Improving the sample complexity using global data

Shahar Mendelson [1]

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.

Key words: learning sample complexity, Glivenko-Cantelli classes, fat-shattering dimension, kernel machines, uniform convexity.

[1] Computer Sciences Laboratory, RSISE, The Australian National University, Canberra 0200, Australia. Email: shahar@csl.anu.edu.au

1 Introduction

In this article we present sample complexity estimates for proper and improper learning problems with respect to different loss functions under the assumption that the classes are not "too large". Unlike previous results, where 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 $p < 2$.

The question we explore is the following: let G be a class of functions defined on a probability space (Ω, μ) such that each $g \in G$ maps Ω into $[0,1]$, and set T to be an unknown function (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 $S_n = (X_1, ..., X_n)$ 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,

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

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

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 the target concept (which, for the sake of simplicity, is assumed to be deterministic), set $1 \leq q < \infty$ and let $F = \{ |g - T|^q \, | g \in G \}$. The Glivenko-Cantelli sample complexity of the class F with respect to accuracy ε and confidence δ is the smallest integer n_0 such that for every $n \geq n_0$,

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

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

Hence, if g is an "almost" minimizer of the empirical loss and if the sample is "large enough" then g is an "almost" minimizer with respect to average loss. 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 *agnostic* case, in which T may be random (for further details, see [2]).

Lately, it was shown [14] 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 $\Theta\big(\varepsilon^{-\max\{2,p\}}\big)$

up to logarithmic factors in ε^{-1}, δ^{-1}. This implies that if $p \geq 2$ the learning sample complexity has the same rate as the GC sample complexity, since the learning sample complexity is $\Omega\big(\mathrm{fat}_{4\varepsilon}(G)\big)$, at least with respect to the quadratic loss [2].

It is important to emphasize that the learning sample complexity may be established by other means rather than via the GC condition. Hence, 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 loss class given by $\{|g-T|^2 - |T-P_GT|^2 \,|\, g \in G\}$, where P_GT is a nearest point to T in G with respect to the $L_2(\mu)$ norm. Assume that there is an absolute constant C such that for every $f \in F$, $\mathbb{E}_\mu f^2 \leq C\mathbb{E}_\mu f$, i.e., that it is possible to control the variance of each loss function using its expectation. In this case, the learning sample complexity with accuracy ε and confidence δ is

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

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 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 $q = 2$, i.e., every loss function 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 up is whether the same kind of a result holds in other L_q spaces.

Here, we combine the ideas used in [12] and in [14] to improve the learning complexity estimates. We show that if G maps Ω into $[0,1]$ such that

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

for $p < 2$, and if either $T \in G$ or if $q \geq 2$ and $G \subset L_q(\mu)$ is compact and convex, then the learning sample complexity with respect to the q-loss is $O(\varepsilon^{1+p/2})$ up to logarithmic factors in $\varepsilon^{-1}, \delta^{-1}$. Recently it was shown in [14] that there an absolute constant C such that

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

therefore, the estimates we obtain improve the $O\big(\varepsilon^{-(1+p)}\big)$ established in [12].

Although we assume that the target concept is a deterministic function, all the results presented here in the improper setup are true even in the agnostic case, where the target concept T may be random. This is due to the fact that learning in the agnostic sense is equivalent to the ability of finding an "almost best approximation" of the conditional expectation of T (see [2]).

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 target concept and $1 \leq q < \infty$, then for every $g \in G$ let $\ell_q(g)$ be its q-loss function. Thus,

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

where $P_G T$ 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 following lemma is at the heart of our discussion. Its proof will be presented in section 4.

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

$$F_\varepsilon = \Big\{ f \in F |\ \mathbb{E}_\mu f^2 < \varepsilon \Big\}, \quad \text{and} \quad H_\varepsilon = \Big\{ h \in H |\ \mathbb{E}_\mu h^2 < B\varepsilon^\alpha \Big\}.$$

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

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

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). This holds provided that the loss class F contains 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 article. 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 below) 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 the $\mathbb{E}_\mu f$. 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, $\mathbb{E}_\mu f^2 \leq B(\mathbb{E}_\mu f)^{2/q}$. We show that those estimates are the key behind the proof of 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 (improper) learning sample complexity of convex hulls of 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 show that the fat-shattering dimension is determined by the rate of decay of the eigenvalues of the kernel, and improve the covering numbers estimates established in [20].

It is important to mention that throughout this article 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 a bound 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 article.

Given a real Banach space X, let $B(X)$ be the unit ball of X. If $B \subset X$ is a ball, set $int(B)$ 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 $[x, y] = \{tx + (1-t)y | 0 \leq t \leq 1\}$.

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 in $\{\omega_i\}$. If F is a class of functions and g is any function, let $|F - g|^q = \{|f - g|^q \, | f \in F\}$.

