - P_G T|^q$. Its proof is based on several ideas commonly used in the field of Convex Geometry and is presented in the Appendix.

Lemma 3.3: Let X be a uniformly convex, smooth Banach space with a modulus of convexity δ_X and let $G \subset X$ be compact and convex. Set $T \notin G$ and put $d = \|T - P_G T\|$. Then, for every $g \in G$

$$\delta_X\left(\frac{\|g - P_G T\|}{d_g}\right) \leq 1 - \frac{d}{d_g}$$

where $d_g = \|T - g\|$.

Corollary 3.4: Let $q \geq 2$ and assume that G is a compact convex subset of $L_q(\mu)$. If F is the q-loss class associated with G, then for every $g \in G$

$$\mathbb{E}_\mu f^2 \leq 4q^2(\mathbb{E}_\mu f)^{\frac{2}{q}}.$$

Proof: Recall that the modulus of uniform convexity of L_q for $q \geq 2$ is $\delta_q(\varepsilon) = 1 - (1 - (\varepsilon/2)^q)^{1/q}$. By Lemma 3.3

$$1 - \left(\frac{\|g - P_G T\|}{2d_g}\right)^q \geq \left(\frac{d}{d_g}\right)^q.$$

Note that $\mathbb{E}_\mu \ell_q(g) = d_g^q - d^q$, hence, for every $f \in F$

$$\mathbb{E}_\mu f = \mathbb{E}_\mu \ell_q(g) = d_g^q - d^q \geq 2^{-q}\mathbb{E}_\mu|g - P_G T|^q.$$

By Lemma 3.1 and since $\|f\|_2 \leq \|f\|_q$

$$\mathbb{E}_\mu f^2 \leq q^2 \mathbb{E}_\mu|g - P_G T|^2 \leq q^2(\mathbb{E}_\mu|g - P_G T|^q)^{\frac{2}{q}} \leq 4q^2(\mathbb{E}_\mu f)^{\frac{2}{q}}.$$

□

IV. Learning Sample Complexity

Unlike the GC sample complexity, the behavior of the *learning sample complexity* is not monotone, in the sense that even if $H \subset G$, it is possible that the learning sample complexity associated with G may be *smaller* than that associated with H. This is due to the fact that a well-behaved geometric structure of the class (e.g., convexity) enables one to derive additional data regarding the loss functions associated with the class. We will show that the learning sample complexity is upper-bounded by the GC sample complexity of classes of functions with the property that $\sup_{h \in H} \mathbb{E}_\mu h^2$ is roughly the same as the desired accuracy in the GC condition.

We formulate our results in two cases. The first theorem deals with proper learning (that is, $T \in G$). In the second, we discuss improper learning in which T may not belong to G. We present a complete proof only to the second claim.

Let us introduce the following notation: for a fixed $\varepsilon > 0$ and given any empirical measure μ_n, let $f^*_{\mu_n}$ be any $f \in F$ such that $\mathbb{E}_{\mu_n} f^*_{\mu_n} \leq \varepsilon/2$. Thus, if $g \in G$ such that $\ell_q(g) = f^*_{\mu_n}$ then g is an "almost minimizer" of the empirical loss. Also, for every $1 \leq q < \infty$, let $\alpha(q) = \max\{1, 2 - 2/q\}$,

Theorem 4.1: Let $G \subset B(L_\infty(\Omega))$ and fix some $T \in G$. Assume that $1 \leq q < \infty$, and let F be the q-loss class associated with G and T. Assume further that $\gamma \geq 2$, $p < 2$, and that for every integer n and any empirical measure μ_n, $\log N(\varepsilon, G, L_2(\mu_n)) \leq \gamma\varepsilon^{-p}$ for every $\varepsilon > 0$. Then, there is a constant $C_{q,p,\gamma}$ such that if

$$n \geq C_{q,p,\gamma}\left(\frac{1}{\varepsilon}\right)^{\alpha(q)(1+\frac{p}{2})} \log\frac{2}{\delta}$$

then $\Pr\{\mathbb{E}_\mu f^*_{\mu_n} \geq \varepsilon\} \leq \delta$.

The same holds if

$$\sup_{\mu_n} \log N(\varepsilon, G, L_2(\mu_n)) \leq \gamma\varepsilon^{-p}\log^2(1/\varepsilon)$$

and if

$$n \geq C_{q,p,\gamma}\left(\frac{1}{\varepsilon}\right)^{\alpha(q)(1+\frac{p}{2})}\left(\log^2\frac{1}{\varepsilon}\right)\log\frac{1}{\delta}.$$

Next, we turn to the improper case.

Theorem 4.2: Let G be as in Theorem 4.1 and set $T \in B(L_\infty(\Omega))$ which satisfies that $T \notin G$. Fix $q \geq 2$, $0 < p < 2$, and $\gamma > 1$; assume that $G \subset L_2(\mu)$ is convex and closed and that F is the q-loss class associated with G and T. Then, there is a constant $C_{q,p,\gamma}$, for which the following holds.

1) For every $f \in F$, $\mathbb{E}_\mu f^2 \leq 4q^2(\mathbb{E}_\mu f)^{q/2}$.
2) If

$$\sup_n \sup_{\mu_n} \log N(\varepsilon, G, L_2(\mu_n)) \leq \gamma\varepsilon^{-p}$$

then

$$C^q_{G,T}(\varepsilon, \delta) \leq C_{q,p,\gamma}\left(\frac{1}{\varepsilon}\right)^{\alpha(q)(1+\frac{p}{2})}\log\frac{2}{\delta}.$$

3) If

$$\sup_n \sup_{\mu_n} \log N(\varepsilon, G, L_2(\mu_n)) \leq \gamma\varepsilon^{-p}\log^2(1/\varepsilon)$$

then

$$C^q_{G,T}(\varepsilon, \delta) \leq C_{q,p,\gamma}\left(\frac{1}{\varepsilon}\right)^{\alpha(q)(1+\frac{p}{2})}\left(\log^2\frac{1}{\varepsilon}\right)\log\frac{2}{\delta}.$$

We begin with the observation that the learning sample complexity is determined by the GC sample complexity of two classes associated with F, but the deviation required in the GC condition is roughly the largest variance of a member of the classes. Recall that this result was formulated in the Introduction.

Lemma 4.3: Let $G \subset B(L_\infty(\Omega))$, set $q \geq 2$ and put F to be the q-loss class associated with G and the target concept $T \in B(L_\infty(\Omega))$. Assume that there is some constant B such that for any $f \in F$, $\mathbb{E}_\mu f^2 \leq B(\mathbb{E}_\mu f)^{2/q}$. Fix $\varepsilon > 0$ and let $\alpha = 2 - 2/q$. Define

$$H = \left\{ \frac{\varepsilon^\alpha f}{\mathbb{E}_\mu f} \middle| f \in F, \mathbb{E}_\mu f \geq \varepsilon, \mathbb{E}_\mu f^2 \geq \varepsilon \right\} \tag{4.1}$$

and set

$$F_\varepsilon = \left\{ f \in F \middle| \mathbb{E}_\mu f^2 < \varepsilon \right\}$$
$$H_\varepsilon = \left\{ h \in H \middle| \mathbb{E}_\mu h^2 < B\varepsilon^\alpha \right\}.$$

Then, for every $0 < \delta < 1$

$$\mathcal{C}^q_{G,T}(\varepsilon, \delta) \leq \max\left\{ S_{F_\varepsilon}\left(\frac{\varepsilon}{2}, \frac{\delta}{2}\right), S_{H_\varepsilon}\left(\frac{\varepsilon^\alpha}{2}, \frac{\delta}{2}\right)\right\}.$$

