

Geometric parameters in Learning Theory

Shahar Mendelson*

Contents

1	Introduction	2
2	Glivenko-Cantelli classes and Learnability	7
2.1	The classical approach	7
2.2	Talagrand's inequality for empirical processes	9
3	Uniform measures of complexity	12
3.1	Metric entropy and the combinatorial dimension	12
3.1.1	Binary valued classes	13
3.1.2	Real valued classes	15
3.2	Random averages and the combinatorial dimension	17
3.3	Phase transitions in GC classes	18
3.3.1	The uniform Central Limit Theorem	18
3.3.2	Uniform ℓ -norm estimates	20
3.3.3	On the size of convex hulls	20
3.4	Concentration of the combinatorial dimension	22
4	Learning Sample complexity and error bounds	24
4.1	Error Bounds	26
4.2	Comparing structures	26
5	Estimating the localized averages	28
5.1	L_2 localized averages	29
5.2	Data dependent bounds	30
5.3	Geometric interpretation	31
6	Bernstein type of L_p loss classes	32
7	Classes of linear functionals	35
8	Concluding Remarks	38

*Research School of Information Sciences and Engineering, Australian National University, Canberra, ACT 0200, Australia, e-mail: shahar.mendelson@anu.edu.au

1 Introduction

In these notes we present some mathematical aspects of Statistical Learning Theory. By no means is this a complete survey which captures the variety of interesting mathematical problems encountered in Machine Learning, but rather, it focuses on one particular problem, called the *sample complexity* problem, which has a geometric flavor. Relevant manuscripts which offer a much broader exposition of Machine Learning (though not necessarily from the mathematical viewpoint) are, for example [2, 14, 15, 36, 62, 64]. We also refer the reader to [7, 6] for results of a similar flavor to the ones we present here.

The surprising fact we wish to stress is that many topics in Machine Learning (and in particular, the sample complexity problem), which have been viewed as problems in Computer Sciences and Nonparametric Statistics, are connected to natural questions arising in the local theory of normed spaces. Our hope is that this article would encourage mathematicians to investigate these seemingly “applied” problems, which are, in fact, linked to interesting theoretical questions.

The starting point of our discussion is the formulation of a *learning problem*. Consider a space Ω endowed with an unknown probability measure μ , let G be a given class of real valued functions defined on Ω , and set T to be a real valued function. The aim of the learner is to approximate T in some sense using a function in G . Throughout these notes we assume that each $g \in G$ and T map Ω into $[0, 1]$. G is called the *base class* and T is referred to as the *target concept*.

The notion of approximation investigated here is with respect to the $L_p(\mu)$ norm (which cannot be computed because the underlying probability measure is unknown). In other words, given $\varepsilon > 0$, the learner attempts to construct a function $g_0 \in G$ which satisfies that

$$\|g_0 - T\|_{L_p(\mu)}^p \leq \inf_{g \in G} \|g - T\|_{L_p(\mu)}^p + \varepsilon. \quad (1.1)$$

The data the learner receives in order to construct the approximating function is a sample $s_n = (X_i, T(X_i))_{i=1}^n$, where $(X_i)_{i=1}^n$ are independent random variables distributed according to μ .

The construction is based on the *learning rule*, which is a mapping that assigns to every sample s_n some $A_{s_n} \in G$. The effectiveness of the learning rule is measured by “how much data” it requires to produce an almost optimal function in the sense of (1.1).

The learning rule we focus on is the *empirical loss minimization*. For the sake of simplicity, assume that the $L_p(\mu)$ distance between T and G is attained at a point denoted by $P_G T$, which is an element in the metric projection of T onto G . Define a new function class which depends on G and T in the following manner; for every $g \in G$, let $\ell_p(g) = |g - T|^p - |P_G T - T|^p$ and set $\mathcal{L}_p(G, T) = \{\ell_p(g) | g \in G\}$. The class $\mathcal{L}_p(G, T)$ is called the p -loss class associated with G and T . Since

$$\mathbb{E}_\mu \ell_p(g) = \|g - T\|_{L_p(\mu)}^p - \inf_{h \in G} \|h - T\|_{L_p(\mu)}^p,$$

it follows that given $\varepsilon > 0$, the learner's aim is to find some $g \in G$ for which $\mathbb{E}_\mu \ell_p(g) < \varepsilon$. Recall that this has to be obtained without apriori knowledge on the measure μ , and thus, with no knowledge on the particular L_p structure one is interested in.

The empirical loss minimization algorithm assigns to every sample a function in G which is "almost optimal" on the data. For every sample (X_1, \dots, X_n) and $\varepsilon > 0$, let $g^* \in G$ be any function which satisfies that

$$\frac{1}{n} \sum_{i=1}^n (g^*(X_i) - T(X_i))^p \leq \inf_{g \in G} \frac{1}{n} \sum_{i=1}^n (g(X_i) - T(X_i))^p + \frac{\varepsilon}{2}. \quad (1.2)$$

Thus, any g^* is an "almost minimizer" of the *empirical L_p distance* between members of G and the target T .

From the definition of the loss class it is evident that if μ_n is the empirical measure $n^{-1} \sum_{i=1}^n \delta_{X_i}$, then $\mathbb{E}_{\mu_n} \ell_p(g^*) \leq \varepsilon/2$, since the second term in every loss function is the same - $|T - P_G T|^p$, and thus the infimum is determined only by the first term $|g - T|^p$. Hence,

$$\mathbb{E}_{\mu_n} \ell_p(g^*) \leq \inf_{f \in \mathcal{L}_p(G, T)} \mathbb{E}_{\mu_n} f + \frac{\varepsilon}{2} \leq \frac{\varepsilon}{2},$$

because $\inf_{f \in \mathcal{L}_p(G, T)} \mathbb{E}_{\mu_n} f \leq 0$, simply by looking at $f = \ell_p(P_G T)$. Therefore, the empirical minimization algorithm is effective if, for any $\varepsilon > 0$, the fact that $\mathbb{E}_{\mu_n} \ell_p(g) < \varepsilon/2$ implies that $\mathbb{E}_\mu \ell_p(g) < \varepsilon$.

Formally, we tackle the following

Question 1.1 *Given a class of functions G , what are the conditions which ensure that for every $\varepsilon > 0$, $0 < \delta < 1$, there exists n_0 such that for every $n \geq n_0$, every target T and any probability measure μ , the following holds. If $(X_i)_{i=1}^n$ are independent and distributed according to μ and g^* is a function which satisfies (1.2), then $\Pr\{\mathbb{E}_\mu \ell_p(g^*) \geq \varepsilon\} \leq \delta$?*

Note that we require an estimate which holds uniformly with respect to all probability measures and all possible targets, because both are unknown. In what follows we shall also be interested in results which are measure dependent.

The reason for the requirement that $\mathbb{E}_\mu \ell_p(g^*)$ is "small" only with high probability, is because it is too much to hope for that any g^* which is almost optimal empirically will be almost optimal with respect to the original L_p structure. In fact, one can encounter arbitrarily large samples which give misleading information on the behavior of the target. The hope is that an affirmative answer can be established with a relatively high probability as the size of the sample increases. The quantitative tradeoff between the desired accuracy ε , the high probability required and the size of the sample is called the *sample complexity problem*.

There are several learning scenarios which are common and which were investigated in the literature. All of them are variations on the theme presented above. The first and simplest case is called proper learning, in which the target concept T belongs to the class. Improper function learning is when T is a

function which does not necessarily belong to the given class. A more general case is called *agnostic learning* in which both the target and μ are represented by a probability measure ν on $\Omega \times [0, 1]$. In this case, the given data points are $(X_i, Y_i)_{i=1}^n \subset \Omega \times [0, 1]$, where $(X_i, Y_i)_{i=1}^n$ are independent random variables distributed according to ν , and the goal is to minimize $\mathbb{E}_\nu |g(x) - y|^p$. The reason for the more general setup is mainly to handle the possibility of noisy observations, which is the standard situation in real-world problems. The analysis we present holds in the agnostic setup with essentially the same proofs. We present the results in the improper case only for the sake of simplicity.

Historically, Learning Theory originated from pattern recognition problems (sometimes referred to as “classification problems”), in which both the base class and the target are assumed to be binary valued, or more generally, to have a finite range. The main tools used in the analysis of these problems were combinatorial in nature (e.g. the Vapnik-Chervonenkis dimension [63]).

In the late 80s the interest in real-valued problems increased, but even then the approach to the sample complexity problem was based on generalized versions of the same tools used in the binary case. In recent years, due to influences from Mathematical Statistics, more modern tools (such as Rademacher and gaussian processes) were introduced in the context of the sample complexity problem. The goal of these notes is to survey these developments from a geometric viewpoint.

In the analysis, we estimate the sample complexity of learning problems using geometric parameters associated with the given class, but we do not attempt to estimate these parameters for particular classes. A more extensive survey would have to contain such estimates, which exist in some cases (see, for example, [2]) but are unknown in many others.

Let us mention some easy observations regarding Question 1.1. Note that any attempt to approximate T with respect to any measure other than the one according to which the sampling is made will not be successful. As an extreme example, if one has two probability measures which are supported on disjoint subsets of Ω , any data obtained by sampling according to one measure will be meaningless when computing distances with respect to the other. Thus, among the various L_p structures possible, the notion of $L_p(\mu)$ approximation is the best one can hope for.

Another observation is that if the class G is “too large” it would be impossible to construct any worthwhile approximating function using empirical data. As an example, assume that G consists of all the measurable functions $g : [0, 1] \rightarrow [0, 1]$, that $T : [0, 1] \rightarrow [0, 1]$ and that μ is the Lebesgue measure on $[0, 1]$. Clearly, there are functions which agree with T completely on any finite sample, but are very far apart with respect to the L_p norm.

The remark above demonstrates a general phenomenon; if G is too large, then for an arbitrarily large sample there are still “too many” very different functions in the class which behave in a similar way to (or even coincide with) the target on the sample. Thus, if one wants an effective learning scheme, additional empirical data (i.e. a larger sample) should decrease the number of class members which are “close” to the target on the sample. Therefore, a

question which emerges is which is the right notion of the “size” of a class?

The approach we present is an outcome of this line of reasoning. Firstly, the easiest notion of size is the ability to approximate means by empirical means. To that end, assume that one can ensure that when the sample size is “large enough”, then for *any* probability measure μ there is a set of “high probability” on which empirical means of members of \mathcal{L}_p are uniformly close to their actual means (that is, if $n > n_0$, with high probability every $f \in \mathcal{L}_p$ satisfies that $|\mathbb{E}_\mu f - \mathbb{E}_{\mu_n} f| < \varepsilon/2$). In particular, since $\mathbb{E}_{\mu_n} \ell_p(g^*) < \varepsilon/2$ then $\mathbb{E}_\mu \ell_p(g^*) < \varepsilon$ as required. This leads to the definition of *uniform Glivenko-Cantelli* classes.

Definition 1.2 *Let F be a class of functions. We say that F is a uniform Glivenko-Cantelli class (uGC class) if for every $\varepsilon > 0$,*

$$\lim_{n \rightarrow \infty} \sup_{\mu} Pr \left\{ \sup_{f \in F} \left| \mathbb{E}_\mu f - \frac{1}{n} \sum_{i=1}^n f(X_i) \right| \geq \varepsilon \right\} = 0,$$

where $(X_i)_{i=1}^\infty$ are independent random variables distributed according to μ .

The uGC condition is simply a uniform version of the law of large numbers, where the uniformity is with respect to all members of F and all probability measures μ .