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 $d(x, Y) = \inf_{y \in Y} d(x, y)$. If X is a Hilbert space and Y is a subspace, then $d(x, y) = \|P_{Y^\perp} x\|$, where $P_{Y^\perp}$ is the orthogonal projection on the ortho-complement of 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 $\mathrm{absconv}(A)$, 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 shall 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. Set $D(\varepsilon, F)$ to 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 $N(\varepsilon, F) \leq D(\varepsilon, F) \leq N(\varepsilon/2, F)$.

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 $G \subset B\big(L_\infty(\Omega)\big)$ and let F be the q-loss class associated with G. Then, for any probability measure μ and every $\varepsilon > 0$,*

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

Proof: For every target concept 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

$$\big||x_1|^q - |x_2|^q\big| \leq q\,|x_1 - x_2|\,.$$

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

$$\Big||g(\omega) - T(\omega)|^q - \big|g'(\omega) - T(\omega)\big|^q\Big| \leq q\,\big|g(\omega) - g'(\omega)\big|\,. \tag{1.3}$$

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 into $[0,1]$, which, combined with (1.3), implies that $|G' - T|^q$ is an ε/q-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 instead of Lagrange's Theorem, simply apply the triangle inequality.

■

Two 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 $\{\omega_1, ..., \omega_n\} \subset \Omega$, if for every $I \subset \{1, ..., n\}$ 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*

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

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

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 $A = \{\omega_1, ..., \omega_n\} \subset \Omega$ is said to be ε–shattered by F if there is some function $s : A \to \mathbb{R}$, such that for every $I \subset \{1, ..., n\}$ there is some $f_I \in F$ for which $f_I(\omega_i) \geq s(\omega_i) + \varepsilon$ if $i \in I$, and $f_I(\omega_i) \leq s(\omega_i) - \varepsilon$ if $i \notin I$. Let*

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

f_I is called the shattering function of the set I and the set $(s_i) = \big\{s(\omega_i) | \omega_i \in A\big\}$ is called a witness to the ε-shattering.

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 [19], while the second was established in [14].

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

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

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

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,*

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

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*

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

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

It is possible to estimate $\bar{R}_{n,\mu}$ of a given class using its $L_2(\mu_n)$ covering numbers. The key result behind this estimate is the following deep theorem which is due to Dudley for Gaussian processes, and was extended to the more general setting of subgaussian processes [19]. We shall formulate it only for Rademacher processes.

Theorem 1.7 *There is an absolute constant C such that for every sample $(X_1, ..., X_n)$,*

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

where $\delta = \sup_{f\in F}\Big(\frac{1}{n}\sum_{i=1}^n f^2(X_i)\Big)^{\frac{1}{2}}$ and μ_n is the empirical measure supported on the given sample $(X_1, ..., X_n)$.

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.

2 Glivenko-Cantelli estimates

The main tool we use in the analysis of the GC sample complexity is a concentration inequality which is due to Talagrand [18]. This concentration result enables one to control the GC sample complexity using two parameters. The first one is the n-th 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*

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

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

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

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

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

In the following section we present a bound on the Rademacher averages using a "global" estimate on the covering numbers, i.e., using the growth rates of the covering numbers.

2.1 Estimating $\bar{R}_{n,\mu}$

As a starting point, the classes we are interested in are very small. This may be seen by the fact that $\bar{R}_{n,\mu}$ are uniformly bounded as a function of n [14]. Our objective here is to estimate $\bar{R}_{n,\mu}$ as a function of $\sup_{f\in F}\mathbb{E}_\mu f^2$.

An important part of our analysis is the following result, due to Talagrand [18], on the expectation of the diameter of F when considered as a subset of $L_2(\mu_n)$.

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

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

Next, we present the estimates on $\bar{R}_{n,\mu}(F)$, using data on τ^2 and on the covering numbers of F in empirical L_2 spaces. Unlike the usual results in Machine Learning literature, we use global data - the growth rates of the covering numbers and not the covering numbers at a specific scale. We present the results in several parts, according to the different growth rates of the covering numbers which are of interest.

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 $\gamma \ge 1$, $d \ge 1$ and $p \ge 1$ such that for every empirical measure μ_n,*

$$\log N\big(\varepsilon, F, L_2(\mu_n)\big) \le d\log^p\Big(\frac{\gamma}{\varepsilon}\Big).$$

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

$$\bar{R}_{n,\mu} \le C_{p,\gamma}\max\Big\{\frac{d}{\sqrt{n}}\log^p\frac{1}{\tau}, \sqrt{d}\tau\log^{\frac{p}{2}}\frac{1}{\tau}\Big\}.$$

Before proving the lemma, we require the next result:

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

$$\int_0^x \log^p\frac{1}{\varepsilon}d\varepsilon \le 2x\log^p\frac{c_{p,\gamma}}{x},$$

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

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

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

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

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

Therefore,

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

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,*

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

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

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

Proof: Again, let $Y = \frac{1}{n} \sup_{f \in F} \sum_{i=1}^n f^2(X_i)$, and given $X_1, ..., X_n$, set μ_n to be the empirical measure supported on $X_1, ..., X_n$. By theorem 1.7 for every fixed sample, there

is an absolute constant C such that,

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

Taking the expectation with respect to μ and by Jensen's inequality and lemma 2.2,

$$\bar{R}_{n,\mu} \leq C_{p,\gamma}\Big(\frac{1}{n}\mathbb{E}_\mu \sup_{f\in F}\sum_{i=1}^n f^2(X_i)\Big)^{\frac{1}{2}(1-\frac{p}{2})} \leq C_{p,\gamma}\big(\tau^2 + \frac{\bar{R}_{n,\mu}}{\sqrt{n}}\big)^{\frac{1}{2}(1-\frac{p}{2})}.$$

Therefore,

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

and the claim follows.