Proof: First, note that

$$\begin{aligned}&\Pr\left\{\mathbb{E}_\mu f^*_{\mu_n} \geq \varepsilon\right\}\\ &\quad\leq \Pr\left\{\exists f \in F, \mathbb{E}_\mu f \geq \varepsilon, \mathbb{E}_\mu f^2 < \varepsilon, \mathbb{E}_{\mu_n} f \leq \varepsilon/2\right\}\\ &\qquad+\Pr\left\{\exists f \in F, \mathbb{E}_\mu f \geq \varepsilon, \mathbb{E}_\mu f^2 \geq \varepsilon, \mathbb{E}_{\mu_n} f \leq \varepsilon/2\right\}\\ &\quad=(1)+(2).\end{aligned}$$

If $\mathbb{E}_\mu f \geq \varepsilon$ then

$$\mathbb{E}_\mu f \geq \frac{1}{2}(\mathbb{E}_\mu f + \varepsilon) \geq \frac{1}{2}\mathbb{E}_\mu f + \mathbb{E}_{\mu_n} f.$$

Therefore,

$$|\mathbb{E}_\mu f - \mathbb{E}_{\mu_n} f| \geq \frac{1}{2}\mathbb{E}_\mu f \geq \varepsilon/2$$

hence,

$$\begin{aligned}(1)+(2) \leq \Pr&\left\{\exists f \in F, \mathbb{E}_\mu f^2 < \varepsilon, |\mathbb{E}_\mu f - \mathbb{E}_{\mu_n} f| \geq \frac{\varepsilon}{2}\right\}\\ &+\Pr\left\{\exists f \in F, \mathbb{E}_\mu f \geq \varepsilon, \mathbb{E}_\mu f^2 \geq \varepsilon,\right.\\ &\qquad\left.|\mathbb{E}_\mu f - \mathbb{E}_{\mu_n} f| \geq \frac{1}{2}\mathbb{E}_\mu f\right\}.\end{aligned}$$

Recall that $\alpha = 2 - 2/q$ and that

$$H = \left\{\frac{\varepsilon^\alpha f}{\mathbb{E}_\mu f}\middle| f \in F, \mathbb{E}_\mu f \geq \varepsilon, \mathbb{E}_\mu f^2 \geq \varepsilon\right\}.$$

Since $q \geq 2$ then $\alpha \geq 1$, and since $\varepsilon < 1$, each $h \in H$ maps Ω into $[0, 1]$. Also, if $\mathbb{E}_\mu f^2 \leq B(\mathbb{E}_\mu f)^{2/q}$ then

$$\mathbb{E}_\mu h^2 \leq B\frac{\varepsilon^{2\alpha}}{(\mathbb{E}_\mu f)^{2-2/q}} \leq B\varepsilon^\alpha.$$

Therefore,

$$\begin{aligned}&\Pr\left\{\mathbb{E}_\mu f^*_{\mu_n} \geq \varepsilon\right\}\\ &\quad\leq \Pr\left\{\exists f \in F, \mathbb{E}_\mu f^2 < \varepsilon, |\mathbb{E}_\mu f - \mathbb{E}_{\mu_n} f| \geq \frac{\varepsilon}{2}\right\}\\ &\qquad+\Pr\left\{\exists h \in H, \mathbb{E}_\mu h^2 \leq B\varepsilon^\alpha, |\mathbb{E}_\mu h - \mathbb{E}_{\mu_n} h| \geq \frac{\varepsilon^\alpha}{2}\right\}\end{aligned} \tag{4.2}$$

which proves our claim. □

The only problem in applying Theorem 2.7 directly to H_ε is the fact that one does not have an *a priori* bound on the covering numbers of that class. The question we need to tackle before proceeding is how to estimate the covering numbers of H_ε, given that the covering numbers of F are well behaved. To that end, we have to use the specific structure of F, namely, that it is a q-loss class associated with the class G. We divide our discussion into two parts. First we deal with proper learning, in which each loss function is given by $f = |g - T|^p$ and no specific assumptions are needed on the structure of G. Then we explore the improper case when G is convex and F is the q-loss class for some $q \geq 2$.

To handle the both cases, we need the following simple definition.

Definition 4.4: Let X be a normed space and let $A \subset X$. We say that A is *star shaped* with center x if for every $a \in A$ the interval $[a, x] \subset A$. Given A and x, denote by $\operatorname{star}(A, x)$ the union of all the intervals $[a, x]$, where $a \in A$.

The next lemma shows that the covering numbers of $\operatorname{star}(A, x)$ are almost the same as those of A.

Lemma 4.5: Let X be a normed space and let $A \subset B(X)$ be totally bounded. Then, for any $\|x\| \leq 1$ and every $\varepsilon > 0$

$$\log N(2\varepsilon, \operatorname{star}(A, x)) \leq \log\frac{2}{\varepsilon} + \log N(\varepsilon, A).$$

Proof: Fix $\varepsilon > 0$ and let $y_1, \ldots, y_k$ be an ε-cover of A. Note that for any $a \in A$ and any $z \in [a, x]$ there is some $z' \in [y_i, x]$ such that $\|z' - z\| < \varepsilon$. Hence, an ε-cover of the union $\bigcup_{i=1}^n [y_i, z]$ is a 2ε-cover for $\operatorname{star}(A, x)$. Since for every i $\|x - y_i\| \leq 2$, it follows that each interval may be covered by $2\varepsilon^{-1}$ balls of radius ε and our claim follows. □

Lemma 4.6: Let G be a class of functions which map Ω into $[0, 1]$, put $T \in G$, set $1 \leq q < \infty$, and let F be the q-loss class associated with G and T. Let $\alpha = 2 - 2/q$ and put H as in (4.1). Then, for every $\varepsilon > 0$ and every empirical measure μ_n

$$\log N(2\varepsilon, H, L_2(\mu_n)) \leq \log\frac{2}{\varepsilon} + \log N\left(\frac{\varepsilon}{q}, G, L_2(\mu_n)\right).$$

Proof: Recall that every $h \in H$ is of the form $h = \kappa_f f$ where $0 < \kappa_f \leq 1$. Thus, $H \subset \operatorname{star}(F, 0)$, and by Lemma 4.5

$$\log N(2\varepsilon, H, L_2(\mu_n)) \leq \log\frac{2}{\varepsilon} + \log N(\varepsilon, F, L_2(\mu_n)).$$

Therefore, our claim follows from Lemma 1.2. □

Now, we estimate the covering numbers even when T might not belong to G.

Lemma 4.7: Let $G \subset B(L_\infty(\Omega))$ be a convex class of functions. Set $T \in B(L_\infty(\Omega))$, put F to be the q-loss class associated with G and T, and let α and H be as in Lemma 4.6. Then, for any $\varepsilon > 0$ and any probability measure μ

$$\log N(\varepsilon, H, L_2(\mu)) \leq \log N\left(\frac{\varepsilon}{4q}, G, L_2(\mu)\right) + 2\log\frac{4}{\varepsilon}.$$

Proof: Again, every member of H is given by $\kappa_f f$, where $0 < \kappa_f < 1$. Hence,

$$H \subset \{\kappa\ell_q(g) | g \in G, \kappa \in [0, 1]\} \equiv \mathcal{Q}.$$

By the definition of the q-loss function, it is possible to decompose $\mathcal{Q} = \mathcal{Q}_1 + \mathcal{Q}_2$, where

$$\mathcal{Q}_1 = \{\kappa|g - T|^q | \kappa \in [0, 1], g \in G\}$$

and

$$\mathcal{Q}_2 = \{-\kappa|T - P_G T|^q | \kappa \in [0, 1]\}.$$

Since T and $P_G T$ map Ω into $[0, 1]$ then $|T - P_G T|^q$ is bounded by 1 pointwise. Therefore, $\mathcal{Q}_2$ is contained in an interval whose radius is at most 1, implying that for any probability measure μ

$$N(\varepsilon, \mathcal{Q}_2, L_2(\mu)) \leq \frac{2}{\varepsilon}.$$

Let $V = \{|g - T|^q | g \in G\}$. Since every $g \in G$ and T map Ω into $[0, 1]$ then $V \subset B(L_\infty(\Omega))$. Hence, by Lemma 1.2 and for every probability measure μ and every $\varepsilon > 0$

$$N(\varepsilon, V, L_2(\mu)) \leq N(\varepsilon/q, G, L_2(\mu)).$$

Also, $\mathcal{Q}_1 \subset \operatorname{star}(V, 0)$, thus for any $\varepsilon > 0$

$$\begin{aligned}N(\varepsilon, \mathcal{Q}_1, L_2(\mu)) &\leq 2\frac{N\left(\frac{\varepsilon}{2}, V, L_2(\mu)\right)}{\varepsilon}\\ &\leq 2\frac{N\left(\frac{\varepsilon}{2q}, G, L_2(\mu)\right)}{\varepsilon}\end{aligned}$$

which suffices, since one can combine the separate covers for $\mathcal{Q}_1$ and $\mathcal{Q}_2$ to form a cover for H. □

Finally, we can prove Theorem 4.2. We present a proof only in the case where the metric entropy is $O(\varepsilon^{-p})$ for some $p < 2$. The proof in the other case is essentially the same and is omitted.