This definition is natural in the context of Learning Theory because the supremum is taken with respect to all probability measures, and this is essential since the learner does not have apriori information on the probability measure according to which the data is sampled.

Let us mention that if F is uncountable, there is a measurability issue that needs to be addressed, and which will be completely ignored in these notes. It can be resolved by imposing mild conditions on F (see [18] for more details).

The definition of the uGC condition has a quantified version. For every $0 < \varepsilon, \delta < 1$, let $S_F(\varepsilon, \delta)$ be the first integer n_0 such that for every $n \geq n_0$ and any probability measure μ ,

$$Pr \left\{ \sup_{f \in F} |\mathbb{E}_\mu f - \mathbb{E}_{\mu_n} f| \geq \varepsilon \right\} \leq \delta, \quad (1.3)$$

where μ_n is the random empirical measure $n^{-1} \sum_{i=1}^n \delta_{X_i}$.

$S_F(\varepsilon, \delta)$ is called the *Glivenko-Cantelli sample complexity* of the class F with accuracy ε and confidence δ .

One can show [1] that if G consists of uniformly bounded functions then it can be used to approximate *any* target T in the agnostic version of (1.1) (and thus uniformly with respect to the measure) if and only if G is a uniform Glivenko-Cantelli class. Therefore, uniform Glivenko-Cantelli classes are the natural base classes that should be investigated, and from the qualitative viewpoint, the uGC property for the base class captures the correct notion of size. However, from the quantitative point of view, this result is not satisfactory. Firstly, exact estimates on the uGC sample complexity are essential, and should depend on a more fine-grained notion of size. Moreover, the uniform Glivenko-Cantelli condition is not necessarily the optimal approach to the

sample complexity problem. Indeed, the uGC condition provides the ability to control the deviation between empirical means and the actual mean of *every* function within the class. This is a very strong property, and is only a (loose!) condition which suffices to ensure that g^* is a “good approximation” of T . In fact, it is enough to show that empirical means are close to the actual means for functions which are produced by the learning algorithm, i.e., functions for which $\mathbb{E}_{\mu_n} f < \varepsilon$.

Formally, for every $\varepsilon > 0$, one would like to estimate

$$\sup_T \sup_{\mu} Pr \{ \exists f \in \mathcal{L}_p(G, T), \mathbb{E}_{\mu_n} f < \varepsilon, \mathbb{E}_{\mu} f \geq 2\varepsilon \}, \quad (1.4)$$

and define the *learning sample complexity* as the first integer n_0 such that for every $n \geq n_0$ the term in (1.4) is smaller than δ . For such a value of n , there is a set of large probability on which any function which is an “almost minimizer” of the empirical loss will be an “almost minimizer” of the actual loss, regardless of the underlying probability measure. Clearly, this would imply that the empirical minimization algorithm was successful.

Let us remark that an estimate on

$$Pr \{ \exists f \in F, \mathbb{E}_{\mu_n} f < \varepsilon, \mathbb{E}_{\mu} f \geq 2\varepsilon \} \quad (1.5)$$

has a geometric interpretation. Indeed, fix a probability measure μ and assume that G is a $\sqrt{\varepsilon}$ separated class in $L_2(\mu)$. For every (x_1, \dots, x_n) , let μ_n be the empirical measure supported on the set, and let $\tilde{G} = \{(g(x_1), \dots, g(x_n)) \mid g \in G\}$ be the projection of G onto the coordinates (x_1, \dots, x_n) , endowed with the $L_2(\mu_n)$ norm. It is natural to ask when such a coordinate projection of G is also separated in $L_2(\mu_n)$ at a scale proportional to $\sqrt{\varepsilon}$. This is exactly determined by (1.5) - if $F = \{(g-h)^2 \mid g \neq h, g, h \in G\}$, then (1.5) is the probability that for a random set (X_1, \dots, X_n) there is some $f \in F$ for which $n^{-1} \sum_{i=1}^n f(x_i) < \varepsilon$, although one knows that $\mathbb{E}_{\mu} f \geq 2\varepsilon$.

The analysis of the sample complexity problem is presented as follows. In the next section we present an outline of the classical and modern probabilistic approaches used to estimate the Glivenko-Cantelli sample complexity. These approaches lead to the introduction of various notions of size, such as the uniform entropy, the combinatorial dimensions and the Rademacher averages, all of which are investigated in section 3. In section 4 we explore the parameters that govern the learning sample complexity needed to improve the bounds obtained via the Glivenko-Cantelli sample complexity. In particular, we show that the learning sample complexity of an arbitrary class can be reduced to “Glivenko-Cantelli” tail estimates for a class of functions with “small variance”, an observation which leads to sharper complexity estimates. We also present a slightly different approach to the sample complexity problem, known as *error bounds* and compare the empirical and actual structures endowed on the class by μ_n and μ respectively using the notion of *isomorphic coordinate projections*, leading to even sharper bounds. Finally, we present a detailed example of a family of particularly interesting function classes, called *Kernel classes*.

We end this introduction with some notation. We denote the class in question by F , which should be thought of as a loss class associated with a base class G and a target T , although all the results we present hold for general classes of functions which are uniformly bounded. All the constants we use are denoted by c , C or K . Their value may change from line to line, or even within the same line. C_p denoted a constant which depends only on p . We write $a \sim b$ if there are absolute constants c and C such that $cb \leq a \leq Cb$. Finally, if X is a Banach space, we denote its unit ball by B_X or $B(X)$.

2 Glivenko-Cantelli classes and Learnability

In this section we investigate two approaches to the problem of bounding the tails of the random variable $\sup_{f \in F} |\mathbb{E}_\mu f - n^{-1} \sum_{i=1}^n f(X_i)|$. The first approach was the basis to the best known sample complexity estimates up to the mid 90s. We then present a more modern result which is based on a sharper concentration inequality.

2.1 The classical approach

Our starting point is a symmetrization argument which originated in the works of Kahane [29] and Hoffman-Jørgensen [24, 25] (see also [27]).

Theorem 2.1 *Let $F \subset B(L_\infty(\Omega))$. Then, for any $\varepsilon > 0$ and every $n \geq 8/\varepsilon^2$,*

$$Pr \left\{ \sup_{f \in F} \left| \mathbb{E}_\mu f - \frac{1}{n} \sum_{i=1}^n f(X_i) \right| > \varepsilon \right\} \leq 4Pr \left\{ \sup_{f \in F} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right| > \frac{n\varepsilon}{4} \right\}, \quad (2.1)$$

where $(\varepsilon_i)_{i=1}^n$ are independent Rademacher random variables (that is, symmetric and $\{-1, 1\}$ -valued), and the probability on the right-hand side is with respect to the product measure.

Having the symmetrization idea in mind, the course of action one can take is clear: restricted to any sample (X_1, \dots, X_n) , F is projected onto a random set $V \subset \mathbb{R}^n$. To control the possibly infinite set V , one uses the notion of covering.

If (Y, d) is a metric space and $F \subset Y$ then for every $\varepsilon > 0$, $N(\varepsilon, F, d)$ is the minimal number of open balls (with respect to the metric d) needed to cover F , and the set of the centers of the balls is called an ε -cover of F . For obvious reasons, we are interested in metrics endowed by samples. Given a sample (X_1, \dots, X_n) let μ_n be the empirical measure supported on that sample, and for $1 \leq p < \infty$, denote by $N(\varepsilon, F, L_p(\mu_n))$ the covering numbers of F at scale ε with respect to the $L_p(\mu_n)$ norm.

To bound the right-hand side of (2.1), fix a sample (X_1, \dots, X_n) which corresponds to fixing a coordinate projection, and set $\mu_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$. It is clear that given a fixed projection one can replace F with an $\varepsilon/8$ cover of F with respect to the $L_1(\mu_n)$ norm. Bounding the probability of the supremum over the cover by the sum of probabilities, all that remains is to estimate

$Pr \left\{ \left| \sum_{i=1}^n \varepsilon_i v_i \right| \geq t \right\}$ for a fixed vector $(v_i)_{i=1}^n \in \mathbb{R}^n$, which is the coordinate projection of an element of the cover of F with respect to the $L_1(\mu_n)$ norm. By Hoeffding's inequality [61],

$$Pr \left\{ \left| \sum_{i=1}^n \varepsilon_i v_i \right| > t \right\} \leq 2e^{-t^2/2\|v\|_2},$$

where $\|v\|_2 = (\sum_{i=1}^n v_i^2)^{1/2}$.

Finally, taking the expectation with respect to μ , the following is evident:

Theorem 2.2 *Let $F \subset B(L_\infty(\Omega))$ and set μ to be a probability measure on Ω . Let $(X_i)_{i=1}^\infty$ be independent random variables distributed according to μ . For every $\varepsilon > 0$ and any $n \geq 8/\varepsilon^2$,*

$$Pr \left\{ \sup_{f \in F} \left| \mathbb{E}_\mu f - \frac{1}{n} \sum_{i=1}^n f(X_i) \right| > \varepsilon \right\} \leq 8\mathbb{E}_\mu [N(\varepsilon/8, F, L_1(\mu_n))] e^{-\frac{n\varepsilon^2}{128}},$$

where μ_n is the (random) empirical measure supported on (X_1, \dots, X_n) .

If one is interested in measure independent bounds, it is possible to replace $\mathbb{E}_\mu N(\varepsilon, F, L_1(\mu_n))$ with the *uniform covering numbers*, which is the first notion of size we investigate.

Definition 2.3 *For every class F , $1 \leq p \leq \infty$ and $\varepsilon > 0$, let*

$$N_p(\varepsilon, F, n) = \sup_{\nu_n} N(\varepsilon, F, L_p(\nu_n)),$$

and

$$N_p(\varepsilon, F) = \sup_n \sup_{\nu_n} N(\varepsilon, F, L_p(\nu_n)),$$

where ν_n is an average of n point-mass measures.

The logarithm of the covering numbers is called the *metric entropy* of the set. $\log N_p(\varepsilon, F)$ is the uniform L_p entropy of F .

Corollary 2.4 *If $F \subset B(L_\infty(\Omega))$ then for any probability measure μ ,*

$$Pr \left\{ \sup_{f \in F} \left| \mathbb{E}_\mu f - \frac{1}{n} \sum_{i=1}^n f(X_i) \right| > \varepsilon \right\} \leq 8N_1(\varepsilon/8, F, n) e^{-\frac{n\varepsilon^2}{128}}.$$

In particular, if F is such that $\log N_1(\varepsilon, F) = O(\varepsilon^{-p})$ for some $0 < p < \infty$, then the uGC sample complexity satisfies that $S_F(\varepsilon, \delta) = O(\varepsilon^{-(2+p)})$ up to logarithmic factors in $1/\varepsilon$ and $1/\delta$.

As an example, consider $F = \mathcal{L}_q(G, T)$, and observe that the entropy of F is controlled by that of G . Indeed, since each $g \in G$ maps Ω into $[0, 1]$, then for every $g_1, g_2 \in G$ and every $\omega \in \Omega$, $|\ell_q(g_1) - \ell_q(g_2)|(\omega) \leq q|g_1 - g_2|(\omega)$. In

particular, if $\log N_1(\varepsilon, G) = O(\varepsilon^{-p})$, then for every target T , $S_{\mathcal{L}_q(G, T)}(\varepsilon, \delta) \leq C_{p, q} \varepsilon^{-(2+p)}$ up to logarithmic factors in $1/\varepsilon$ and $1/\delta$.