■

In a similar fashion to the proofs in lemma 2.3 and lemma 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 $\varepsilon < 1$,*

$$\log N\big(\varepsilon, F, L_2(\mu_n)\big) \leq \frac{\gamma}{\varepsilon^p}\log^2\frac{1}{\varepsilon}.$$

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

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

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

2.2 Deviation estimates

After bounding $\bar{R}_{n,\mu}$ 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,*

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

then there is a constant $C_{p,\gamma}$ such that for every $k > 0$

$$S_F(k\tau^2, \delta) \leq C_{p,\gamma} d \max\{k^{-1}, k^{-2}\}\Big(\frac{1}{\tau^2} \log^p \frac{\gamma}{\tau}\Big) \log \frac{1}{\delta}.$$

2. *If there are $\gamma \geq 2$ and $p < 2$ such that for any empirical measure,*

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

then there is a constant $C_{p,\gamma}$ such that for every $k > 0$

$$S_F(k\tau^2, \delta) \leq C_{p,\gamma} \max\{k^{-1}, k^{-2}\}\Big(\frac{1}{\tau^2}\Big)^{1+\frac{p}{2}}\Big(1 + \log \frac{1}{\delta}\Big).$$

3. *If there are $\gamma \geq 2$ and $p < 2$ such that for any empirical measure μ_n,*

$$\log N\big(\varepsilon, F, L_2(\mu_n)\big) \leq \frac{\gamma}{\varepsilon^p} \log^2 \frac{1}{\varepsilon},$$

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

$$S_F(k\tau^2, \delta) \leq C_{p,\gamma} \max\{k^{-1}, k^{-2}\}\Big(\frac{1}{\tau^2}\Big)^{1+\frac{p}{2}}\Big(\log^2 \frac{1}{\tau}\Big)\Big(1 + \log \frac{1}{\delta}\Big),$$

Since the proof is a straightforward (yet tedious) calculation and follows immediately from theorem 2.1, we omit the details.

3 Dominating the variance

The key assumption used in the proof of learning sample complexity 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 (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 $f \in F$, $\mathbb{E}_\mu f^2 \leq B(\mathbb{E}_\mu f)^{\frac{q}{2}}$. Our proof is based on a geometric characterization of the nearest point

map onto a convex subset of an L_q space. 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 $B\big(L_\infty(\Omega)\big)$ 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.

Recall that it is possible to show (see appendix A) 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, i.e., 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,*

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

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 $\mathbb{E}_\mu \left|g - P_G T\right|^2$ 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 for every $x, y \in X$ such that $\|x\|, \|y\| = 1$, $\|x + y\| < 2$. X is called uniformly convex if there is a positive function $\delta(\varepsilon)$ such 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,*

$$\delta(\varepsilon) = \inf\Big\{1 - \frac{1}{2}\|x + y\| \;\Big|\; \|x\|, \|y\| \leq 1,\ \|x - y\| \geq \varepsilon\Big\}.$$

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. Also, 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 $\delta_q(\varepsilon) = 1 - \big(1 - (\varepsilon/2)^q\big)^{1/q}$.

The next lemma enables one to prove the desired bound. Its proof is based on several ideas commonly used in the field of Convex Geometry and is presented in appendix A.

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 $g \in G$,*

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

Proof: Recall that the modulus of uniform convexity of L_q for $q \geq 2$ is $\delta_q(\varepsilon) = 1 - \left(1 - (\varepsilon/2)^q\right)^{1/q}$. 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 $f \in F$,

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

By lemma 3.1 and since $\|f\|_2 \leq \|f\|_q$,

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

■

4 Learning Sample Complexity

Unlike the GC sample complexity, the behaviour 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 determined by the GC sample complexity of classes of functions such that $\sup_{h \in H} \mathbb{E}_\mu h^2$ is roughly the same as the desired accuracy in the GC condition.

We shall formulate results in two cases. The first theorem deals with proper learning (i.e. $T \in G$). In the second, we discuss improper learning. 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.

Theorem 4.1 *Let $G \subset B\big(L_\infty(\Omega)\big)$ and fix $T \in G$. 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, $\log N\big(\varepsilon, G, L_2(\mu_n)\big) \leq \gamma\varepsilon^{-p}$ for every $\varepsilon > 0$. Then, there is a constant $C_{q,p,\gamma}$ such that if*

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

then $Pr\Big\{\mathbb{E}_\mu f^*_{\mu_n} \geq \varepsilon\Big\} \leq \delta$.
The same holds if $\sup_{\mu_n} \log N\big(\varepsilon, G, L_2(\mu)\big) \leq \gamma\varepsilon^{-p}\log^2(1/\varepsilon)$ and if

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

Next, we turn to the improper case.