Proof of Theorem 4.2: Fix $0 < \varepsilon, \delta < 1$ and let α, F_ε, H and H_ε be as in Lemma 4.3. Note that $F_\varepsilon \subset F$ and $H_\varepsilon \subset H$. Thus, by Lemma 4.7, for every $\rho > 0$ and any probability measure μ_n

$$\log N(\rho, F_\varepsilon, L_2(\mu_n)) \leq \frac{\gamma}{\rho^p}$$

and

$$\log N(\rho, H_\varepsilon, L_2(\mu_n)) \leq \frac{C_{q,p,\gamma}}{\rho^p}.$$

The assertion follows by combining Lemma 4.3 and Theorem 2.7. □

Remark 4.8: It is possible to prove an analogous result to Theorem 4.2 when the covering numbers of G are polynomial; indeed, if there are $\gamma > 1$, $d \geq 1$, and $p > 0$ such that for every $0 < \varepsilon < 1$

$$\sup_n \sup_{\mu_n} \log N(\varepsilon, G, L_2(\mu_n)) \leq C\, d \log^p \frac{\gamma}{\varepsilon}$$

then for every $0 < \varepsilon, \delta < 1$

$$\mathcal{C}^q_{G,T}(\varepsilon, \delta) \leq C_{q,p,\gamma}\, d \left(\varepsilon^{-\alpha(q)} \log^{\max\{1,p\}} \frac{2}{\varepsilon}\right) \log \frac{2}{\delta}$$

where $\alpha(q) = \max\{1, 2 - 2/q\}$.

V. Basic Examples

We present several examples in which one may estimate the learning sample complexity of proper and improper learning problems. All the results are based on estimates on the covering numbers which are obtained either directly or via the fat-shattering dimension. The reason for presenting these examples is to indicate that there are many interesting classes which are both "relatively small" and convex, hence fit our improper learning framework. Although some of the results to follow may not be new, we still think that presenting them in this context emphasizes the fact that the theory developed here covers interesting ground.

A. Proper Learning

The two examples presented in this section are proper learning problems for classes which are either VC classes or classes with polynomial fat-shattering dimension with exponent $p < 2$. By Theorem 1.5, it follows that there is an absolute constant C which satisfies that if G is a VC class for which $\mathrm{VC}(G) = d$, then for every $0 < \varepsilon < 1$

$$\sup_n \sup_{\mu_n} \log N(\varepsilon, G, L_2(\mu_n)) \leq C\, d \log \frac{2}{\varepsilon}$$

whereas if $\mathrm{fat}_\varepsilon(G) \leq \gamma \varepsilon^{-p}$ then there is a constant $c_{p,\gamma}$ such that for every $0 < \varepsilon < 1$

$$\sup_n \sup_{\mu_n} \log N(\varepsilon, G, L_2(\mu_n)) \leq \frac{c_{p,\gamma}}{\varepsilon^p} \log^2 \frac{2}{\varepsilon}.$$

Therefore, applying Theorem 4.2, we can derive the sample complexity estimates for such classes.

Theorem 5.1: Let $G \subset B(L_\infty(\Omega))$, assume that $T \in G$ and that $1 \leq q < \infty$. Set $\alpha(q) = \max\{1, 2 - 2/q\}$.

1) If $\mathrm{VC}(G) = d$, there is a constant C_q such that for every $0 < \varepsilon, \delta < 1$
$$\mathcal{C}^q_{G,T}(\varepsilon, \delta) \leq C_q d \left(\varepsilon^{-\alpha(q)} \log \frac{2}{\varepsilon}\right) \log \frac{2}{\delta}.$$
2) If $\mathrm{fat}_\varepsilon(G) \leq \gamma \varepsilon^{-p}$ for some $\gamma \geq 2$ and $p < 2$, there is a constant $C_{q,p,\gamma}$ such that for every $0 < \varepsilon, \delta < 1$
$$\mathcal{C}^q_{G,T}(\varepsilon, \delta) \leq C_{q,p,\gamma} \left(\frac{1}{\varepsilon}\right)^{\alpha(q)(1+\frac{p}{2})} \left(\log^2 \frac{2}{\varepsilon}\right) \log \frac{2}{\delta}.$$

B. Improper Learning

Recall that if one wishes to use the results in the improper learning setup, one must assume that the concept class is convex. Hence, the most natural starting point is to take the convex hulls of "small" classes. Unfortunately, convex hulls of classes with polynomial fat-shattering dimension are "too large." Even if the fat-shattering dimension of original class is polynomial with an exponent $p < 2$, the covering numbers of its convex hull may be as bad as $\Omega(\varepsilon^{-2} \log^{1-2/p}(1/\varepsilon))$ [5], [16]. Thus, we are left with convex hulls of VC classes. Estimating the covering numbers of VC classes was a well-known problem which was investigated by Dudley [7] and then by Carl and Van-der Vaart and Wellner [4], [20]. The following is a modification of the result in [20], which was presented in [16].

Theorem 5.2: Let G be the convex hull of a class of $\{0, 1\}$-valued functions, denoted by G_0, and assume that $\mathrm{VC}(G_0) = d$. Then, there is an absolute constant C such that for every probability measure μ and every $\varepsilon > 0$

$$\log N(\varepsilon, G, L_2(\mu)) \leq C\, d \left(\frac{1}{\varepsilon}\right)^{\frac{2d}{d+2}}.$$

Corollary 5.3: Let G be as in Theorem 5.2, set $T \in B(L_\infty(\Omega))$ and put $2 \leq q < \infty$. Then, there is a constant $C_{q,d}$ such that for every $0 < \varepsilon, \delta < 1$

$$\mathcal{C}^q_{G,T}(\varepsilon, \delta) \leq C_{q,d} \left(\frac{1}{\varepsilon}\right)^{\beta} \log^2 \frac{2}{\delta}$$

where $\beta = (2 - 2/q)(1 + \frac{d}{d+2})$.

Functions With Bounded Oscillation: There are many important classes of sufficiently smooth functions which appear naturally in learning problems. Such classes of functions fit our setup perfectly, since they usually are convex and uniformly bounded. Though in many problems it is possible to obtain bounds on the covering numbers of such classes directly (see, e.g., [20]), we wish to formulate an estimate on the fat-shattering dimension of a class using data on the ability of members of the class to change quickly. Natural parameters which come to mind in this context are the *variation* of the function and the *oscillation function of the class*. The latter is the supremum of the modulus of continuity of functions in F, that is, for every $\delta > 0$

$$\mathrm{ocs}_F(\delta) = \sup_{f \in F} \sup_{\|x-y\| \leq \delta} |f(x) - f(y)|.$$

Before proving a connection between the "smoothness" properties of the class and its fat-shattering dimension, we require the following property of the fat-shattering dimension of classes which are both convex and symmetric.

Lemma 5.4: Let F be a convex and symmetric class of functions on Ω. If $\{\omega_1, \ldots, \omega_n\}$ is ε-shattered by F then $(s_i)_{i=1}^n = (0, 0, \ldots, 0)$ may be selected as a witness to the shattering.