Although these tail bounds will be shown to be loose, the uniform entropy does characterize uniform Glivenko-Cantelli classes. This line of investigation - connecting the metric entropy to the uniform Glivenko-Cantelli property originated in the work of Vapnik and Chervonenkis [63], and then extended by various authors.

Theorem 2.5 [19] *A class $F \subset B(L_\infty(\Omega))$ is a uniform Glivenko-Cantelli class if and only if the following holds. There is some $1 \leq p \leq \infty$ such that for every $\varepsilon > 0$*

$$\lim_{n \rightarrow \infty} \frac{\log N_p(\varepsilon, F, n)}{n} = 0.$$

The fact that uniform entropy can be used to characterize uniform Glivenko-Cantelli classes (and thus, classes in which any target can be approximated) demonstrates its significance. Unfortunately, in most cases it is not easy to estimate the uniform entropy directly. It turns out that one can control the uniform entropy of the class using combinatorial dimensions which measure the “richness” of the class (such as the VC dimension) and those are sometimes easier to estimate (see [2] for some examples). These *uniform measures of complexity* will be discussed in the next section.

It is clear that the main source of possible looseness in the uGC sample complexity bounds is due to the weak concentration result used in the proof. In order to obtain sharper bounds one has to find better ways of controlling the supremum of the random variable $\sup_{f \in F} |\mathbb{E}_\mu f - n^{-1} \sum_{i=1}^n f(X_i)|$.

2.2 Talagrand’s inequality for empirical processes

Let us begin by recalling Bernstein’s inequality in a form analogous to Talagrand’s inequality below [37, 61].

Theorem 2.6 *Let μ be a probability measure on Ω and set X_1, \dots, X_n to be independent random variables distributed according to μ . Given a function $f : \Omega \rightarrow \mathbb{R}$, set $Z = \sum_{i=1}^n f(X_i)$, let $b = \|f\|_\infty$ and put $u = n\mathbb{E}_\mu f^2$. Then,*

$$\Pr\{|Z - \mathbb{E}_\mu Z| \geq x\} \leq 2e^{-\frac{x^2}{2(u+bx/3)}}.$$

This exponential inequality is sharper than Hoeffding’s inequality when one has additional information on the variance of the random variable Z . It has been a long standing question whether a similar result can be obtained when Z is replaced by $\sup_{f \in F} |\sum_{i=1}^n (f(X_i) - \mathbb{E}_\mu f)|$. Such an inequality was first established by Talagrand [59] and later modified by Ledoux [34], Massart [37], Rio [52] and Bousquet [10] who currently has the best estimates on the constants appearing in the inequality. The following is Bousquet’s version of Talagrand’s inequality.

Theorem 2.7 [10] Let F be a class of functions defined on a probability space (Ω, μ) such that $\sup_{f \in F} \|f\|_\infty \leq 1$. Let $(X_i)_{i=1}^n$ be independent random variables distributed according to μ , put $\sigma^2 \geq \sup_{f \in F} \text{Var}[f(X_1)]$ and set $Z = \sup_{f \in F} \left| \sum_{i=1}^n (f(X_i) - \mathbb{E}_\mu f) \right|$. Then, for every $x > 0$,

$$\Pr \{Z \geq \mathbb{E}Z + x\} \leq \exp\left(-vh\left(\frac{x}{v}\right)\right),$$

where $v = n\sigma^2 + 2\mathbb{E}Z$ and $h(x) = (1+x)\log(1+x) - x$. Moreover, for every $x > 0$,

$$\Pr \left\{ Z \geq \mathbb{E}Z + \sqrt{2xv} + \frac{x}{3} \right\} \leq e^{-x}. \quad (2.2)$$

Hence, if F consists of functions which are bounded by 1, then by selecting $x = n\varepsilon^2/4$ it follows that with probability larger than $1 - e^{-\frac{n\varepsilon^2}{4}}$,

$$\sup_{f \in F} \left| \mathbb{E}_\mu f - \frac{1}{n} \sum_{i=1}^n f(X_i) \right| \leq 2\mathbb{E} \sup_{f \in F} \left| \mathbb{E}_\mu f - \frac{1}{n} \sum_{i=1}^n f(X_i) \right| + \frac{3\varepsilon}{4}.$$

The dominating term in (2.2) is the expectation of the random variable Z , and this expectation $\mathbb{E}Z$ can be estimated by a symmetrization argument.

Definition 2.8 Let μ be a probability measure on Ω and set F to be a class of uniformly bounded functions on Ω . For every integer n , let

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

where $(X_i)_{i=1}^n$ are independent random variables distributed according to μ and $(\varepsilon_i)_{i=1}^n$ are independent Rademacher random variables (which are also independent of X_1, \dots, X_n). $R_n(F, \mu)$ are the (global) Rademacher averages associated with the class F and the measure μ .

Proposition 2.9 Let μ be a probability measure and set $F \subset B(L_\infty(\Omega))$. Denote

$$H = \sup_{f \in F} \left| \mathbb{E}_\mu f - \frac{1}{n} \sum_{i=1}^n f(X_i) \right| = \frac{Z}{n},$$

where $(X_i)_{i=1}^n$ are independent random variables distributed according to μ . Then,

$$\mathbb{E}_\mu H \leq 2 \frac{R_n(F, \mu)}{\sqrt{n}} \leq 4\mathbb{E}_\mu H + 2 \left| \sup_{f \in F} \mathbb{E}_\mu f \right| \cdot \mathbb{E}_\varepsilon \left| \frac{1}{n} \sum_{i=1}^n \varepsilon_i \right| = 4\mathbb{E}_\mu H + O\left(\frac{1}{\sqrt{n}}\right).$$

Combining this with Talagrand's inequality yields sharp estimates on the uGC sample complexity.

Theorem 2.10 Let μ be a probability measure on Ω , set $F \subset B(L_\infty(\Omega))$ and put $\sigma^2 = \sup_{f \in F} \text{var}(f(X_1))$. There is an absolute constant $C \geq 1$ such that for every $x > 0$, there is a set of probability larger than $1 - e^{-x}$ on which

$$\sup_{f \in F} \left| \mathbb{E}_\mu f - \frac{1}{n} \sum_{i=1}^n f(X_i) \right| \leq \frac{4R_n(F, \mu)}{\sqrt{n}} + C \left(\sigma \sqrt{\frac{x}{n}} + \frac{x}{n} \right). \quad (2.3)$$

In particular, there is an absolute constant C such that if

$$n \geq \frac{C}{\varepsilon^2} \max \left\{ R_n^2(F, \mu), \log \frac{1}{\delta} \right\},$$

then $\Pr \left\{ \sup_{f \in F} \left| \mathbb{E}_\mu f - \frac{1}{n} \sum_{i=1}^n f(X_i) \right| \geq \varepsilon \right\} \leq \delta$.

We will show in the following sections that although the Glivenko-Cantelli sample complexity is governed by the global Rademacher averages, the learning sample complexity can be controlled by several local versions, one of which is

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

where $F_r = \{h = \lambda f \mid f \in F, 0 \leq \lambda \leq 1, \mathbb{E}_\mu h^2 \leq r\}$. In other words, F_r is the intersection of the *star-shaped hull* of F and 0 with an $L_2(\mu)$ ball of radius \sqrt{r} .

It is evident that the averages associated with every fixed sample $(X_i)_{i=1}^n$ can be written as $\mathbb{E}_\varepsilon \left\| \sum_{i=1}^n \varepsilon_i e_i \right\|_{(F/s_n)^\circ}$, where

$$F/s_n = \left\{ \frac{1}{\sqrt{n}} \sum_{i=1}^n f(X_i) e_i \mid f \in F \right\} \subset \ell_2^n$$

is the natural coordinate projection of F onto ℓ_2^n endowed with the Euclidean structure of $L_2(\mu_n)$, and $(F/s_n)^\circ$ is the polar of F/s_n . This can be interpreted as a decoupling of $R_n(F, \mu)$: the “richness” associated with every sample is determined by the geometry of the coordinate projection F/s_n , and the complexity of the underlying measure μ is the “average size” (with respect to μ) of individual coordinate projections.

Sometimes it is useful to use the gaussian averages instead of the Rademacher ones, i.e.,

$$\mathbb{E}_\mu \mathbb{E}_g \left\| \sum_{i=1}^n g_i e_i \right\|_{(F/s_n)^\circ} = \mathbb{E}_\mu \frac{1}{\sqrt{n}} \mathbb{E}_g \sup_{f \in F} \left| \sum_{i=1}^n g_i f(X_i) \right|,$$

where $(g_i)_{i=1}^n$ are independent standard gaussian random variables. The gaussian average associated with a fixed coordinate projection is simply the ℓ -norm of the polar body of the projected class.

Recall that for any index set T , the expectation of the supremum of the gaussian process indexed by T dominates the expectation of the supremum of

the Rademacher process (up to an absolute multiplicative constant), and thus the gaussian averages may be used instead of the Rademacher ones to obtain upper bounds.

The reason for the normalizing factor of $1/\sqrt{n}$ in the definition of the random averages (instead of the more transparent $1/n$) becomes clearer in the gaussian case. If $T \subset \ell_2$, the isonormal process indexed by T is the gaussian process which has the covariance structure endowed by the inner product in ℓ_2 , i.e., for every $t_1, t_2 \in T$, $\mathbb{E}X_{t_1}X_{t_2} = \langle t_1, t_2 \rangle$. Under the $1/\sqrt{n}$ normalization, the gaussian process indexed by F/s_n is the isonormal process indexed by F when considered as a subset of $L_2(\mu_n)$, since the covariance structure of the process is endowed by the inner product in $L_2(\mu_n)$. As such, all the usual tools of gaussian processes, and in particular, the Dudley-Sudakov inequalities may be applied, and the underlying metric entropy is the natural empirical L_2 entropy.

Note that Theorem 2.10 is measure dependent. Hence, to obtain uniform estimates one would require measure independent estimates on the Rademacher averages $R_n(F, \mu)$.

3 Uniform measures of complexity

As stated above, the sample complexity problem has two determining factors. The first is the probability measure according to which the samples are selected and the other is the geometry of coordinate projections of F . The probability measure determines which samples, or equivalently - which projections are encountered frequently and thus influence the “size” of the class with respect to that measure. To analyze the worst case scenario (which would hold for all probability measures) one has to estimate the size of the largest coordinate projection of F .

In this section we investigate various geometric parameters of the projected sets which are uniform with respect to the measure. In section 5 we present measure dependent bounds which can be computed using the given sample. The notions of size we focus on are the random averages, the combinatorial dimension and the metric entropy.

3.1 Metric entropy and the combinatorial dimension

The interest in the metric entropy is two-fold. Firstly, the “classical” approach shows that the uniform metric entropy $N_p(\varepsilon, F)$ is essential in obtaining the (loose) complexity bound in Theorem 2.2. Secondly, the Rademacher/gaussian averages associated with the class (which lead to sharper bounds) can be estimated via Dudley’s entropy integral. Indeed, for every empirical measure μ_n supported on the sample $s_n = (X_1, \dots, X_n)$,

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

where C is an absolute constant.

Below, we present estimates on the uniform L_2 entropy of a class based on its *combinatorial dimension*.

3.1.1 Binary valued classes

Let F be a class of $\{0, 1\}$ -valued functions. One possible way of measuring how “rich” a class is, is by identifying the largest dimension of a combinatorial cube that can be found in a coordinate projection of F . This combinatorial dimension was introduced by Vapnik and Chervonenkis and is known as the VC dimension.