Theorem 4.2 *Let G be as in theorem 4.1, and assume that $T \notin G$. Assume further that $q \geq @$ and that G is convex, and let F be the q-loss class associated with G and T. Then, for every $f \in F$, $\mathbb{E}_\mu f^2 \leq 4q^2(\mathbb{E}_\mu f)^{q/2}$ and the assertion of theorem 4.1 remains true.*

Remark 4.3 *The key assumption here is that $\mathbb{E}_\mu f^2 \leq B(\mathbb{E}_\mu f)^{2/q}$ for every loss function f. Recall that in section 3 we explored this issue, and in fact, the first part of theorem 4.2 is a reformulation of corollary 3.4. Combining theorem 4.1 (resp. theorem 4.2) with the results of section 3 gives us the promised estimate on the learning sample complexity.*

The first step we take is the important 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.

Lemma 4.4 *Let $G \subset B\big(L_\infty(\Omega)\big)$, set $q \geq 2$ and let F be the q-loss class associated with G and the target concept $T \in B\big(L_\infty(\Omega)\big)$. 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 put

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

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

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

Proof: First, note that

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

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,

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

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

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

Therefore,

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

which proves our claim.

■

The only problem in applying theorem 2.7 directly to the class 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 to 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 4.5 *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* $\mathrm{star}(A,x)$ *the union of all the intervals $[a,x]$, where $a \in A$.*

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

Lemma 4.6 *Let X be a normed space and let $A \subset B(X)$. Then, for any $\|x\| \leq 1$ and every $\varepsilon > 0$, $N\big(2\varepsilon, \mathrm{star}(A,x)\big) \leq 2N\big(\varepsilon, A\big)/\varepsilon$.*

Proof: Fix $\varepsilon > 0$ and let $y_1, ..., y_k$ 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 $\cup_{i=1}^{n}[y_i, z]$ is a 2ε-cover for $\mathrm{star}(A,x)$. 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 4.7 *Let G be a class of functions 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,*

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

Proof: Recall that every $h \in H$ is of the from $h = \kappa_f f$ where $0 < \kappa_f \leq 1$. Thus, $H \subset \mathrm{star}(F, 0)$, and by lemma 4.6,

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

Therefore, our claim follows from lemma 1.2.

■

Now, we estimate the covering numbers in the improper case.

Lemma 4.8 *Let $G \subset B\big(L_\infty(\Omega)\big)$ be a convex class of functions. Set $T \in B\big(L_\infty(\Omega)\big)$, put F to be the q-loss class associated with G and T and let α and H be as in lemma 4.7. Then, for any $\varepsilon > 0$ and any probability measure μ,*

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

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

$$\mathcal{Q}_1 = \Big\{ \kappa \, |g - T|^q \, \Big| \kappa \in [0,1], \ g \in G \Big\} \ \text{ and } \ \mathcal{Q}_2 = \Big\{ -\kappa \, |T - P_G T|^q \, \Big| \kappa \in [0,1] \Big\}.$$

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

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

Let $V = \{|g - T|^q \, | g \in G\}$. Since every $g \in G$ and T map Ω into $[0,1]$ then $V \subset B\big(L_\infty(\Omega)\big)$. Hence, by lemma 1.2 and for every probability measure μ and every $\varepsilon > 0$, $N\big(\varepsilon, V, L_2(\mu)\big) \leq N\big(\varepsilon/q, G, L_2(\mu)\big)$. Also, $\mathcal{Q}_1 \subset \mathrm{star}(V, 0)$, thus for any $\varepsilon > 0$,

$$N\big(\varepsilon, \mathcal{Q}_1, L_2(\mu)\big) \leq 2 \frac{N\big(\frac{\varepsilon}{2}, V, L_2(\mu)\big)}{\varepsilon} \leq 2 \frac{N\big(\frac{\varepsilon}{2q}, G, L_2(\mu)\big)}{\varepsilon},$$

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 4.4. Note that $F_\varepsilon \subset F$ and $H_\varepsilon \subset H$. Thus, by lemma 4.8, for every $\rho > 0$ and any probability measure μ_n,

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

The assertion follows by combining lemma 4.4 and theorem 2.7.

■

Remark 4.9 *It is possible to prove an analogous result to theorem 4.2 when the covering numbers of G are polynomial; indeed, if there are $\gamma, d \geq 1$ and $p > 0$ such that for every $\varepsilon > 0$*

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

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

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

5 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 main reason for presenting these example 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.