Proof: Assume that $(s_i)_{i=1}^n$ is a witness to the shattering, and for every $I \subset \{1, \ldots, n\}$, let f_I be the function which shatters the set I. Therefore, for every such I and every $i \in I$

$$f_I(\omega_i) - f_{I^c}(\omega_i) \geq s_i + \varepsilon - s_i + \varepsilon = 2\varepsilon$$

and if $i \notin I$

$$f_I(\omega_i) - f_{I^c}(\omega_i) \leq s_i - \varepsilon - (s_i + \varepsilon) = -2\varepsilon.$$

For every I, let $\tilde{f}_I = (f_I - f_{I^c})/2$. Since F is convex and symmetric, each $\tilde{f}_I$ belongs to F and the set $(\tilde{f}_I)$ ε-shatters $\{\omega_1, \ldots, \omega_n\}$ with $(0, 0, \ldots, 0)$ as a witness. □

Using this observation, it is easy to connect the fat-shattering dimension of a class of functions on Ω with its oscillation and the packing numbers of Ω.

Lemma 5.5: Let G be a convex and symmetric class of functions on a metric space (Ω, ρ). Then, for every $\delta > 0$ and every $\varepsilon > \operatorname{ocs}_\delta(G)/2$, $\operatorname{fat}_\varepsilon(G) \leq D(\delta, \Omega, \rho)$.

Proof: Assume that there are $\delta > 0$ and $\varepsilon > \operatorname{ocs}_\delta(G)/2$ such that $\operatorname{fat}_\varepsilon(G) > D(\delta, \Omega, \rho)$. Thus, there is a set $\{\omega_1, \ldots, \omega_n\}$ which is ε-shattered, such that there are two indexes $i \neq j$, for which $\rho(\omega_i, \omega_j) < \delta$. By Lemma 5.4, we may assume that $(0, 0, \ldots 0)$ is a witness to the shattering. Hence, there is some $g \in G$ such that $|g(\omega_i) - g(\omega_j)| \geq 2\varepsilon$, which is impossible. □

Remark 5.6: Note that a class of functions which is defined by a constraint on its oscillation function is necessarily convex and symmetric, since for every $\delta > 0$, $\operatorname{osc}_G(\delta) = \operatorname{osc}_{\operatorname{absconv}(G)}(\delta)$.

Example 5.7: Let $\Omega \subset B(\mathbb{R}^d)$ and set $G \subset B(L_\infty(\Omega))$ to be a class of functions on Ω such that for every $\delta > 0$, $\operatorname{osc}_G(\delta) < \gamma\delta^p$ for some $p > 0$. In particular, we may assume that F is convex and symmetric. Note that with respect to the Euclidean metric, $D(\delta, \Omega) \leq C\delta^{-d}$. Thus, there is some absolute constant C such that for every $\varepsilon > 0$

$$\operatorname{fat}_\varepsilon(G) \leq C\left(\frac{\gamma}{\varepsilon}\right)^{\frac{d}{p}}$$

which implies that if $d/p < 2$, then for every $T \in B(L_\infty(\Omega))$ and every $q \geq 2$

$$\mathcal{C}_{G,T}^q = O(\varepsilon^{-(1+d/2p)(2-2/q)})$$

up to logarithmic factors in ε^{-1} and δ^{-1}.

A natural example of a family of functions which have a power type oscillation function is the unit ball of certain Sobolev spaces (see [1] for more details).

The second family of functions we shall be interested in is the family of functions with bounded variation.

Definition 5.8: Given $\alpha > 0$, we say that a function f: $[a, b] \to \mathbb{R}$ has an α bounded variation if

$$V_\alpha(f) = \sup \sum_{i=1}^n |f(\omega_i) - f(\omega_{i-1})|^\alpha < \infty$$

where the supremum is taken with respect to all integers n and all the partitions $\{a = \omega_0 < \omega_1 < \cdots < \omega_n = b\}$.

Example 5.9: Let $1 \leq \alpha < 2$ and set $G = \{g | V_\alpha(g) \leq 1\}$. It is easy to see that G is convex and symmetric. Assume that $\{\omega_1, \ldots, \omega_n\}$ is ε-shattered and recall that we may take $(0)_{i=1}^n$ as a witness to the shattering. Thus, there is some $g \in G$ such that for every $2 \leq i \leq n$, $|g(\omega_i) - g(\omega_{i-1})| \geq 2\varepsilon$. The variation of this g satisfies that

$$(2\varepsilon)^\alpha (n-1) \leq V_\alpha(g) \leq 1$$

therefore,

$$\operatorname{fat}_\varepsilon(G) \leq \left(\frac{1}{2\varepsilon}\right)^\alpha + 1.$$

Hence, for every $T \in B(L_\infty(\Omega))$ and every $q \geq 2$

$$\mathcal{C}_{G,T}^q = O(\varepsilon^{-(1+\alpha/2)(2-2/q)})$$

up to logarithmic factors in ε^{-1} and δ^{-1}.

VI. Application: Kernel Machines

In this final section, we present an application of our results to affine functionals on ellipsoids in Hilbert spaces, and in particular, we focus on kernel machines. We present new bounds on the fat-shattering dimension of such classes, which yields an estimate on their covering numbers. We chose to present the results in a separate section since kernel machines are very important in Machine Learning and deserve special attention.

The bounds we present improve some of the bounds appearing in [21]. After presenting our results, we compare them to the ones established in [21].

A. Affine Functionals on ℓ_2

Let A: $\ell_2 \to \ell_2$ be a diagonal operator with eigenvalues $a_1 \geq a_2 \geq \cdots \geq 0$. Set $\Omega = A(B(\ell_2))$ and put $\mathcal{H}$ to be the set of affine functions $h(\omega) = x^*(\omega) + b$, where $||x^*||_{\ell_2} \leq 1$ and $|b| \leq 1$. Our goal is to estimate the fat-shattering dimension of the class $\mathcal{H}$ when considered as functions on Ω.

Tight estimates on the fat-shattering dimension of the class of linear functionals on the unit ball of a Banach space were presented in [8], [14], [16]. In [14], [16] it was shown that if X is infinite-dimensional, the fat-shattering dimension $\operatorname{fat}_\varepsilon(B(X^*), B(X))$ is determined by a geometric property of X, called *type*. The technique used in the proof of that estimate is based on the fact that the domain of the function class is a bounded subset of the Banach space. Intuitively, $A(B(\ell_2))$ should be "much smaller" than a ball (depending, of course, on $(a_i)_{i=1}^\infty$). Hence, there is hope one may be able to obtain an improved bound. Another issue one must address is that we investigate *affine* functions and not just linear ones. Thus, the first order of business is to show that the affine case may be easily reduced to the linear one.

Note that we can embed Ω and $\mathcal{H}$ in ℓ_2. Indeed, each $\omega \in \Omega$ is given by $Ax = (a_i x_i)_{i=1}^\infty$, where $\|x\|_{\ell_2} \leq 1$. We map ω to $\tilde{\omega} = (1, a_1 x_1, a_2 x_2 \cdots)$. The affine function $h = x^* + b$ is mapped to $\tilde{h} = (b, x_1^*, x_2^*, \ldots)$. Therefore, for every h and ω, $\tilde{h}(\tilde{\omega}) = f(\omega)$, and $\|\tilde{h}\|_{\ell_2} \leq 2$. Moreover, $\{\tilde{\omega} | \omega \in \Omega\}$ is the image of the ℓ_2 unit ball under the diagonal operator given by $Te_1 = e_1$, and $Te_i = a_{i-1}e_i$ for $i \geq 2$, where $(e_i)_{i=1}^\infty$ are the unit vectors in ℓ_2. Thus, the class $\tilde{\mathcal{H}}$ is a class of uniformly

bounded linear functionals, and we consider it as a set of functions on a domain $\tilde{\Omega}$, which is the image of unit ball by a diagonal operator with one additional "large" eigenvalue. To simplify things, we will abuse notation and denote our "new" class of linear functionals by $\mathcal{H}$ and the "new" domain by Ω.