Definition 3.1 *Let F be a class of $\{0, 1\}$ -valued functions on a space Ω . We say that F shatters $\{x_1, \dots, x_n\} \subset \Omega$, if for every $I \subset \{1, \dots, n\}$ there is a function $f_I \in F$ for which $f_I(x_i) = 1$ if $i \in I$ and $f_I(x_i) = 0$ if $i \notin I$. Let*

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

$\text{vc}(F, \Omega)$ is called the VC dimension of F , but when the underlying space is clear we denote the Vapnik-Chervonenkis dimension by $\text{vc}(F)$.

It is easy to see that $s_n = \{x_1, \dots, x_n\}$ is shattered if and only if

$$\{(f(x_1), \dots, f(x_n)) \mid f \in F\} = \{0, 1\}^n.$$

For every sample σ denote by $P_\sigma F$ the coordinate projection of F , that is,

$$P_\sigma F = \{(f(x_i))_{x_i \in \sigma} \mid f \in F\},$$

and thus, the VC dimension is the largest cardinality of $\sigma \subset \Omega$ such that $P_\sigma F$ is the combinatorial cube of dimension $|\sigma|$.

The following lemma is known as the Sauer-Shelah Lemma ([54, 56, 63], see also [30] for a generalization of that result).

Lemma 3.2 *Let F be a class of binary valued functions and set $d = \text{vc}(F)$. Then, for every finite subset $\sigma \subset \Omega$ of cardinality n ,*

$$|P_\sigma F| \leq \sum_{i=0}^n \binom{n}{i} \leq \left(\frac{en}{d}\right)^d,$$

where the last inequality holds for $n \geq d$. In particular, for every $\varepsilon > 0$, $N(\varepsilon, F, L_\infty(\sigma)) \leq |P_\sigma F| \leq (en/d)^d$.

Using the Sauer-Shelah Lemma one can characterize the uniform Glivenko-Cantelli property of classes of binary valued functions in terms of the VC dimension.

Theorem 3.3 *A class of binary valued functions is a uniform Glivenko-Cantelli class if and only if it has a finite VC dimension.*

Proof: Assume that $\text{vc}(F) = \infty$ and fix an integer $d \geq 2$. Then, there is a set $\sigma \subset \Omega$, $|\sigma| = d$ such that $P_\sigma F = \{0, 1\}^d$, and let μ^σ be the uniform measure on σ (assigns a weight of $1/d$ to every point). For any $A \subset \sigma$ of cardinality $n \leq d/2$, let μ_n^A be the empirical measure supported on A . Since there is some $f_A \in F$ which is 1 on A and vanishes on $\sigma \setminus A$ then $|\mathbb{E}_{\mu^\sigma} f_A - \mathbb{E}_{\mu_n^A} f_A| = |1 - n/d| \geq 1/2$, and in particular, for any empirical measure μ_n supported on a subset of σ of cardinality smaller than $d/2$,

$$\sup_{f \in F} |\mathbb{E}_{\mu^\sigma} f - \mathbb{E}_{\mu_n} f| \geq \frac{1}{2}.$$

Hence, for any $n \leq d/2$,

$$\sup_{\mu} Pr \left\{ \sup_{f \in F} |\mathbb{E}_{\mu_n} f - \mathbb{E}_{\mu} f| \geq 1/2 \right\} = 1,$$

and since d can be made arbitrarily large, F is not a uGC class.

To prove the converse, recall that for every $0 < \varepsilon < 1$ and every empirical measure μ_n supported on a set σ of cardinality $n \geq d$, $N(\varepsilon, F, L_\infty(\sigma)) \leq |P_\sigma F| \leq (en/d)^d$. Since the empirical L_1 entropy is bounded by the empirical L_∞ one, $\log N_1(\varepsilon, F, n) \leq d \log(en/d)$. Thus, for every $\varepsilon > 0$, $\log N_1(\varepsilon, F, n) = o(n)$, showing that F is a uGC class. ■

Although the L_∞ entropy estimate depends on n and thus on the dimension of the coordinate projection, it is possible to derive dimension free L_p entropy bounds for $1 \leq p < \infty$. The first such bound was proved by Dudley [16] and is based on a combination of an extraction principle and the Sauer-Shelah Lemma. The extraction argument shows that if $K \subset F$ is “well separated” in $L_1(\mu_n)$ in the sense that every two points are different on a number of coordinates which is proportional to n , one can find a much smaller set of coordinates (whose size depends of the cardinality of K) onto which the coordinate projection of K is one-to-one.

Recall that a set is ε -separated with respect to a metric d if the distance between every two distinct points in the set is larger than ε . It is easy to see that the cardinality of a maximal ε -separated subset of F (denoted by $D(\varepsilon, F, d)$) is equivalent to the covering numbers of F , namely, for every $\varepsilon > 0$, $N(\varepsilon, F, d) \leq D(\varepsilon, F, d) \leq N(\varepsilon/2, F, d)$.

Theorem 3.4 *Let F be a class of binary valued functions and assume that $\text{vc}(F) = d$. Then, for every $1 \leq p < \infty$ and every $0 < \varepsilon < 1$,*

$$N_p(\varepsilon, F) \leq \left(c_p \log \frac{2}{\varepsilon} \right)^d \left(\frac{1}{\varepsilon} \right)^{pd}.$$

Proof: Since the functions in F are $\{0, 1\}$ -valued it is enough to prove the claim for $p = 1$. The general case follows since for any $f, g \in F$ and any probability measure μ , $\|f - g\|_{L_p(\mu)}^p = \|f - g\|_{L_1(\mu)}$.

Fix x_1, \dots, x_n and $0 < \varepsilon < 1$, and let $\mu_n = n^{-1} \sum_{i=1}^n \delta_{x_i}$. Set K_ε to be any ε -separated subset of F with respect to the $L_1(\mu_n)$ norm and denote its cardinality by D .

Let $V = \{f_i - f_j \mid f_i \neq f_j \in K_\varepsilon\}$; thus, $|V| \leq D^2$ and since K_ε is ε -separated then every $v \in V$ has at least $n\varepsilon$ coordinates which belong to $\{-1, 1\}$.

Set $(X_i)_{i=1}^t$ to be independent $\{x_1, \dots, x_n\}$ -valued random variables, such that for every $1 \leq i \leq t$ and $1 \leq j \leq n$, $Pr(X_i = x_j) = 1/n$. For any $v \in V$, $Pr(\{\forall i, v(X_i) = 0\}) = \prod_{i=1}^t Pr(\{v(X_i) = 0\}) \leq (1 - \varepsilon)^t$, implying that

$$Pr(\{\exists v \in V, \forall i, v(X_i) = 0\}) \leq |V|(1 - \varepsilon)^t \leq D^2(1 - \varepsilon)^t.$$

If $\sigma = (X_1, \dots, X_t)$ then

$$Pr(\{P_\sigma K_\varepsilon \text{ is one-to-one}\}) \geq 1 - D^2(1 - \varepsilon)^t,$$

and if the latter is greater than 0, there is a set $\sigma \subset \{1, \dots, n\}$ such that $|\sigma| \leq t$ and

$$|P_\sigma K_\varepsilon| = |\{(f(x_i))_{i \in \sigma} \mid f \in K_\varepsilon\}| = D.$$

Selecting $t = \frac{2 \log D}{\varepsilon}$ suffices to ensure the existence of such a set σ . By the Sauer-Shelah Lemma,

$$D = |P_\sigma K_\varepsilon| \leq |P_\sigma F| \leq \left(\frac{e|\sigma|}{d}\right)^d \leq \left(\frac{2e \log D}{d\varepsilon}\right)^d. \quad (3.1)$$

To complete the proof, note that if $\alpha \geq 1$ and $\alpha \log^{-1} \alpha \leq \beta$ then $\alpha \leq \beta \log(e\beta \log \beta)$. By (3.1), $\log D \leq d \log\left(\frac{2e^2}{d\varepsilon} \log\left(\frac{2e}{d\varepsilon}\right)\right)$, as claimed. ■

This result was strengthened by Haussler in [23] (see also [61] for a different exposition).

Theorem 3.5 *There are constants C_p which satisfy that for every class F of binary valued functions with $vc(F) = d$, any $1 \leq p < \infty$ and every $0 < \varepsilon < 1$,*

$$N_p(\varepsilon, F) \leq C_p d (4e)^d \varepsilon^{-pd}.$$

3.1.2 Real valued classes

Combinatorial parameters associated with real-valued classes can be obtained by extending the notion of a shattered set. Unlike the $\{0, 1\}$ case, where every cube found in a coordinate projection of the class is the combinatorial cube of the appropriate dimension, in the real-valued case the size of the cube found in the projection is very important. The combinatorial dimension measures the tradeoff between the size of such a cube and its dimension.

Definition 3.6 *For every $\varepsilon > 0$, a set $\sigma = \{x_1, \dots, x_n\} \subset \Omega$ is said to be ε -shattered by F if there is some function $s : \sigma \rightarrow \mathbb{R}$, such that for every*

$I \subset \{1, \dots, n\}$ there is some $f_I \in F$ for which $f_I(x_i) \geq s(x_i) + \varepsilon$ if $i \in I$, and $f_I(x_i) \leq s(x_i) - \varepsilon$ if $i \notin I$. Let

$$\text{vc}(F, \Omega, \varepsilon) = \sup \{|\sigma| \mid \sigma \subset \Omega, \sigma \text{ is } \varepsilon\text{-shattered by } F\}.$$

f_I is called the *shattering function* of the set I and the set $\{s(x_i) \mid x_i \in \sigma\}$ is called a *witness to the ε -shattering*. In cases where the underlying space is clear we denote the combinatorial dimension by $\text{vc}(F, \varepsilon)$.

Observe that if F is convex and symmetric and if $\sigma = \{x_1, \dots, x_n\}$ is ε -shattered by F , then

$$\varepsilon B_\infty^n \subset \{(f(x_1), \dots, f(x_n)) \mid f \in F\}.$$

Indeed, if σ is ε -shattered with a witness $(s(x_i))_{i=1}^n$ then for every $I \subset \{1, \dots, n\}$ there is some f_I such that $f_I(x_i) \geq s(x_i) + \varepsilon$ if $i \in I$ and $f_I(x_i) \leq s(x_i) - \varepsilon$ for $i \in I^c$. For every such I , define $g_I = (f_I - f_{I^c})/2$ and since F is convex and symmetric, $g_I \in F$. It follows that $\{(f(x_1), \dots, f(x_n)) \mid f \in F\} \supset \varepsilon B_\infty^n$, as claimed.

The first estimate on the empirical L_∞ covering numbers in terms of the combinatorial dimension was established in [1] and was used to show that a class of uniformly bounded functions is a uGC class if and only if it has a finite combinatorial dimension for every ε . The proof that if F is a uGC class it has a finite combinatorial dimension for every ε follows from a similar argument to the one used in the binary valued case. For the converse, one requires empirical L_∞ entropy estimates combined with Theorem 2.2.

Other results on the L_∞ and L_p entropy estimates in terms of the combinatorial dimension were derived in [51, 4, 39, 58, 60].

The following are the best known bounds on the uniform L_2 and L_∞ entropy in terms of the combinatorial dimension. The L_p estimates are from [47] and the L_∞ ones can be found in [53].