5.1 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 if G is a VC class for which $VC(G) = d$, then for every $0 < \varepsilon < 1$,

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

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

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

Therefore, applying theorem 4.2, we can derive the sample complexity estimates for such classes:

Theorem 5.1 *Let $G \subset B\big(L_\infty(\Omega)\big)$, assume that $T \in G$ and that $1 \leq q < \infty$.*

1. *If $VC(G) = d$, there is a constant C_q such that for every $0 < \varepsilon, \delta < 1$,*

$$\mathcal{C}^q_{G,T}(\varepsilon, \delta) \leq C_q d \Big(\frac{1}{\varepsilon} \log \frac{2}{\varepsilon}\Big) \log \frac{1}{\delta}.$$

2. *If $\mathrm{fat}_\varepsilon(G) \leq \gamma \varepsilon^{-p}$ for some $\gamma \geq 2$ and $p < 2$, then there is a constant $C_{q,p,\gamma}$ such that for every $0 < \varepsilon, \delta < 1$,*

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

5.2 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, 15]. 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, 19]. The following is a modification of the result in [19], which was presented in [15].

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

$$\log N\big(\varepsilon, G, L_2(\mu)\big) \leq Cd\Big(\frac{1}{\varepsilon}\Big)^{\frac{2d}{d+2}}.$$

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

$$\mathcal{C}^q_{G,T}(\varepsilon,\delta) \leq C_{q,d}\Big(\frac{1}{\varepsilon}\Big)^{1+\frac{d}{d+2}}\log^2\frac{2}{\delta}.$$

5.2.1 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., [19]), 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, i.e., for every $\delta > 0$,

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

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 $\{\omega_1, ..., \omega_n\}$ is ε-shattered by F, then $(s_i)_{i=1}^n = (0, 0, ..., 0)$ 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 $I \subset \{1, ..., n\}$, let f_I be the functions 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 $\{\omega_1, ...\omega_n\}$ with $(0, 0, ..., 0)$ as a witness.

■

Using this observation, it is easy to connect the fat-shattering dimension of a class of functions on Ω with osc_F 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 $\varepsilon > \mathrm{ocs}_\delta(G)/2$, $\mathrm{fat}_\varepsilon(G) \leq D(\delta, \Omega, \rho)$.*

Proof: Assume that there are $\delta > 0$ and $\varepsilon > \mathrm{ocs}_\delta(G)/2$ such that $\mathrm{fat}_\varepsilon(G) > D(\delta, \Omega, \rho)$. Thus, there is a set $\{\omega_1, ..., \omega_n\}$ which is ε-shattered, such that there are two indices $i \neq j$, for which $\rho(\omega_i, \omega_j) < \delta$. By lemma 5.4 we may assume that $(0, 0, ...0)$ 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 $\delta > 0$, $\mathrm{osc}_G(\delta) = \mathrm{osc}_{\mathrm{absconv}(G)}(\delta)$.*

Example 5.7 *Let $\Omega \subset B(\mathbb{R}^d)$ and let $G \subset B\big(L_\infty(\Omega)\big)$ be a class of functions on Ω such that for every $\delta > 0$, $\mathrm{osc}_G(\delta) < \gamma\delta^p$ 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 $\varepsilon > 0$,*

$$\mathrm{fat}_\varepsilon(G) \leq C\Big(\frac{\gamma}{\varepsilon}\Big)^{\frac{d}{p}},$$

which implies that if $d/p < 2$, then for every $T \in B\big(L_\infty(\Omega)\big)$ and every $q \geq 2$, $\mathcal{C}_{G,T}^q = O\big(\varepsilon^{-(1+d/2p)}\big)$, 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 $\{a = \omega_0 < \omega_1 < ... < \omega_n = b\}$.

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 $\{\omega_1, ..., \omega_n\}$ 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,

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

Hence, for every $T \in B\big(L_\infty(\Omega)\big)$ and every $q \geq 2$, $\mathcal{C}_{G,T}^q = O(\varepsilon^{-(1+\alpha/2)})$ up to logarithmic factors in ε^{-1} and δ^{-1}.

6 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 for two reasons. Firstly, kernel machines are very important in Machine Learning and deserve special attention. Secondly, the bound on the fat-shattering dimension of kernel machines is obtained using a new geometric idea we wish to highlight.

The bounds we present improve some of the bounds appearing in [20]. After presenting our results, we compare them to the ones established in [20].

6.1 Affine functionals on ℓ_2

Let $A : \ell_2 \to \ell_2$ be a diagonal operator with eigenvalues $a_1 \geq a_2 \geq .... \geq 0$. Set $\Omega = A\big(B(\ell_2)\big)$ and put F to be the set of affine functions $f(\omega) = x^*(\omega) + b$, where $\|x^*\|_{\ell_2} \leq 1$

and $|b| \leq 1$. Our goal is to estimate the fat-shattering dimension of the class F 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, 13, 15]. In [13, 15] it was shown that the fat-shattering dimension $\mathrm{fat}_\varepsilon\big(B(X^*), B(X)\big)$ 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, $A\big(B(\ell_2)\big)$ 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 F 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 $\tilde{\omega} = (1, a_1 x_1, a_2 x_2 ...)$. The affine function $f = x^* + b$ is mapped to $\tilde{f} = (b, x_1^*, x_2^*, ...)$. Hence, for every f and ω, $\tilde{f}(\tilde{\omega}) = f(\omega)$, and $\|\tilde{f}\|_{\ell_2} \leq 2$. 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 $i \geq 2$, were $(e_i)_{i=1}^\infty$ are the unit vectors in ℓ_2. Thus, the class $\tilde{F}$ 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 F and the "new" domain by Ω.