The next step in our analysis is to translate the fact that a set $\{x_1, \ldots, x_n\} \subset \Omega$ is ε-shattered to a geometric language.

Lemma 6.1: If $A = \{x_1, \ldots, x_n\}$ is ε-shattered by $B(\ell_2)$ then $\varepsilon B_n \subset \operatorname{absconv}(A)$, where $B_n = B(\ell_2) \cap \operatorname{span}(A)$.

Proof: Assume that the set $\{x_1, \ldots, x_n\}$ is ε-shattered by $B(\ell_2)$. Since $B(\ell_2)$ is convex and symmetric, then by Lemma 5.4, we may assume that $(0)_{i=1}^n$ is a witness to the shattering. Let $(a_i)_{i=1}^n \subset \mathbb{R}$, set $I = \{i | a_i \geq 0\}$, and put x_I^* to be the functional shattering of the set I. Note that for every such I and every $i \in I$

$$x_I^*(x_i) - x_{I^c}^*(x_i) \geq 2\varepsilon$$

and if $i \notin I$

$$x_I^*(x_i) - x_{I^c}^*(x_i) \leq -2\varepsilon.$$

Thus,

$$\begin{aligned}\left\|\sum_{i=1}^n a_i x_i\right\| &= \sup_{x^* \in B(\ell_2)} \left|x^*\left(\sum_{i=1}^n a_i x_i\right)\right| \\ &\geq \frac{1}{2} \sup_{x^*, \tilde{x}^* \in B(\ell_2)} \left|x^*\left(\sum_{i=1}^n a_i x_i\right) - \tilde{x}^*\left(\sum_{i=1}^n a_i x_i\right)\right| \\ &= (*).\end{aligned}$$

Selecting $x^* = x_I^*$ and $\tilde{x}^* = x_{I^c}^*$

$$\begin{aligned}(*) &\geq \frac{1}{2}\left|x_I^*\left(\sum_{i\in I} a_i x_i + \sum_{i \in I^c} a_i x_i\right)\right. \\ &\qquad \left. - x_{I^c}^*\left(\sum_{i\in I} a_i x_i + \sum_{i \in I^c} a_i x_i\right)\right| \\ &= \frac{1}{2}\left|\sum_{i\in I} a_i (x_I^*(x_i) - x_{I^c}^*(x_i))\right. \\ &\qquad \left. + \sum_{i \in I^c} (-a_i)(x_{I^c}^*(x_i) - x_I^*(x_i))\right| \\ &\geq \varepsilon \sum_{i=1}^n |a_i|.\end{aligned}$$

Note that every point on the boundary of $\operatorname{absconv}(A)$ is given by $\sum_{i=1}^n a_i x_i$, where $\sum_{i=1}^n |a_i| = 1$. Hence, by the inequality above, $\|\sum_{i=1}^n a_i x_i\| \geq \varepsilon$, implying that the ℓ_2 norm of every point on the boundary of $\operatorname{absconv}(A)$ is larger than ε. Thus, $\varepsilon B_n \subset \operatorname{absconv}(A)$, as claimed. □

The geometric interpretation of our situation is as follows: first, the set Ω corresponds to an ellipsoid $\mathcal{E}$, which is the image of the ℓ_2 unit ball under a positive semidefinite operator. If Ω contains a set which is ε-shattered by the dual unit ball, then it contains a set A, consisting of n elements, such that $\operatorname{absconv}(A)$ contains an n-dimensional Euclidean ball of radius ε. This brings up the next question we have to face: what is the geometric structure of a set $\{x_1, \ldots, x_n\} \subset \mathcal{E}$ such that its symmetric convex hull contains $\varepsilon B(\ell_2^n)$? Intuitively, one would suspect that if the facets of $\operatorname{absconv}\{x_1, \ldots, x_n\}$ are "far away" from 0, then "most" of the vertices must have a considerably larger norm and should be "close" to orthogonal in some sense. On the other hand, they are restricted by the structure of the ellipsoid $\mathcal{E}$. The analysis of this situation follows from a volumetric argument which requires some knowledge in convex geometry and falls beyond the scope of this article. We refer the interested reader to [17] for a detailed discussion and a proof of a more general claim than the one we require, which is presented in what follows.

Theorem 6.2: Let $\mathcal{E} \subset \ell_2$ be an ellipsoid with principle axes of lengths (a_i) arranged in a nonincreasing order. Assume that there is a set $\{x_1, \ldots, x_n\} \subset \mathcal{E}$ which is ε-shattered by $B(\ell_2)$, let $K = \operatorname{absconv}(x_1, \ldots, x_n)$ and $E = \operatorname{span}(x_1, \ldots, x_n)$. Then, there are absolute constants C and c such that

$$\operatorname{vol}^{\frac{1}{n}}(K) \geq c\varepsilon/\sqrt{n}$$

and

$$\operatorname{vol}^{\frac{1}{n}}(K) \leq \frac{\operatorname{vol}^{\frac{1}{n}}(\mathcal{E} \cap E)}{\sqrt{n}} \leq C \frac{\left(\prod_{i=1}^n a_i\right)^{\frac{1}{n}}}{n} \tag{6.1}$$

where $\operatorname{vol}(\cdot)$ denotes the n-dimensional Lebesgue measure.

In particular

$$\left(\prod_{i=1}^n a_i\right)^{\frac{1}{n}} \geq C\varepsilon\sqrt{n}.$$

The first inequality is a consequence of Lemma 6.1 and standard estimates on the volume of the Euclidean ball. The proof of (6.1) is considerably more difficult (see [17]).

Theorem 6.3: Let A: $\ell_2 \to \ell_2$ be a diagonal operator with eigenvalues $a_1 \geq a_2 \geq \cdots \geq 0$, and set $\mathcal{E} = A(B(\ell_2))$.

1) If there are p, $\gamma > 0$ such that for every integer n, $a_n \leq \gamma/n^p$, there is an absolute constant C such that for every $\varepsilon > 0$
$$\operatorname{fat}_\varepsilon(B(\ell_2), \mathcal{E}) \leq C\left(\frac{\gamma}{\varepsilon}\right)^{\frac{2}{1+2p}}.$$
2) If there are p, $\gamma > 0$ such that for every integer n, $a_n \leq \exp(-\gamma n^p)$, there is an absolute constant C such that for every $\varepsilon > 0$
$$\operatorname{fat}_\varepsilon(B(\ell_2), \mathcal{E}) \leq C\gamma^{-\frac{1}{p}} \log^{\frac{1}{p}} \frac{1}{\varepsilon}.$$

Proof: For the first part, fix $\varepsilon > 0$ and assume that $\{x_1, \ldots x_n\} \subset \mathcal{E}$ is ε-shattered. By Theorem 6.2, there is an absolute constant C such that $(\prod_{i=1}^n a_i)^{1/n} \geq C\varepsilon\sqrt{n}$. On the other hand, using the estimate on the growth rate on (a_i) and Stirling's approximation

$$C\varepsilon\sqrt{n} \leq \left(\prod_{i=1}^n a_i\right)^{\frac{1}{n}} \leq \gamma(n!)^{-\frac{p}{n}} \leq C\gamma\left(\frac{e}{n}\right)^p$$

thus,

$$n \leq C\left(\frac{\gamma}{\varepsilon}\right)^{\frac{2}{1+2p}}$$

as claimed.

The second claim follows since

$$C\varepsilon\sqrt{n} \le \left(\prod_{i=1}^{n} e^{-\gamma i^p}\right)^{\frac{1}{n}} \le e^{-\frac{\gamma}{p+1}(n^p+1)}. \qquad \square$$

Corollary 6.4: Let A be as in Theorem 6.3 and put $\mathcal{E} = A(B(\ell_2))$. Set

$$G = \{x^* + b \,|\, \|x^*\|_{\ell_2} \le 1,\ |b| \le 1\}$$

to be a class of affine functions on $\mathcal{E}$ and let μ to be a probability measure on $\mathcal{E}$.