Theorem 3.7 *There are constants K_p and c_p depending only on p such that the following holds. For every $F \subset B(L_\infty(\Omega))$, every probability measure μ , every $1 \leq p < \infty$ and any $0 < \varepsilon < 1$,*

$$N(\varepsilon, F, L_p(\mu)) \leq \left(\frac{2}{\varepsilon}\right)^{K_p \text{vc}(F, c_p \varepsilon)}.$$

Moreover, for any $0 < \varepsilon, \delta < 1$ and every set $\sigma \subset \Omega$, $|\sigma| = n$,

$$\log N(\varepsilon, F, L_\infty(\sigma)) \leq K \cdot \text{vc}(F, c\varepsilon\delta) \log^{1+\delta} \left(\frac{n}{\delta\varepsilon}\right),$$

where K and c are absolute constants.

It is easy to show that the L_p bounds are optimal (up to the absolute constants) and that the optimal L_∞ bound (as a function of ε and n) is proportional to $\text{vc}(F, \varepsilon) \cdot \log(2n/\varepsilon)$.

By selecting empirical measures supported on the shattered sets, the uniform entropy can be bounded from below, demonstrating the sharpness of the upper bounds (up to a single power of a logarithm).

Theorem 3.8 *There are absolute constants K, K', c, c' such that for any $F \subset B(L_\infty(\Omega))$ and every $0 < \varepsilon < 1$,*

$$K' \cdot \text{vc}(F, c'\varepsilon) \leq \log N_2(\varepsilon, F) \leq K \cdot \text{vc}(F, c\varepsilon) \log\left(\frac{2}{\varepsilon}\right).$$

3.2 Random averages and the combinatorial dimension

In the past, the only known way to obtain complexity estimates was by bounding the uniform entropy via the combinatorial dimensions. Thus, much effort was put into the investigation of the combinatorial dimension of many interesting classes (see e.g. [2]).

It was noted in [39] that there are nontrivial connections between the combinatorial dimension and the Rademacher/gaussian averages associated with a class, and that the latter could be used to obtain sharper complexity bounds. To estimate the random averages using the combinatorial dimension, we introduce a uniform version of the gaussian averages. Let

$$\ell_n(F) \equiv \sup_{s_n} \ell(F/s_n) = \sup_{\{x_1, \dots, x_n\} \subset \Omega} \frac{1}{\sqrt{n}} \mathbb{E} \sup_{f \in F} \left| \sum_{i=1}^n g_i f(x_i) \right|,$$

that is, the largest gaussian average associated with a coordinate projection on n points. Since $F \subset B(L_\infty(\Omega))$ then $F/s_n \subset n^{-1/2}B_\infty^n$, and the largest projection one might encounter is when $F/s_n = n^{-1/2}B_\infty^n$, in which case $\ell(F/s_n) \sim \sqrt{n}$. We define a scale-sensitive parameter which measures for every $\varepsilon > 0$ the largest cardinality of a projection which has a “large” ℓ -norm. Let

$$t(F, \varepsilon) = \sup\{n \in \mathbb{N} \mid \ell_n(F) \geq \varepsilon\sqrt{n}\}.$$

$t(F, \varepsilon)$ was introduced in [39] where its connection to the uniform Glivenko-Cantelli condition and complexity estimates were investigated.

One can show that if $\ell(F/s_n) \sim \varepsilon\sqrt{n}$, there exists a set $\sigma \subset s_n$ of cardinality proportional to n , which is $c\varepsilon$ shattered by F , and thus the two “worst case” scale-sensitive parameters - the combinatorial dimension and $t(F, \varepsilon)$ are comparable.

Theorem 3.9 [48] *There are absolute constants K and c such that for any $F \subset B(L_\infty(\Omega))$ and every $\varepsilon > 0$,*

$$\text{vc}(F, c'\varepsilon) \leq t(F, \varepsilon) \leq K \frac{\text{vc}(F, c\varepsilon)}{\varepsilon^2}.$$

This result can be used to estimate the combinatorial dimension of a convex hull of a class.

Corollary 3.10 [48] *There are absolute constants K and c such that for any $F \subset B(L_\infty(\Omega))$ and every $\varepsilon > 0$,*

$$\text{vc}(\text{conv}(F), \varepsilon) \leq \frac{K \cdot \text{vc}(F, c\varepsilon)}{\varepsilon^2}.$$

Proof. Since the ℓ -norm of a set and of its convex hull are the same, then for any $\varepsilon > 0$, $t(F, \varepsilon) = t(\text{conv}(F), \varepsilon)$. By Theorem 3.9,

$$\text{vc}(\text{conv}(F), \varepsilon) \leq t(\text{conv}(F), \varepsilon) = t(F, \varepsilon) \leq \frac{K \text{vc}(F, c\varepsilon)}{\varepsilon^2}.$$

■

It is possible to prove sharper estimates on the combinatorial dimension of convex hulls of F if one knows more on the structure of the class. For example, if F is a VC class of functions, then one has “global” information - the uniform entropy of the class at every scale ε . In the general case presented above, the information is “local” - the combinatorial dimension at a single scale, in which case it is possible to construct an example proving that the estimate in Corollary 3.10 is sharp. Indeed, in [44] it was shown that if $K_{n,N} \subset B_\infty^n$ is the symmetric convex hull of a random $\{0, 1\}$ polytope $F_{n,N}$ with N vertices, and if $N \geq cn^2$ then

$$\text{vc}(K_{n,N}, \varepsilon) \sim \min \left\{ \frac{\log N \varepsilon^2}{\varepsilon^2}, n \right\}.$$

One can also show that $\text{vc}(F_{n,N}) \sim \log N$ from which the sharpness of Corollary 3.10 follows.

3.3 Phase transitions in GC classes

The final remark in the previous section reveals a very significant (though not surprising) fact. If one is interested in scale-sensitive parameters (e.g. the uniform entropy) one can obtain much sharper bounds if one has some global information on the class. In this section we relate the combinatorial dimension to the random averages, given the combinatorial dimension at every scale. We focus on classes which have a power type p , that is, there exists a constant C such that for every $0 < \varepsilon < 1$, $\text{vc}(F, \varepsilon) \leq C\varepsilon^{-p}$.

Although it is natural to think that parameters which describe the complexity of uniform Glivenko-Cantelli classes behave smoothly in terms of the power type, there is a clear phase transition which occurs at $p = 2$. In some sense, if the class has power type p for $0 < p < 2$ it is “small”, whereas if $p > 2$ the class is considerably larger. We will demonstrate this phenomenon in several examples.

3.3.1 The uniform Central Limit Theorem

The first example of a phase transition can be seen in a uniform version of the Central Limit Theorem.

Definition 3.11 [18] *Let $F \subset B(L_\infty(\Omega))$, set μ to be a probability measure on Ω and assume G_μ to be a gaussian process indexed by F , which has mean 0 and covariance*

$$\mathbb{E}G_\mu(f)G_\mu(g) = \int fg d\mu - \int f d\mu \int g d\mu.$$

F is called a *universal Donsker class* if for any probability measure μ , the law G_μ is tight in $\ell_\infty(F)$ and $\nu_n^\mu = n^{1/2}(\mu_n - \mu) \in \ell_\infty(F)$ converges in law to G_μ in $\ell_\infty(F)$, where μ_n a random empirical measure selected according to μ .

Stronger than the universal Donsker property is the uniform Donsker property. For such classes, ν_n^μ converges to G_μ uniformly in μ in some sense (see [18, 61] for more details). The following result of Giné and Zinn [28] is a relatively simple characterization of uniform Donsker classes.

For every probability measure μ on Ω , let $\rho_\mu^2(f, g) = \mathbb{E}_\mu(f-g)^2 - (\mathbb{E}_\mu(f-g))^2$, and set for every $\delta > 0$, $F_\delta = \{f - g | f, g \in F, \rho_\mu(f, g) \leq \delta\}$.

Theorem 3.12 [28] *A class F is a uniform Donsker class if and only if the following holds. For every probability measure μ on Ω , G_μ has a version with bounded, ρ_μ -uniformly continuous sample paths, and for these versions,*

$$\sup_\mu \mathbb{E} \sup_{f \in F} |G_\mu(f)| < \infty, \quad \lim_{\delta \rightarrow 0} \sup_\mu \mathbb{E} \sup_{h \in F_\delta} |G_\mu(h)| = 0.$$

The main tool in the analysis of uniform Donsker classes is the Koltchinskii-Pollard entropy integral.

Theorem 3.13 [18] *If $F \subset B(L_\infty(\Omega))$ satisfies that*

$$\int_0^\infty \sup_n \sup_{\mu_n} \sqrt{\log N(\varepsilon, F, L_2(\mu_n))} d\varepsilon < \infty,$$

then it is a uniform Donsker class, where μ_n are empirical measures supported on at most n points.

On the other hand one can show [18] that if F is a uniform Donsker class then there is some constant C such that for every $\varepsilon > 0$,

$$\sup_n \sup_{\mu_n} \log N(\varepsilon, F, L_2(\mu_n)) \leq \frac{C}{\varepsilon^2}. \quad (3.2)$$

Therefore, the gap between the two estimates is at most logarithmic.

Corollary 3.14 [39] *Let $F \subset B(L_\infty(\Omega))$. If there is some constant C such that $\text{vc}(F, \varepsilon) \leq C\varepsilon^{-p}$ for $0 < p < 2$ then F is a uniform Donsker class. On the other hand, if $\text{vc}(F, \varepsilon) \geq c\varepsilon^{-p}$ for $p > 2$ then F is not a uniform Donsker class.*

Proof: The first part of our claim follows from Lemma 3.7, since F satisfies the Koltchinskii-Pollard entropy condition. For the second part, recall that by Theorem 3.8, $N_2(F, \varepsilon) \geq C/\varepsilon^p$ for some $p > 2$. On the other hand, if F is a uniform Donsker class then $N_2(F, \varepsilon) = O(\varepsilon^{-2})$ which is a contradiction. ■

3.3.2 Uniform ℓ -norm estimates

Another clear phase transition can be seen in the asymptotic behavior of $\ell_n(F)$.

Theorem 3.15 *Let $F \subset B(L_\infty(\Omega))$ and assume that there is some $\gamma > 0$ such that for any $\varepsilon > 0$, $\text{vc}(F, \varepsilon) \leq \gamma\varepsilon^{-p}$. Then, there are constants C_p which depend only on p , such that*

$$\ell_n(F) \leq \gamma^{1/2} C_p \begin{cases} 1 & \text{if } 0 < p < 2, \\ \log^{3/2} n & \text{if } p = 2, \\ n^{1/2-1/p} \log^{1/p} n & \text{if } p > 2. \end{cases}$$

The proof is based on Dudley's entropy integral bound for $0 < p < 2$ and on a discrete version of that theorem for $p \geq 2$, combined with the uniform entropy estimate (see [39, 43]). Note that in the case $p < 2$ and $p > 2$ the bound is tight up to a logarithmic factor. This is obvious for $p < 2$. As for $p > 2$, if $\text{vc}(F, \varepsilon) = \gamma\varepsilon^{-p}$ then $t(F, \varepsilon) \geq K\gamma\varepsilon^{-p}$ for some absolute constant K . Hence, there is an integer $n \geq K\gamma\varepsilon^{-p}$ and a set s_n of cardinality n such that $\ell(F/s_n) \geq \varepsilon\sqrt{n} \geq (K\gamma)^{\frac{1}{p}} n^{\frac{1}{2}-\frac{1}{p}}$. The exact growth rate in the case $p = 2$ is not known.

Combining Theorem 3.15 with Theorem 2.10, we obtain the following estimate on the uGC sample complexity, which, as lower bounds show (see, e.g. [2]) is optimal rate wise, up to logarithmic factors.