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

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

Proof: Assume that the set $\{x_1, ..., x_n\}$ 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 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,

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

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

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

Note that every point on the boundary of absconv(A) is given by $\sum_{i=1}^n a_i x_i$, where $\sum_{i=1}^n |a_i| = 1$. Hence, $\|\sum_{i=1}^n a_i x_i\| \geq \varepsilon$, which proves our claim.

■

The geometric interpretation of our situation is as follows: firstly, the set Ω corresponds to an ellipsoid which is the image of the ℓ_2 unit ball under a positive semi-definite 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 absconv(A) contains an n-dimensional Euclidean ball of radius ε. This brings up the next question we have to tackle: what is the geometric structure of a set $\{x_1, ..., x_n\} \subset \ell_2^n$ which contains $\varepsilon B(\ell_2^n)$? Intuitively, one would suspect that if the facets of absconv$(x_1, ..., x_n)$ 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. Indeed, the following lemma shows that our intuition is correct.

Lemma 6.2 *Let $A = \{x_1, ..., x_n\} \subset \ell_2^n$ and assume that $\varepsilon B(\ell_2^n) \subset$ absconv(A). For every $n \geq 2$, let h_i be the distance of x_i to* span$\{x_1, ..., x_{i-1}\}$ *and set $h_1 = \|x_1\|$. Then, there is an absolute constant C such that*

$$\Big(\prod_{i=1}^n h_i\Big)^{\frac{1}{n}} \geq C\varepsilon\sqrt{n}.$$

Proof: The proof will follow from volume estimates. Since ℓ_2^n is naturally identified with $\mathbb{R}^n$, it is endowed with the Lebesgue measure on $\mathbb{R}^n$. Hence, if $A \subset \mathbb{R}^n$, let vol(A) be the Lebesgue measure of A. Let $B_\varepsilon \equiv \varepsilon B(\ell_2^n)$ and since $S =$ absconv(A) contains an ε ball, its volume must be larger than vol(B_ε). Now, the volume of S can be computed inductively. Indeed, if H is an $n-1$ dimensional subspace of ℓ_2^n and if $B \subset H$ and $x \notin H$, then

$$\text{vol}\big(\text{conv}(B \cup \{x\})\big) = \frac{h}{n}\,\text{vol}(B),$$

where vol(B) is the $n-1$ dimensional volume of B and h is the distance of x to H. Hence,

$$\text{vol}(S) = \frac{2^n}{n!}\prod_{i=1}^n h_i.$$

On the other hand, $\mathrm{vol}(B_\varepsilon) = \varepsilon^n V_n$ where V_n is the volume of the n-dimensional Euclidean unit ball. It is possible to show [16] that there are absolute constants c and C such that for every integer n

$$\frac{c}{\sqrt{n}} \leq V_n^{\frac{1}{n}} \leq \frac{C}{\sqrt{n}}.$$

Comparing the two volumes and by Stirling's approximation,

$$\frac{c\varepsilon}{\sqrt{n}} \leq \big(\mathrm{vol}(B_\varepsilon)\big)^{\frac{1}{n}} \leq \big(\mathrm{vol}(S)\big)^{\frac{1}{n}} \leq \frac{C}{n}\Big(\prod_{i=1}^n h_i\Big)^{\frac{1}{n}},$$

and our claim follows.

■

Corollary 6.3 *Let $\mathcal{E}$ be an ellipsoid with principle axes of length $a_1, ..., a_n$. If there is a set $\{x_1, ..., x_n\} \subset \mathcal{E}$ which is ε-shattered, then there is an absolute constant C such that*

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

Proof: By lemma 6.2 (and using the notation of that lemma), there is an absolute constant C such that

$$\Big(\prod_{i=1}^n h_i\Big)^{\frac{1}{n}} \geq C\varepsilon\sqrt{n}.$$

Since $\mathcal{E}$ is an ellipsoid, then for every subspace H and every $x \in \mathcal{E}$, the orthogonal projection onto H satisfies that $P_H x \in \mathcal{E}$. Hence, one can construct inductively a set $(y_i)_{i=1}^n \subset \mathcal{E}$ of orthogonal vectors, such that for every $1 \leq m \leq n$, $\mathrm{span}(x_1, ..., x_m) = \mathrm{span}(y_1, ..., y_m)$ and $\|y_i\| = h_i$. On the other hand, the maximal product of an orthogonal system contained in an ellipsoid is attained by the principle axes, so $\prod_{i=1}^n a_i \geq \prod_{i=1}^n h_i$.