1) If there are $\gamma \ge 2$ and $p > 0$ such that for every integer n, $a_n \le \gamma n^{-p}$, there is an absolute constant C such that for every $\varepsilon > 0$

$$\log N(\varepsilon, G, L_2(\mu)) \le C\left(\frac{\gamma}{\varepsilon}\right)^{\frac{2}{1+2p}} \log^2 \frac{\gamma}{\varepsilon}.$$

2) If there are $\gamma \ge 2$ and $p > 0$ such that for every integer n, $a_n \le \exp(-\gamma n^p)$, then there is an absolute constant C such that for every $\varepsilon > 0$

$$\log N(\varepsilon, G, L_2(\mu)) \le C\gamma^{-\frac{1}{p}} \log^{2+\frac{1}{p}} \frac{1}{\varepsilon}.$$

Proof: Recall that by the argument presented in the beginning of this section, one may consider G to be a class of linear functionals, which was denoted by $\tilde{G}$. The price one pays is that $\tilde{G}$ is contained in a ball of radius 2 centered at the origin and the "new" domain is an ellipsoid $\tilde{\mathcal{E}}$ which has an additional eigenvalue $a_0 = 1$. Thus, our result follows immediately from Theorem 6.3. $\square$

B. Kernels

One of the most interesting family of function classes appearing in modern Learning Theory is the family of *kernel machines*. In this setup, one is given a positive–definite function $K(-, -)$ defined on $X \times X$, where X is a probability space. Consider a probability measure μ on X and let T_K: $L_2(\mu) \to L_2(\mu)$ be the integral operator defined by K and μ. Thus,

$$T_K f = \int K(x, y) f(y)\, d\mu(y).$$

By Mercer's theorem, T_K has a diagonal representation as an operator on $L_2(\mu)$. Moreover, let $(\phi_n(x))$ be the sequence of eigenvectors of the integral operator T_K and set (λ_n) to be the nonincreasing sequence of eigenvalues associated with the eigenvectors. It is possible to show [18], [6] that (ϕ_i) are orthogonal in $L_2(\mu)$ and that under suitable assumptions on the measure μ

$$K(x, y) = \sum_{i=1}^{\infty} \lambda_i \phi_i(x) \phi_i(y) \tag{6.2}$$

for every $x, y \in X$.

Also, one can define the *reproducing kernel Hilbert space* associated with T_K, which will be denoted by $\mathcal{H}_K$. One of the properties of this Hilbert space is that for every $x \in X$, $K(x, -) \in \mathcal{H}_K$, and for every $h \in \mathcal{H}_K$

$$h(x) = \langle h, K(x, -)\rangle_{\mathcal{H}_K}.$$

We focus on the case in which the eigenvectors of T_K are uniformly bounded functions (i.e., there exists some M such that for every integer n and every $x \in X$, $|\phi_n(x)| \le M$—which is the case, for example, for translation invariant kernels). In that case, every $h \in \mathcal{H}_K$ may be represented by $z_h^* \in \ell_2$ and every $x \in X$ may be represented by some $z_x \in \ell_2$ such that

$$h(x) = \langle h, K(x, -)\rangle_{\mathcal{H}_K} = \langle z_h^*, z_x\rangle_{\ell_2} \tag{6.3}$$

where $\|z_h^*\|_{\ell_2} = \|h\|_{\mathcal{H}_K}$, and there is an ellipsoid $\mathcal{E} \subset \ell_2$ which contains every z_x. The "size" of the ellipsoid $\mathcal{E}$ is determined by the eigenvalues of T_K, as described in the following lemma.

Lemma 6.5 [6], [21]: Let μ be a measure on Ω and set K to be a positive-definite kernel such that the eigenvectors of T_K satisfy that $(\phi_n) \subset B(L_\infty(X))$. Assume further that (6.2) holds, where $(\lambda_n)_{n=1}^{\infty}$ is the nonincreasing sequence of the eigenvalues of T_K. Set $(a_n)_{n=1}^{\infty} \in \ell_2$ to be such that

$$(b_n)_{n=1}^{\infty} = (\sqrt{\lambda_n}/a_n)_{n=1}^{\infty} \in \ell_2$$

and put $R = \|(b_n)\|_{\ell_2}$. If A: $\ell_2 \to \ell_2$ is defined by $Ae_i = Ra_i e_i$, and if $\mathcal{E} = A(B(\ell_2))$, then for every $x \in X$, $z_x \in \mathcal{E}$.

Any such sequence $(a_i)_{i=1}^{\infty}$ is called a scaling sequence, and it determines the lengths of the principle axes of the ellipsoid $\mathcal{E}$.

Example 6.6 [21]: Let K and $(\lambda_n)_{n=1}^{\infty}$ be as in Lemma 6.5, and assume that there are C and $\alpha > 0$ such that for any integer n, $\lambda_n \le Cn^{-(\alpha+1)}$. Then, the scaling sequence $(a_n)_{i=1}^{\infty}$ may be selected as $(a_n)_{i=1}^{\infty} = (n^{-\tau/2})_{i=1}^{\infty}$ for any $\tau < \alpha$. An example of such a kernel is the convolution kernel generated by $k(t) = e^{-t}$.

Example 6.7 [21]: Let K and $(\lambda_n)_{n=1}^{\infty}$ be as in Lemma 6.5 and assume that there are positive B, α and p such that for every integer n, $\lambda_n \le Be^{-\alpha n^p}$. Then, the scaling sequence may be selected as $a_n = e^{-\tau n^p/2}$ for any $\tau < \alpha$. An example of such a kernel is the convolution kernel generated by $k(t) = e^{-t^2}$.

Let us define the class of functions we shall be interested in. Each function consists of a "linear part" which is an element in the unit ball of the reproducing kernel Hilbert space, and an "affine part" which will be a constant in $[-1, 1]$. One can show that this class has the following representation.

Definition 6.8: Let K and $(\lambda_n)_{n=1}^{\infty}$ be as in Lemma 6.5. Set

$$G = \left\{ \sum_{i=1}^{n} \alpha_i K(x_i, -) + b \;\middle|\; n \in \mathbb{N},\ (x_i)_{i=1}^{n} \subset X,\ |b| \le 1,\ \sum_{i,j=1}^{n} \alpha_i \alpha_j K(x_i, x_j) \le 1 \right\}. \tag{6.4}$$

Hence, each "linear part" of a function $g \in G$ is a finite combination of basis functions subject to a constraint on the coefficients, which ensures that it belongs to the unit ball in the reproducing kernel Hilbert space (see the proof below).

Theorem 6.9: Let K, $(\lambda_n)_{n=1}^{\infty}$, and G be as in the definition above, and denote $|K| = \sup_x K(x, x)$.

1) If there are B and $\alpha > 0$ for which $\lambda_n \le Bn^{-(\alpha+1)}$, then for any probability measure μ and any $\tau < \alpha$ there is a constant $C = C_{|K|, B, \tau}$ such that for every $0 < \varepsilon < 1$

$$\log(\varepsilon, G, L_2(\mu)) \le C\left(\frac{1}{\varepsilon}\right)^{\frac{2}{1+\tau}} \log^2 \frac{2}{\varepsilon}.$$

In particular, for every $T \in B(L_\infty(\Omega))$

$$\mathcal{C}^2_{G,T} = O(\varepsilon^{-(1+1/(1+\tau))})$$

up to logarithmic factors in ε^{-1} and δ^{-1}.

2) If there are positive B, α and p such that for every integer n, $\lambda_n \le Be^{-\alpha n^p}$, then for any probability measure μ and every $\tau < \alpha$ there is a constant $C = C_{|K|,B,p,\tau}$, such that for every $0 < \varepsilon < 1$

$$\log(\varepsilon, G, L_2(\mu)) \le C \log^{2+\frac{2}{p}} \frac{2}{\varepsilon}.$$

In particular, for every $T \in B(L_\infty(\Omega))$, $\mathcal{C}^2_{G,T} = O(\varepsilon^{-1})$, up to logarithmic factors in ε^{-1} and δ^{-1}.