Corollary 3.16 *Let $F \subset B(L_\infty(\Omega))$ and assume that $\text{vc}(F, \varepsilon) \leq \gamma\varepsilon^{-p}$. Then, there is a constant $C_{p,\gamma}$ such that*

$$S_F(\varepsilon, \delta) \leq C_{p,\gamma} \max\left\{\frac{1}{\varepsilon^p}, \frac{1}{\varepsilon^2} \log \frac{1}{\delta}\right\}$$

if $p \neq 2$. If $p = 2$ there is an additional logarithmic factor in $\frac{1}{\varepsilon}$.

In particular, if each $g \in G$ maps Ω into $[0, 1]$ and $\text{vc}(G, \varepsilon) \leq \gamma\varepsilon^{-p}$ for some $p \neq 2$, then

$$\sup_{T:\Omega \rightarrow [0,1]} S_{\mathcal{L}_q(G,T)}(\varepsilon, \delta) \leq C_{p,q,\gamma} \max\left\{\frac{1}{\varepsilon^p}, \frac{1}{\varepsilon^2} \log \frac{1}{\delta}\right\}.$$

The proof of the latter part of the corollary follows from the fact that for every probability measure μ , $N(\varepsilon, \mathcal{L}_q(G, T), L_2(\mu)) \leq N(\varepsilon/q, G, L_2(\mu))$. Thus, the same asymptotic estimates on $\ell_n(G)$ hold for $\ell_n(\mathcal{L}_q(G, T))$, regardless of the specific selection of T .

Recall that the sample complexity bound obtained by the "classical" approach (Corollary 2.4) was $O(\varepsilon^{-(2+p)})$, which is considerably worse than the above result.

3.3.3 On the size of convex hulls

Investigating the structure of convex hulls of classes has a natural origin in Machine Learning. In each learning procedure one has to solve, in one way or

another, an optimization problem (locating an “almost” empirical minimizer). This is a reasonable task when the class is convex. Apriori, taking the convex hull of a nonconvex class might make it too complex for an efficient solution from the probabilistic point of view, in the sense that the sample complexity becomes huge. From the definition of the Glivenko-Cantelli condition it is clear that the uGC sample complexity of a class and of its convex hull are the same, but this is no longer true for the learning sample complexity, and thus it is important to at least obtain upper bounds on the way the combinatorial dimension, the metric entropy and the localized random averages grow when one takes the convex hull of F .

There are sharp estimates on the growth rate of the L_2 entropy of the convex hull of a set, given information on the growth rate of the entropy of the given set. These results are summarized in the next theorem, which is divided into two parts. The first is when the covering numbers of the class are polynomial in $1/\varepsilon$ while in the second they are exponential.

Theorem 3.17 *Let $F \subset B(L_\infty(\Omega))$ and set H to be the convex hull of F .*

1. *If there are $\gamma, p > 0$ such that $N(\varepsilon, F, L_2(\mu)) \leq \gamma\varepsilon^{-p}$ for every $\varepsilon > 0$, then there is an absolute constant C such that for every $\varepsilon > 0$,*

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

2. *If there are $\gamma > 0$ and p such that $\log N(\varepsilon, F, L_2(\mu)) \leq \gamma\varepsilon^{-p}$ for every $\varepsilon > 0$, then there is some constant $C_{p,\gamma}$ such that*

$$\log N(\varepsilon, H, L_2(\mu)) \leq C_{p,\gamma} \begin{cases} \varepsilon^{-2} \log^{1-2/p}(1/\varepsilon) & \text{if } 0 < p < 2 \\ \varepsilon^{-2} \log^2(1/\varepsilon) & \text{if } p = 2 \\ \varepsilon^{-p} & \text{if } p > 2, \end{cases}$$

and all the estimates are sharp (up to the constants).

The fact that the rate of the entropy in part (1) is $\varepsilon^{-2p/(p+2)}$ was established by Dudley [17], and the constants are taken from [38]. Part (2) for $p \neq 2$ was proved in [12] (see [38] for a different proof), and for $p = 2$ in [26]. Another proof for the complete range $0 < p < \infty$ may be found in [11].

As in the previous examples, the phase transition is apparent. If $0 < p < 2$ the entropy of the convex hull grows at a rate which is slightly better than $1/\varepsilon^2$. For “large” classes the power of the entropy does not change. Observe that there is a slight looseness in the analysis of Donsker classes using the entropy. If F is a Donsker class then its convex hull is also Donsker [18]. On the other hand, for any $2/3 < p < 2$ there is a probability measure μ and class $F \subset L_2(\mu)$ for which the $L_2(\mu)$ entropy is $O(\varepsilon^{-p})$, but the entropy of the convex hull could be of the order of $\varepsilon^{-2} \log^{1-2/p}(1/\varepsilon)$. In particular, the entropy integral diverges.

As a corollary to Theorem 3.17, we improve the estimates on the combinatorial dimension of convex hulls. For example, let F be a binary valued class with $\text{vc}(F) = d$. Then, for any probability measure μ

$$\log N(\varepsilon, \text{conv}(F), L_2(\mu)) \leq Cd \left(\frac{1}{\varepsilon} \right)^{\frac{2d}{1+d}},$$

where C is an absolute constant. In particular, by Theorem 3.8

$$\text{vc}(\text{conv}(F), \varepsilon) \leq K \cdot \log N_2(c\varepsilon, \text{conv}(F)) \leq K \cdot d \left(\frac{1}{\varepsilon} \right)^{\frac{2d}{1+d}} = K \cdot \frac{\text{vc}(F)}{\varepsilon^{2d/d+1}},$$

in which the exponent of ε is strictly smaller than 2.

Similar results can be established for classes with $\text{vc}(F, \varepsilon) = O(\varepsilon^{-p})$, namely, that $\text{vc}(\text{conv}(F), \varepsilon) = O(\varepsilon^{-\max\{2, p\}})$ for $p \neq 2$, while for $p = 2$ there is an additional logarithmic factor in $1/\varepsilon$. This estimate is better than the one in Corollary 3.10, because one has “global” information on the given class.

3.4 Concentration of the combinatorial dimension

This section is devoted to the concentration of measure phenomenon for the combinatorial dimension. The fact that the combinatorial dimension is highly concentrated (in a sense which will be explained below) is rather surprising and is due to Boucheron, Lugosi and Massart [9]. It is based on the following exponential inequality for general functions.

Theorem 3.18 [9] *Let X_1, \dots, X_n be independent random variables defined on some set Ω and set $Z = f(X_1, \dots, X_n)$ to be a random variable. Assume that there is a function $g : \Omega^{n-1} \rightarrow \mathbb{R}$ such that for all $\{x_1, \dots, x_n\} \subset \Omega$,*

1. $0 \leq f(x_1, \dots, x_n) - g(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n) \leq 1$ for all $1 \leq i \leq n$,
2. $\sum_{i=1}^n (f(x_1, \dots, x_n) - g(x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_n)) \leq f(x_1, \dots, x_n)$,

then, for every $t > 0$,

$$\Pr\{Z \geq \mathbb{E}Z + t\} \leq \exp\left[-\frac{t^2}{2\mathbb{E}Z + 2t/2}\right],$$

$$\Pr\{Z \leq \mathbb{E}Z - t\} \leq \exp\left[-\frac{t^2}{2\mathbb{E}Z}\right].$$

For any fixed $\varepsilon > 0$ let $Z = \text{vc}(F, \{X_1, \dots, X_n\}, \varepsilon)$ be the combinatorial dimension of the class F at scale ε on the random set $\{X_1, \dots, X_n\}$, where $(X_i)_{i=1}^n$ are independent, distributed according to a measure μ . It is easy to see that by setting $g(x_1, \dots, x_{n-1}) = \text{vc}(F, \{x_1, \dots, x_{n-1}\}, \varepsilon)$, both conditions of Theorem 3.18 are satisfied. Indeed, removing one point can change the combinatorial dimension by at most one. By the same argument, the number of removals of points in $\{x_1, \dots, x_n\}$ which change the combinatorial dimension is exactly $\text{vc}(F, \{x_1, \dots, x_n\}, \varepsilon)$. Thus, both conditions are verified and $Z = \text{vc}(F, \{X_1, \dots, X_n\}, \varepsilon)$ satisfies the required concentration inequality.

Some other applications of Theorem 3.18 can be found in [8, 9].

It is natural to ask whether the two other measures of complexity, the Rademacher averages and the L_p empirical entropy, display similar concentration phenomena. The first can be verified, simply by applying Talagrand's inequality (a fact which will be used in section 7). Whether the random variable $Z = \log N(\varepsilon, F, L_p(\mu_n))$ is concentrated is not known in general. As the following example [45] shows, the L_∞ log-covering numbers when the centers are taken from the set are not concentrated, in the sense that the best estimate on the concentration of the covering numbers is given by Markov's inequality. More precisely, Let $\tilde{N}(K, \varepsilon B_\infty^n)$ be the covering numbers of K by translates of εB_∞^n , centered at point in K . Then, we have the following

Theorem 3.19 *There exists an absolute constant C for which the following holds. For any $t \geq 2$ and $\varepsilon > 0$ there exist $N, n \in \mathbb{N}$ and a set $K \subset B_\infty^N$ such that the random variable*

$$Z(X_1, \dots, X_n) = \log \tilde{N}(P_\sigma K, \varepsilon B_\infty^n)$$

satisfies $\mathbb{E}Z \leq 2 \log t$ and

$$Pr(Z \geq t) \leq \frac{C}{t},$$

where $(X_i)_{i=1}^n$ are independent, distributed according to the uniform measure on $\{1, \dots, N\}$.

Proof. Without loss of generality assume that $d = t/\log 2 \in \mathbb{N}$. Set $N = 20d^2$ and let

$$K_1 = \left\{ \left\{ -\frac{2\varepsilon}{3}, \frac{2\varepsilon}{3} \right\}^d \times (0, \dots, 0) \right\} \subset B_\infty^N$$

and $K_2 = \{e_1, \dots, e_d\}$ where $(e_i)_{i=1}^d$ is the standard unit vector basis in \mathbb{R}^d . Define $K = K_1 \cup K_2$, observe that if $\{1, \dots, d\} \subset \sigma$ then all the points in $P_\sigma K$ are ε -separated, and thus $N(P_\sigma K, \varepsilon B_\infty^\sigma) \geq 2^d$. On the other hand, if there is some $i \in \{1, \dots, d\} \setminus \sigma$ then $P_\sigma e_i = 0$, and $\varepsilon B_\infty^\sigma$ is centered at $P_\sigma e_i$ and contains $P_\sigma K$, implying that $N(P_\sigma K, \varepsilon B_\infty^\sigma) \leq d$. Let μ be the uniform probability measure on $\{1, \dots, N\}$ and denote

$$\alpha = \alpha_n(d, N) = Pr(\{1, \dots, d\} \subset \{X_1, \dots, X_n\}).$$

There exists an integer n such that

$$\frac{1}{20d} \leq \alpha \leq \frac{1}{10d}. \quad (3.3)$$

If the inequality (3.3) holds, then $Z \geq \log 2 \cdot d = t$ with probability $\alpha \geq \frac{1}{20d}$, while

$$\mathbb{E}Z \leq \alpha \cdot \log(2^d + d) + (1 - \alpha) \cdot \log d \leq 2 \log t,$$

which implies the Theorem.