■

Theorem 6.4 *Let $A : \ell_2 \to \ell_2$ be a diagonal operator with eigenvalues $a_1 \geq a_2 \geq ... \geq 0$, and set $\mathcal{E} = A\big(B(\ell_2)\big)$.*

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 $\varepsilon > 0$,*

$$\mathrm{fat}_\varepsilon\big(B(\ell_2), \mathcal{E}\big) \leq C\Big(\frac{\gamma}{\varepsilon}\Big)^{\frac{2}{1+2p}}.$$

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 $\varepsilon > 0$,*

$$\mathrm{fat}_\varepsilon\big(B(\ell_2), \mathcal{E}\big) \leq C\gamma^{-\frac{1}{p}} \log^{\frac{1}{p}} \frac{1}{\varepsilon}.$$

Proof: For the first part, fix $\varepsilon > 0$ and assume that $\{x_1, ...x_n\} \subset \mathcal{E}$ is ε-shattered. By corollary 6.3 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,

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

thus,

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

as claimed.

The second claim follows since

$$C\varepsilon\sqrt{n} \leq \Big(\prod_{i=1}^n e^{-\gamma i^p}\Big)^{\frac{1}{n}} \leq e^{-\frac{\gamma}{p+1}(n^p+1)}.$$

■

Corollary 6.5 *Let A be as in theorem 6.4 and put $\mathcal{E} = A\big(B(\ell_2)\big)$. Set $F = \{x^* + b| \, \|x^*\|_{\ell_2} \leq 1, \ |b| \leq 1\}$ 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 $\gamma \geq 2$ and $p > 0$ such that for every integer n, $a_n \leq \gamma n^{-p}$, there is an absolute constant C such that for every $\varepsilon > 0$,*

$$\log N\big(\varepsilon, F, L_2(\mu)\big) \leq C\Big(\frac{\gamma}{\varepsilon}\Big)^{\frac{2}{1+2p}} \log^2 \frac{\gamma}{\varepsilon}.$$

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

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

Proof: Recall that by the argument presented in the beginning of this section, one may consider F to be a class of linear functionals, which was denoted by $\tilde{F}$. The price one pays is that $\tilde{F}$ 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 6.4

■

6.2 Kernels

One of the most interesting family of function classes 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. Let $\big(\phi_n(x)\big)$ be the sequence of eigenvectors of the integral operator defined by K and set (λ_n) to be the non increasing sequence of eigenvalues associated with the eigenvectors. It is possible to show [17, 6] that for every $x, y \in X$,

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

Also, if $\mathcal{H}_K$ is the *Reproducing Kernel Hilbert space* associated with the kernel K, then for every $x \in X$, $K(x,-) \in \mathcal{H}_K$, and for every $f \in \mathcal{H}_K$,

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

We focus on the case in which the eigenvectors of K are uniformly bounded functions (i.e., there exists some M such that for every integer n and every $x \in X$, $|\phi_n(x)| \leq M$). In that case, every $f \in \mathcal{H}_K$ may be represented by $z_f^* \in \ell_2$ and every x may be represented by some $z_x \in \ell_2$ such that

$$f(x) = \big\langle f, K(x,-) \big\rangle_{\mathcal{H}_K} = \big\langle z_f^*, z_x \big\rangle_{\ell_2}, \tag{6.1}$$

where $\|z_f^*\|_{\ell_2} = \|f\|_{\mathcal{H}_K}$, 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 K, as described in the following lemma:

Lemma 6.6 *[6, 20] Let K be a positive definite kernel such that the eigenvectors satisfy that $(\phi_n) \subset B\big(L_\infty(X)\big)$. Let $(\lambda_n)_{n=1}^\infty$ be the non increasing sequence of the eigenvalues of 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 $\mathcal{E} = A\big(B(\ell_2)\big)$, 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 length of the principle axes of the ellipsoid $\mathcal{E}$.

Example 6.7 *[20] Let K and $(\lambda_n)_{n=1}^\infty$ be as in lemma 6.6, and assume that there are C and $\alpha > 0$ such that for any integer n, $\lambda_n \leq Cn^{-(\alpha+1)}$. 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 6.8 *[20] Let K and $(\lambda_n)_{n=1}^\infty$ be as in lemma 6.6, and assume that there are positive B, α and p such that for every integer n, $\lambda_n \leq Be^{-\alpha n^p}$. 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 $k(t) = e^{-t^2}$.*

Theorem 6.9 *Let K and $(\lambda_n)_{n=1}^{\infty}$ be as in 6.6 and denote $|K| = \sup_x K(x,x)$. Let*

$$G = \Big\{ \sum_{i=1}^{n} \alpha_i K(x,-) + b \Big| n \in \mathbb{N},\ |b| \le 1,\ \sum_{i,j=1}^{m} \alpha_i \alpha_j k(x_i,x_j) \le 1 \Big\}. \tag{6.2}$$

Then,

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

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

   In particular, for every $T \in B\big(L_\infty(\Omega)\big)$ and $q \ge 2$, $\mathcal{C}_{G,T}^q = O\big(\varepsilon^{-(1+1/(1+\tau))}\big)$, 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, $\lambda_n \le B e^{-\alpha n^p}$, then for any probability measure μ and every $\tau < \alpha$ there is a constant $C = C_{|K|,B,p,\tau}$, such that for every $0 < \varepsilon < 1$,*

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

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

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

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

Thus, the "linear part" $g_f = \sum_{i=1}^{n} \alpha_i K(x_i,-)$ of every $f \in F$ is contained in the unit ball of $\mathcal{H}_K$. Again, by the reproducing kernel property (6.1) and lemma 6.6, each g_f may be viewed as a linear functional on an ellipsoid defined by the scaling sequence $(a_i)_{i=1}^{\infty}$. By a similar argument to the one used in section 6.1 we can identify each $f \in F$ 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, (example 6.7 and example 6.8), the covering numbers of the ellipsoid defined by the scaling sequences (corollary 6.5) and theorem 4.2.