Clearly, one can obtain similar bounds for other values of $q > 2$.

Proof: Let $\mathcal{H}_K$ be the reproducing kernel Hilbert space associated with T_K. By the reproducing kernel property it follows that if $h = \sum_{i=1}^n \alpha_i K(x_i, -)$, then

$$\|h\|^2_{\mathcal{H}_K} = \sum_{i,j=1}^n \alpha_i \alpha_j K(x_i, x_j).$$

Thus, the "linear part" $g_h = \sum_{i=1}^n \alpha_i K(x_i, -)$ of every $g \in G$ is contained in the unit ball of $\mathcal{H}_K$. Again, by reproducing the kernel property (6.3) and Lemma 6.5, each g_h may be viewed as a linear functional on an ellipsoid defined by the scaling sequence $(a_i)_{i=1}^\infty$. Applying a similar argument to the one used in Section VI-A, we can identify each $g \in G$ as a linear functional on an ellipsoid which has one additional "large" eigenvalue. Hence, our result follows immediately from the selection of the scaling sequence (Examples 6.6 and 6.7), the covering numbers of the ellipsoid defined by the scaling sequences (Corollary 6.4) and Theorem 4.2. □

Remark 6.10: The condition in (6.4) is imposed simply to ensure that the "linear" part of every $g \in G$ is contained in the unit ball of the reproducing kernel Hilbert space associated with K. This could also be obtained by imposing a convex constraint, namely, that $\sum_{i=1}^n |\alpha_i| = 1$. In that case, every $g = \sum_{i=1}^n \alpha_i K(x_i, -)$ satisfies that $\|g\|_{\mathcal{H}_K} \le |K|$.

It is worthwhile to compare our results with those obtained in [21]. First, note that for generalization estimates, the norm used in [21] is too strong, yielding poorer covering results. Indeed, the authors were able to bound the entropy numbers of the scaling operator A, hence, they provided an ℓ_2-covering numbers estimate on the ellipsoid $\Omega = \mathcal{E}$. When translated to covering numbers of the class $\mathcal{H}$ on the domain Ω, these are, in fact, $L_\infty(\Omega)$ estimates. Indeed, if h is represented by $z_h \in B(\ell_2)$ and every x is represented by $z_x = Ay$, then

$$h(x) = \langle z_h, Ay \rangle = \langle A^* z_h, y \rangle.$$

Hence, the class $\mathcal{H}$ may be viewed as a class of linear functionals contained in $\mathcal{E}^* = A^*(B(\ell_2))$ on a domain which is $B(\ell_2)$. Let $\{x_1^*, \dots, x_n^*\} \subset \mathcal{E}^*$ be an ε-cover of $\mathcal{E}^*$. Thus, $n \le N(\varepsilon/2, \mathcal{E}^*, \ell_2)$. If $\|x^* - x_i^*\| < \varepsilon$, then for every $x \in B(\ell_2)$

$$|x^*(x) - x_i^*(x)| \le \|x^* - x_i^*\| \, \|x\| < \varepsilon.$$

Therefore, for every $\varepsilon > 0$

$$N(\varepsilon, \mathcal{H}, L_\infty(\Omega)) \le N\left(\frac{\varepsilon}{2}, \mathcal{E}^*, \ell_2\right) = N\left(\frac{\varepsilon}{2}, \mathcal{E}, \ell_2\right).$$

Our bounds are $L_2(\mu_n)$ bounds, which suffice for the generalization results and are considerably smaller. For example, if the eigenvalues of the kernel have a polynomial decay with exponent $-(\alpha+1)$, the covering numbers rate obtained in [21] is $O(\varepsilon^{-2/\tau})$ for every $0 < \tau < \alpha/2$, while here we get (up to logarithmic factors) $O(\varepsilon^{-2/(1+\tau)})$.

When the decay rate is exponential, our bound is essentially the same as that in [21], since in both cases the "dominant part" of the covering numbers is the "affine" part (the "$+b$") of the functions, which means that the covering numbers cannot be better than $\Omega(\varepsilon^{-1})$. In our analysis there is an additional effect, which is due to some looseness in the bound on the covering numbers in terms of the fat-shattering dimension. On the other hand, this byproduct has little influence on the complexity bounds, since the dominant term in the learning sample complexity estimate will always be at least of the order of ε^{-1}.

VII. Concluding Remarks

There are several points which deserve closer attention and were not treated here. First, there is the question of the rates of the generalization bounds. Though we believe that the learning sample complexity estimates presented here are optimal with respect to the polynomial scale (i.e., $O(\varepsilon^{-(1+p/2)})$), we have not proved it. Moreover, it is possible that there is some looseness in logarithmic factors in ε^{-1}. Of course, it is important to provide estimates on the constants, an issue which was completely ignored here.

Secondly, we dealt with approximation in L_q for $q \ge 2$. It seems that our analysis does not extend to $1 < q < 2$, since the modulus of convexity of L_q behaves differently for these values of q.

Finally, although we investigated the fat-shattering dimension of uniformly bounded functionals when considered as functions on an ellipsoid in ℓ_2, a major part of the puzzle is still missing. We have not presented the connection between the geometry of the space X, the properties of the operator A, and $\mathrm{fat}_\varepsilon(B(X^*), A(B(X)))$, where $A: X \to X$ is a bounded operator. The only case presented here is when $A = I_X$, in which the fat-shattering dimension is determined by the type of X. The general case is analyzed in [17].

Appendix
Convexity

In this appendix, we present the definitions and preliminary results needed for the proof of Lemma 3.3. All the definitions are standard and may be found in any basic textbook in functional analysis, e.g., [10].

Definition A.1: Given $A, B \subset X$ we say that a nonzero functional $x^* \in X^*$ separates A and B if

$$\inf_{a \in A} x^*(x) \ge \sup_{b \in B} x^*(b).$$

It is easy to see that x^* separates A and B if and only if there is some $\alpha \in \mathbb{R}$ such that for every $a \in A$ and $b \in B$, $x^*(b) \leq \alpha \leq x^*(a)$. In that case, the hyperplane $H = \{x | x^*(x) = \alpha\}$ separates A and B. We denote the closed "positive" halfspace $\{x | x^*(x) \geq \alpha\}$ by H^+ and the "negative" one by H^-. By the Hahn–Banach theorem, if A and B are closed, convex, and disjoint there is a hyperplane (equivalently, a functional) which separates A and B.

Definition A.2: Let $A \subset X$, we say that the hyperplane H supports A in $a \in A$ if $a \in H$ and either $A \subset H^+$ or $A \subset H^-$.

By the Hahn–Banach theorem, if $B \subset X$ is a ball then for every $x \in \partial B$ there is a hyperplane which supports B in x. Equivalently, there is some x^*, $||x^*|| = 1$, and $\alpha \in \mathbb{R}$ such that $x^*(x) = \alpha$ and for every $y \in B$, $x^*(y) \geq \alpha$.

Given a line $V = \{tx + (1-t)y | t \in \mathbb{R}\}$, we say it supports a ball $B \subset X$ in z if $z \in V \cap B$ and $V \cap \operatorname{int}(B) = \emptyset$. By the Hahn–Banach theorem, if V supports B in z, there is a hyperplane which contains V and supports B in z.

Definition A.3: We say that a Banach space X is smooth if for any $x \in X$ there is a unique functional $x^* \in X^*$, such that $||x^*|| = 1$ and $x^*(x) = ||x||$.

Thus, a Banach space is smooth if and only if for every x such that $||x|| = 1$, there is a unique hyperplane which supports the unit ball in x. It is possible to show [10] that for every $1 < q < \infty$, L_q is smooth. On the other hand, ℓ_1^n is not smooth, since there are many hyperplanes supporting its unit ball in the unit vector $e_1 = (1, 0, \ldots, 0)$.

We shall be interested in the properties of the nearest point map onto a compact convex set in "nice" Banach spaces, which is the subject of the following lemma.