To verify (3.3), observe that for fixed integers d and N , $\lim_{n \rightarrow \infty} \alpha_n(d, N) = 1$. Let $\sigma_n = \{X_1, \dots, X_n\}$, set B_n be the event $\{\{1, \dots, d\} \subset \sigma_n\}$ and put $A_i = \{1, \dots, d\} \setminus \{i\}$. Clearly,

$$\begin{aligned} \alpha_{n+1}(d, N) &= Pr(B_{n+1} \cap B_n) + Pr(B_{n+1} \cap B_n^c) \\ &= \alpha_n(d, N) + \sum_{i=1}^d Pr(B_{n+1} \cap \{\{X_1, \dots, X_n\} \cap \{1, \dots, d\} = A_i\}) \\ &= \alpha_n(d, N) + \frac{d}{N} \alpha_n(d-1, N). \end{aligned}$$

so,

$$0 \leq \alpha_{n+1}(d, N) - \alpha_n(d, N) \leq d/N = \frac{1}{20d}.$$

■

4 Learning Sample complexity and error bounds

Bounding the learning sample complexity using the Glivenko-Cantelli condition is not the optimal strategy. In fact, the best that one can hope for using this approach is a sample complexity estimate of $\Omega(1/\varepsilon^2)$ (up to logarithmic factors), since this is the minimal number of examples needed to ensure that $|\mathbb{E}_\mu f - n^{-1} \sum_{i=1}^n f(X_i)| < \varepsilon$ with sufficiently high probability for a single function with a nontrivial variance.

The reason for the sub-optimality of this approach is that the random variable $\sup_{f \in F} |\mathbb{E}_\mu f - n^{-1} \sum_{i=1}^n f(X_i)|$ measures the worst deviation of empirical means from actual ones, going over the entire class; for the learning sample complexity problem, it suffices to control the deviation with respect to functions which are potential empirical minimizers (that is, functions which satisfy that $\mathbb{E}_{\mu_n} f$ is “small”), and this is a relatively small subclass of F . An added complication arises because this is a *random* subset of F .

The ability to obtain improved complexity estimates is based on two essential ingredients. Firstly, the class has to be “relatively small”, otherwise, it is impossible to improve the uGC based bounds. Secondly, the class should have the property that the variance of each class member can be controlled by a power of its expectation.

Definition 4.1 *F is called a Bernstein class of type $0 < \alpha \leq 1$ with respect to the measure μ , if there is some constant B such that for any $f \in F$, $\mathbb{E}_\mu f^2 \leq B(\mathbb{E}_\mu f)^\alpha$.*

Note that this definition is measure dependent. To obtain uniform results one requires the class to have a Bernstein type α for every measure, and with the same constant.

Obviously, if F consists of functions bounded by \sqrt{B} it has “Bernstein type 0” with a constant B . It is also evident that if F consists of nonnegative

functions bounded by B then F has Bernstein type 1 with constant B with respect to any probability measure μ . Therefore, if the target T belongs to G then for any $1 \leq p < \infty$, $\mathcal{L}_p(G, T)$ has Bernstein type 1 with constant 1. In section 6 we show that even if $T \notin G$ and G is convex, then for any probability measure μ , $\mathcal{L}_p(G, T)$ has a Bernstein type 1 for any $1 < p \leq 2$ and Bernstein type $2/p$ if $2 < p < \infty$, and the constants do not depend on the measure μ .

The improved complexity bounds are based on the fact that if F has Bernstein type 1 with a constant B , the learning sample complexity at scale ε is controlled by the GC sample complexity of the intersection of the *star-shaped* hull of F with 0 (i.e., $\{\lambda f | 0 \leq \lambda \leq 1, f \in F\}$) with the ball $\{f | \mathbb{E}_\mu f^2 \leq B\varepsilon\}$. Similar results hold for any class with any nontrivial Bernstein type [41].

Lemma 4.2 [41] *Let μ be a probability measure on Ω and assume that $F \subset B(L_\infty(\Omega))$ has Bernstein type 1 with respect to μ , with a constant $B \geq 1$. Then, for any $\varepsilon > 0$,*

$$\Pr\left\{\exists f \in F, \frac{1}{n} \sum_{i=1}^n f(X_i) \leq \frac{\varepsilon}{2}, \mathbb{E}_\mu f \geq \varepsilon\right\} \leq 2\Pr\left\{\sup_{h \in \text{star}(F, 0), \mathbb{E}_\mu h^2 \leq B\varepsilon} |\mathbb{E}_\mu h - \mathbb{E}_{\mu_n} h| \geq \frac{\varepsilon}{2}\right\},$$

where $\text{star}(F, 0) = \{\lambda f | 0 \leq \lambda \leq 1, f \in F\}$.

Proof: Set $A = \{\exists f \in F, \frac{1}{n} \sum_{i=1}^n f(X_i) \leq \varepsilon/2, \mathbb{E}_\mu f \geq \varepsilon\}$ and observe that

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

because $\mathbb{E}_\mu f \geq \varepsilon$ and $\mathbb{E}_{\mu_n} f \leq \varepsilon/2$ imply that $|\mathbb{E}_\mu f - \mathbb{E}_{\mu_n} f| \geq \frac{1}{2}\mathbb{E}_\mu f \geq \varepsilon/2$. Let $H = \left\{\frac{\varepsilon f}{\mathbb{E}_\mu f} \mid f \in F, \mathbb{E}_\mu f \geq \varepsilon, \mathbb{E}_\mu f^2 \geq \varepsilon\right\}$ which is a subset of $\text{star}(F, 0)$ since $0 < \varepsilon/\mathbb{E}_\mu f \leq 1$. Every $f \in F$ satisfies that $\mathbb{E}_\mu f^2 \leq B\mathbb{E}_\mu f$, and thus for every $h \in H$, $\mathbb{E}_\mu h^2 \leq B\varepsilon$. In particular,

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

which proves our claim. ■

The fact that the learning sample complexity can be interpreted as a GC estimate of a class consisting of functions with variance of the same order of magnitude as the required deviation, yields improved tail bounds by a straightforward application of Talagrand's inequality.

Theorem 4.3 [43] *There is an absolute constant C such that the following holds. Let $F \subset B(L_\infty(\Omega))$ be a class of Bernstein type 1 with a constant $B \geq 1$. Set $\mathcal{H} = \text{star}(F, 0)$ and for every $\varepsilon > 0$ let $\mathcal{H}_\varepsilon = \mathcal{H} \cap \{h | \mathbb{E}_\mu h^2 \leq \varepsilon\}$. Then for every $0 < \varepsilon, \delta < 1$,*

$$Pr\left\{\exists f \in F, \frac{1}{n} \sum_{i=1}^n f(X_i) \leq \varepsilon/2, \mathbb{E}_\mu f \geq \varepsilon\right\} \leq \delta$$

Provided that

$$n \geq C \max\left\{\frac{R_n^2(\mathcal{H}_\varepsilon, \mu)}{\varepsilon^2}, \frac{B \log \frac{2}{\delta}}{\varepsilon}\right\},$$

where $(X_i)_{i=1}^n$ are selected according to μ and recall that

$$R_n(\mathcal{H}_\varepsilon, \mu) = \frac{1}{\sqrt{n}} \mathbb{E} \sup_{h \in \mathcal{H}_\varepsilon} \left| \sum_{i=1}^n \varepsilon_i h(X_i) \right|.$$

4.1 Error Bounds

An alternative presentation to the sample complexity problem is to use *error bounds*. The idea is to ensure that with high probability, every function in the class satisfies that $\mathbb{E}_\mu f \leq \frac{C}{n} \sum_{i=1}^n f(X_i) + \text{“penalty term”}$, which measures the richness of the given class. If $C = 1$, this is simply a one-sided GC condition and the penalty term can not decay faster than σ_F/\sqrt{n} , for $\sigma_F^2 = \sup_{f \in F} \text{var}(f)$. On the other hand, if $C > 1$ one can obtain faster rates. The rate by which the penalty term tends to 0 as a function of n is called the *error rate* and is analogous to the learning sample complexity. Indeed, if the error bound holds for any $f \in F = \mathcal{L}_p(G, T)$, it holds for any function which satisfies that $\mathbb{E}_{\mu_n} \ell_p(g) \leq \varepsilon/2$. For such functions $\mathbb{E}_\mu \ell_p(g) \leq C\varepsilon$ provided that n is large enough to ensure that the penalty term is sufficiently small. It can be shown (see, for example, [20, 2, 14]) that the best error rate possible is $O(1/n)$ (which corresponds to a sample complexity estimate of $O(1/\varepsilon)$), and this holds for “very small” classes. For example, this is the rate of decay (up to a logarithmic factor) for losses associated with binary valued classes with a finite VC dimension.

There are various known error bounds in literature, and modern estimates for the penalty terms are based on the random averages associated with the class. It turns out that even if the class is small, in order to establish error rates faster than $1/\sqrt{n}$ (which can be obtained via a GC-style argument if the class is sufficiently small) one has to use the localized random averages instead of the global ones as penalty terms, as will be explained below. Recently, in [5], a new method of obtaining error bounds was developed, based on the idea of comparing the empirical and the actual structures on the class.

4.2 Comparing structures

One can consider the uniform law of large numbers as a way of comparing the empirical (random) structure and the structure endowed by μ uniformly over

the class. Indeed, a typical statement would be that with probability larger than $1 - \delta$, for every $f \in F$, $|\mathbb{E}_\mu f - n^{-1} \sum_{i=1}^n f(X_i)| \leq \lambda_n(\delta)$. Hence, this condition is an additive notion of similarity between the two structures. It is possible to use a multiplicative notion of similarity - that for “most” functions in the class, a random coordinate projection is a good isomorphism. By this we mean that for a subset $G \subset F$ containing functions with “large” expectations and $0 < \rho < 1$, then with high probability, every $g \in G$ satisfies that

$$(1 - \rho)\mathbb{E}_\mu g \leq \frac{1}{n} \sum_{i=1}^n g(X_i) \leq (1 + \rho)\mathbb{E}_\mu g.$$

In this case we say that the coordinate projection P_σ onto (X_1, \dots, X_n) is a ρ -isomorphism of G . The connection to our problem is simple: suppose that $G_n = \{f \in F : \mathbb{E}f \geq \lambda_n\}$ and assume that P_σ is a ρ -isomorphism on G_n . Then, for every $f \in F$

$$\mathbb{E}f \leq \max \left\{ \frac{1}{1 - \rho} \mathbb{E}_n f, \lambda_n \right\},$$

(since if $\mathbb{E}f \geq \lambda_n$ then f is projected isomorphically onto σ), and this yields the desired error bound. The name ρ -isomorphism comes from the fact that if F consists of nonnegative functions then a projection is a ρ -isomorphism if and only if $P_\sigma : (G, L_1(\mu)) \rightarrow (G, L_1(\mu_n))$ is a bi-Lipschitz function.