■

Remark 6.10 *The condition in (6.2) 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 $\|g\|_{\mathcal{H}_K} \leq |K|$.*

It is worthwhile to compare our results with those obtained is [20]. Firstly, note that for generalization estimates, the norm used in [20] in 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 F on the domain Ω, these are in fact $L_\infty(\Omega)$ estimates. Indeed, if f is represented by $z_f \in B(\ell_2)$ and every x is represented by $z_x = Ay$, then

$$f(x) = \langle z_f, Ay \rangle = \langle A^* z_f, y \rangle.$$

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

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

Therefore, for every $\varepsilon > 0$

$$N\big(\varepsilon, F, L_\infty(\Omega)\big) \leq N\big(\frac{\varepsilon}{2}, \mathcal{E}^*, \ell_2\big) = N\big(\frac{\varepsilon}{2}, \mathcal{E}, \ell_2\big).$$

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 [20] 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 [20], 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 can not 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 ε^{-1}.

7 Concluding remarks

There are several points which deserve closer attention and were not treated here. Firstly, 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 $q \geq 2$. 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, the majority of the puzzle is still missing. We do not have a clear understanding of the connection between the geometry of the space X, the properties of the operator A and $\mathrm{fat}_\varepsilon\big(B(X^*), A(B(X))\big)$, where $A : X \to X$ is a bounded operator. The only case which is fully understood is when $A = I_X$, in which the fat-shattering dimension is determined by the type of X.

References

[1] R.A. Adams: *Sobolev Spaces*, Pure and Applied Mathematics series 69, Academic Press 1975.

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

[3] B. Beauzamy: *Introduction to Banach spaces and their Geometry*, Math. Studies, vol 86, North-Holland, 1982.

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

[5] B. Carl, I.Kyrezi, A. Pajor: Metric entropy of convex hulls in Banach spaces, J. London. Math. Soc. 60, 2, 871-896, 1999.

[6] F. Cucker, S. Smale: On the mathematical foundations of learning, preprint.

[7] R.M. Dudley: Universal Donsker classes and metric entropy, Annals of Probability 15, 1306-1326, 1987.

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

[9] O. Hanner: On the uniform convexity of L^p and l^p, Ark. Math. 3, 239-244, 1956.

[10] P. Habala, P. Hajek, V. Zizler: *Introduction to Banach spaces* vol I and II, matfyz-press, Univ. Karlovy, Prague, 1996.

[11] W.S. Lee: *Agnostic learning and single hidden layer neural networks*, Ph.D. thesis, The Australian National University, 1996.

[12] W.S. Lee, P.L. Bartlett, R.C. Williamson: The Importance of Convexity in Learning with Squared Loss, IEEE Transactions on Information Theory 44 5, 1974-1980, 1998.

[13] S. Mendelson: Learnability in Hilbert Spaces with Reproducing Kernels, to appear in Journal of Complexity.

[14] S. Mendelson: Rademacher averages and phase transitions in Glivenko–Cantelli classes, preprint.

[15] S. Mendelson: On the size of convex hulls of small sets, preprint.

[16] G. Pisier: *The volume of convex bodies and Banach space geometry*, Cambridge University Press, 1989.

[17] S. Saitoh: *Integral Transforms, Reproducing Kernels and their applications*, Pitman research notes in Mathematics 369, Addison Wesley 1997.

[18] M. Talagrand: Sharper bounds for Gaussian and empirical processes, Annals of Probability, 22(1), 28-76, 1994.

[19] A.W. Van–der–Vaart, J.A. Wellner: *Weak convergence and Empirical Processes*, Springer-Verlag, 1996.

[20] R.C. Williamson, A.J. Smola, B. Schölkopf: Generalization performance of regularization networks and support vectors machines via entropy numbers of compact operators, to appear in IEEE transactions on Information Theory.

A Convexity

In this section we present the definitions and preliminary results needed for the proof of lemma 3.3. All the definitions are standard and my 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* $\inf_{a\in A} x^*(x) \geq \sup_{b\in B} x^*(b)$.

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 $V \cap \text{int}(B) = \emptyset$. 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.

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 can not 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, i.e., 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 corollary 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 $g \in G$,*

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

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

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

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

Let $\tilde{z}_1, \tilde{z}_2 \in \big(B(T, d_g) \cap H^+\big)$, 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,

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

Hence,

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

and our claim follows.

■