Lemma A.4: Let X be a strictly convex space and let $G \subset X$ be convex and compact. Then every $x \in X$ has a unique nearest point in G.

Proof: Fix some $x \in X$ and set $R = \inf_{g \in G} ||g - x||$. By the compactness of G and the fact that the norm is continuous, there is some $g_0 \in G$ for which the infimum is attained, i.e., $R = ||g_0 - x||$.

To show uniqueness, assume that there is some other $g \in G$ for which $||g - x|| = R$. Since G is convex then

$$g_1 = (g + g_0)/2 \in G.$$

By the strict convexity of the norm, $||g_1 - x|| < R$, which is impossible. □

Next, we turn to an important property of the nearest point map onto compact convex sets in strictly convex, smooth spaces.

Lemma A.5: Let X be a strictly convex, smooth Banach space and let $G \subset X$ be compact and convex. Let $x \notin G$ and set $y = P_G x$ to be the nearest point to x in G. If $R = ||x - y||$, then the hyperplane supporting the ball $B = B(x, R)$ at y separates B and G.

Proof: Clearly, we may assume that $x = 0$ and that $R = 1$. Therefore, if x^* is the normalized functional which supports B at y then for every $x \in B$, $x^*(x) \leq 1$. Let $H = \{x | x^*(x) = 1\}$, set H^- to be the open halfspace $\{x | x^*(x) < 1\}$, and assume that there is some $g \in G$ such that $x^*(g) < 1$. Since G is convex, then for every $0 \leq t < 1$, $ty + (1-t)g \in G \cap H^-$. Moreover, since y is the unique nearest point to 0 in G and since X is strictly convex, $[g, y] \cap B = \{y\}$, otherwise, there would have been some $g_1 \in G$ such that $||g_1 - x|| < 1$. Hence, the line $V = \{ty + (1-t)g | t \in \mathbb{R}\}$ supports B in y. By the Hahn–Banach theorem, there is a hyperplane which contains V and supports B in y. However, this hyperplane cannot be H because it contains g. Thus, B was two different supporting hyperplanes at y, contrary to the assumption that X is smooth. □

In the following lemma, our goal is to be able to "guess" the location of some $g \in G$ based on the its distance from $T \notin G$. The idea is that since G is convex and since the norm of X is both strictly convex and smooth, the intersection of a ball centered at the target and G are contained within a "slice" of a ball, that is, the intersection of a ball and a certain halfspace. Formally, we claim the following.

Lemma A.6: Let X be a strictly convex, smooth Banach space and let $G \subset X$ be compact and convex. For any $T \notin G$, let $P_G T$ be the nearest point to T in G and set $d = ||T - P_G T||$. Let x^* be the functional supporting $B(T, d)$ in $P_G T$ and put

$$H^+ = \{x | x^*(x) \geq d + x^*(T)\}.$$

Then, every $g \in G$ satisfies that $g \in B(T, d_g) \cap H^+$, where $d_g = ||f - T||$.

The proof of this lemma is straightforward and is omitted.

Finally, we arrive to the proof of the main claim. We shall estimate the diameter of the "slice" of G using the modulus of uniform convexity of X. This was formulated as Lemma 3.3 in the main text.

Lemma A.7: Let X be a uniformly convex, smooth Banach space with a modulus of convexity δ_X and let $G \subset X$ be compact and convex. If $T \notin G$ and $d = ||T - P_G T||$ then for every $g \in G$

$$\delta_X\left(\frac{||g - P_G T||}{d_g}\right) \leq 1 - \frac{d}{d_g}$$

where $d_g = ||T - g||$.

Proof: Clearly, we may assume that $T = 0$. Using the notation of Lemma A.6

$$||g - P_G T|| \leq \operatorname{diam}\left(B(T, d_g) \cap H^+\right).$$

Let $\tilde{z}_1, \tilde{z}_2 \in (B(T, d_g) \cap H^+)$, put $\varepsilon = ||\tilde{z}_1 - \tilde{z}_2||$ and set $z_i = \tilde{z}_i / d_g$. Hence, $||z_i|| \leq 1$, $||z_1 - z_2|| = \varepsilon / d_g$, and $x^*(z_i) \geq d/d_g$. Thus,

$$\frac{1}{2}||z_1 + z_2|| \geq \frac{1}{2}x^*(z_1 + z_2) \geq \frac{d}{d_g}.$$

Therefore,

$$\frac{d}{d_g} \leq \frac{1}{2}||z_1 + z_2|| \leq 1 - \delta_X\left(\frac{\varepsilon}{d_g}\right)$$

and our claim follows. □

REFERENCES

[1] R. A. Adams, *Sobolev Spaces*, ser. Pure and Applied Mathematics no. 69. New York: Academic, 1975.

[2] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*. Cambridge, U.K.: Cambridge Univ. Press, 1999.

[3] B. Beauzamy, *Introduction to Banach Spaces and Their Geometry*. Amsterdam, The Netherlands: North-Holland, 1982, vol. 86, Math. Studies.

[4] B. Carl, "Metric entropy of convex hulls in Hilbert spaces," *Bull. London Math. Soc.*, vol. 29, pp. 452–458, 1997.

[5] B. Carl, I. Kyrezi, and A. Pajor, "Metric entropy of convex hulls in Banach spaces," *J. London. Math. Soc.*, vol. 60, no. 2, pp. 871–896, 1999.

[6] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Bull. Amer. Math. Soc.*, vol. 39, no. 1, pp. 1–49, 2002.

[7] R. M. Dudley, "Universal Donsker classes and metric entropy," *Ann. Probab.*, vol. 15, pp. 1306–1326, 1987.

[8] L. Gurvits, "A note on the scale-sensitive dimension of linear bounded functionals in Banach spaces," NEC Res. Inst., Tech. Rep., 1997.

[9] O. Hanner, "On the uniform convexity of L^p and l^p," *Ark. Math.*, vol. 3, pp. 239–244, 1956.

[10] P. Habala, P. Hajek, and V. Zizler, *Introduction to Banach Spaces*. Prague, Czech Rep.: Univ. Karlovy, matfyzpress, 1996, vol. I and II.

[11] W. S. Lee, "Agnostic learning and single hidden layer neural networks," Ph.D. dissertation, Australian Nat. Univ., Canberra, 1996.

[12] W. S. Lee, P. L. Bartlett, and R. C. Williamson, "The importance of convexity in learning with squared loss," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1974–1980, Sept. 1998.

[13] P. Massart, "About the constants in Talagrand's concentration inequalities for empirical processes," *Ann. Probab.*, vol. 28, no. 2, pp. 863–884, 2000.

[14] S. Mendelson, "Learnability in Hilbert spaces with reproducing kernels," *J. Complexity*, vol. 18, no. 1, pp. 152–170, 2002.

[15] ——, "Rademacher averages and phase transitions in Glivenko–Cantelli classes," *IEEE Trans. Inform. Theory*, vol. 48, pp. 251–263, Jan. 2002.

[16] ——, "On the size of convex hulls of small sets," *J. Machine Learning Res.*, vol. 2, pp. 1–18, 2001.

[17] ——, Geometric parameters of kernel machines. Preprint. [Online]. Available: http://axiom.anu.edu.au/~shahar

[18] S. Saitoh, *Integral Transforms, Reproducing Kernels and Their Applications*. Reading, MA: Addison-Wesley, 1997, Pitman Research Notes in Mathematics no. 369.

[19] M. Talagrand, "Sharper bounds for Gaussian and empirical processes," *Ann. Probab.*, vol. 22, no. 1, pp. 28–76, 1994.

[20] A. W. Van-der-Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes*. Berlin, Germany: Springer-Verlag, 1996.

[21] R. C. Williamson, A. J. Smola, and B. Schölkopf, "Generalization performance of regularization networks and support vectors machines via entropy numbers of compact operators," *IEEE Trans. Inform. Theory*, vol. 47, pp. 2516–2532, Sept. 2001.