In what follows we shall denote

$$\mathbb{E} \|\mu - \mu_n\|_F = \mathbb{E} \sup_{f \in F} \left| \mathbb{E}_\mu f - n^{-1} \sum_{i=1}^n f(X_i) \right|.$$

Theorem 4.4 [5] *There is an absolute constant c for which the following holds. Let F be a class of functions, such that for every $f \in F$, $\mathbb{E}f = \lambda$ and $\|f\|_\infty \leq b$. Assume that F is a Bernstein class of type $0 < \beta < 1$ with a constant B , and suppose that $0 < \rho < 1$ and $0 < \alpha < 1$ satisfy*

$$\lambda \geq c \max \left\{ \frac{bx}{n\alpha^2\rho}, \left(\frac{Bx}{n\alpha^2\rho^2} \right)^{1/(2-\beta)} \right\}.$$

1. *If $\mathbb{E} \|\mu - \mu_n\|_F \geq (1 + \alpha)\lambda\rho$, then*

$$\Pr \{P_\sigma \text{ is not an } \rho\text{-isomorphism of } F\} \geq 1 - e^{-x}.$$

2. *If $\mathbb{E} \|\mu - \mu_n\|_F \leq (1 - \alpha)\lambda\rho$, then*

$$\Pr \{P_\sigma \text{ is an } \rho\text{-isomorphism of } F\} \geq 1 - e^{-x}.$$

Observe that if the class is star-shaped around 0, and if the set $F_\lambda = \{f \in F : \mathbb{E}_\mu f = \lambda\}$ is ρ -isomorphically projected, then the same holds for the set $\{f \in F : \mathbb{E}_\mu f \geq \lambda\}$. This leads to the following error bound:

Corollary 4.5 *There is an absolute constant c for which the following holds. Let F be a class of functions bounded by b , which is star-shaped around 0 and is a Bernstein class of type β with a constant B . For $0 < \rho, \lambda, \alpha < 1$ and $x > 0$, if*

$$\lambda \geq c \max \left\{ \frac{bx}{n\alpha^2\rho}, \left(\frac{Bx}{n\alpha^2\rho^2} \right)^{1/(2-\beta)} \right\},$$

and if $\mathbb{E} \|\mu - \mu_n\|_{F_\lambda} \leq (1 - \alpha)\lambda\rho$, then with probability at least $1 - e^{-x}$, every $f \in F$ satisfies

$$\mathbb{E}f \leq \max \left\{ \frac{\mathbb{E}_n f}{1 - \rho}, \lambda \right\}.$$

In particular, if f^* is an ε -empirical minimizer (that is, if $n^{-1} \sum_{i=1}^n f^*(X_i) < \varepsilon$) then $\mathbb{E}f^* \leq \max \left\{ \frac{\varepsilon}{1 - \rho}, \lambda \right\}$.

Note that this approach improves on the results of previous section, mainly because the localized averages here are indexed by the constraint $F_r = \{f \in F : \mathbb{E}_\mu f = r\}$. By the Bernstein assumption, $F_r \subset \{f \in F : \mathbb{E}_\mu f^2 \leq Br\}$. In extreme cases, the resulting gap between the averages indexed by the F_r and $F \cap \{f : \mathbb{E}_\mu f^2 \leq r\}$ could be considerable.

Although the star-shape assumption is very mild, it means that the class is very regular in some sense. Indeed, every function encountered at level r will have a scaled down version for any $0 < t < r$. Thus, as the r decreases the sets F_r become richer in some sense - up to a critical value of r , beyond which it becomes impossible to compare the empirical and actual structures. Let us mention that the fact that a set can not be isomorphically projected should have a geometric interpretation, a point we shall return to later.

Although this approach yields the best known error bounds, and as such, the best estimate on the error of the empirical minimizer based on structural results, it is possible to obtain even better bounds on $\mathbb{E}f^*$. In [5] it was shown that the error rate obtained using the structural argument presented above is a λ_n defined by $\mathbb{E} \|\mu_n - \mu\|_{F_r} \sim r$, while sharp estimates on $\mathbb{E}f^*$ are of the order of r' which minimizes the function $\mathbb{E} \|\mu_n - \mu\|_{F_r} - r$.

5 Estimating the localized averages

It is clear from the sample complexity estimates that the behavior of $\mathbb{E} \|\mu_n - \mu\|_F$ or alternatively, that of the Rademacher/gaussian averages is the determining factor in the tail estimates which lead to sharp bounds for both the GC and the learning sample complexities. When global averages are used, that is, when the set indexing the process is the entire class, one can estimate the averages via Dudley's entropy integral. This is particularly useful because there are natural connections between the entropy of the base class G and the entropy of the loss class $\mathcal{L}_p(G, T)$, implying a similar upper bound on the global averages. Another possibility is to use comparison theorems for the averages (e.g. Slepian's Lemma), since p -loss classes are Lipschitz images of the base class.

There are many interesting examples of base classes used in practice that can be written as compositions of simpler base classes [2]. In most cases, there are structural results which enable one to bound the global random averages associated with G using those of the original classes \mathcal{H}_i without resorting to entropy estimates. The most extreme case, in which an entropic argument would clearly be the wrong approach, is when G is the convex hull of elements of \mathcal{H} . The random averages associated with G are the same as those of \mathcal{H} - whereas the metric entropy of G is much larger than the metric entropy of \mathcal{H} .

The situation becomes more difficult when one has to estimate the localized averages, that is, the expectation of the empirical or Rademacher process indexed by $F_r = \{f \in F : \mathbb{E}_\mu f = r\}$. As a first step in the analysis, one can use the assumption that F is a Bernstein class and replace the “ L_1 ” constraint with an L_2 one, which is easier to handle in the context of subgaussian processes. Indeed, as mentioned before, if F is a Bernstein class of type 1 with a constant B then

$$F_r \subset \{f \in F : \mathbb{E}_\mu f^2 \leq Br\},$$

but the latter set could be considerably larger than the former. The next section is devoted to the question of estimating the expectation of the process indexed by $\{f \in F : \mathbb{E}_\mu f^2 \leq r\}$.

5.1 L_2 localized averages

The complexity of the learning problem is determined by the function which measures the localized averages and not the global one. Unfortunately, there are no clear structural properties which enable one to compare the localized averages of a “complicated class” to those of simpler classes used to compose it. Let us give two examples to demonstrate this point; firstly, suppose that G is the convex hull of \mathcal{H} , which is assumed to be star-shaped around 0 and symmetric. Given an L_2 structure on the indexing sets \mathcal{H} and G , the aim is to compare

$$\mathbb{E} \sup_{\{h \in \mathcal{H} : \|h\| \leq r\}} X_h, \text{ and } \mathbb{E} \sup_{\{g \in G : \|g\| \leq r\}} X_g,$$

where $\{X_t : t \in T\}$ is the gaussian process indexed by T .

The two quantities are clearly different; the former corresponds to the averages associated with an index set generated by the convex hull of the intersection of \mathcal{H} and a ball of radius r , while the latter is generated by the intersection of the convex hull of \mathcal{H} with a ball of radius r , and may be considerably larger.

Comparing the two averages is equivalent to the problem of estimating the modulus of continuity of a gaussian process indexed by a set, and the one indexed by its convex hull. There are some results which connect the two (see, for example, [11]), all of which are based on additional parameters of the original class. For example, in the approach presented in [11], additional information on the Kolmogorov numbers or the metric entropy was required. It is not clear whether such an additional information is really needed to compare the two averages.

The other example is when $G = \phi(\mathcal{H})$, where ϕ is a Lipschitz function. Although it is possible to apply Slepian's inequality and thus compare the two expectations, one has to identify the pre image $\phi^{-1}\{g \in G \mid \|g\| \leq r\}$ which may be difficult in practical examples.

When it is possible to estimate the uniform entropy of the class, the L_2 localized averages can be bounded by combining Dudley's entropy bound and an estimate on the average diameter of a random coordinate projection. Since the upper limit in Dudley's entropy integral can be taken to be the radius of a ball centered at the origin which contains the set, then

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

where $Y = \frac{1}{n} \sup_{\{f \in F: \mathbb{E}_{\mu} f^2 \leq r\}} \sum_{i=1}^n f^2(X_i)$. The expected value of the radius can be estimated by the following inequality from [59], which states that

$$\mathbb{E}Y \leq r + \frac{8}{\sqrt{n}} \mathbb{E} \sup_{\{f \in F: \mathbb{E}_{\mu} f^2 \leq r\}} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right|,$$

and the localized averages may be bounded by bootstrapping. This line of argumentation together with Theorem 4.3 yields the following upper bound.

Corollary 5.1 [3, 43] *Let G be a convex class of functions which map Ω into $[0, 1]$, and assume that $\log N_2(\varepsilon, G) \leq \gamma \varepsilon^{-p}$ for some $0 < p < \infty$. Then, the error rate of the class $\mathcal{L}_2(G, T)$ is $O(n^{-2/(2+p)})$ [resp. a learning sample complexity estimate of $O(\varepsilon^{-(1+p/2)})$] if $0 < p < 2$ and of $O(n^{-1/p})$ [resp. $O(\varepsilon^{-p})$] if $p > 2$. Both results are uniform in the target T .*

This result is independent of the underlying measure μ and as such, is sharp. If $0 < p < 2$, the error rate is faster than the $1/\sqrt{n}$ (which one can prove via a uGC argument), while if $p > 2$ the "localized" approach yields the same rates as the uGC one.

5.2 Data dependent bounds

Next, we turn our attention to data dependent error bounds. Recall that the critical value of r in the isomorphic coordinate projections approach is given by the solution r' to the equation $\xi(r) \sim r$ for $\xi(r) = \mathbb{E} \|\mu - \mu_n\|_{F_r}$. If F happens to be a Bernstein class of type 1 with a constant B , and if we set

$$\psi(r) = \mathbb{E} \sup_{\{f \in F: \mathbb{E}_{\mu} f^2 \leq Br^2\}} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right|,$$

then the equation $\psi(r) = r$ has a unique positive solution, which we denote by r^* , and satisfies that $r^* \geq r'$. Thus, r^* is an upper bound on the error rate established in Corollary 4.5. From the practical point of view, it is advantageous

to bound r' (or r^*) using only empirical data (that is, the values of class members on the given sample). This is a useful strategy because one can, in principle, discover the error rate without knowing the “size” of the class. The fact that the “ L_1 ” constraint in the definition of the indexing set F_r was replaced with an L_2 constraint (which is used in the definition of ψ) enables one to estimate r^* efficiently [3]. To that end, we introduce random functions $\hat{\psi}$ which have similar properties to ψ , but can also be computed from the given data. Moreover, these functions have the following property: if \hat{r}^* is the random variable defined as the unique positive solution of the equations $\hat{\psi}(r) = r$, then \hat{r}^* satisfies that $r^* \leq \hat{r}^* + 1/n$ with high probability.

One can show [3] that the random functions

$$\hat{\psi}_{X_1, \dots, X_n}(r) = \frac{C_1}{n} \mathbb{E}_\varepsilon \sup_{\{f \in \text{star}(F, 0) : n^{-1} \sum_{i=1}^n f^2(X_i) \leq r\}} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right| + \frac{C_2 r}{n}$$

have the desired properties, where C_1, C_2 are appropriately selected absolute constants. These functions are computable since they depend on the empirical Rademacher averages associated with the intersection of the star-shaped hull of F and 0 with an *empirical ball* of radius \sqrt{r} (unlike ψ where the ball was an $L_2(\mu)$ ball). Hence, $\hat{\psi}$ is determined by the Euclidean structure endowed on F by the coordinate projection, and thus is potentially computable.

From the statistical point of view, there is a practical method of estimating \hat{r}^* , called the *iterative procedure*, an idea which was introduced in [33] and refined in [3].

Set $\hat{r}_{k+1} = \min\{1, \hat{\psi}(\hat{r}_k)\}$ where $\hat{r}_0 = 1$. It is possible to show that this (computable) sequence converges quickly to \hat{r}^* ; it takes $N = O(\log \log n)$ steps to ensure that $\hat{r}^* \leq \hat{r}_N \leq \hat{r}^* + 1/n$. Thus, with high probability, the iterative procedure gives a quick estimate of r^* .

Passing to the L_2 constraint implies that $\hat{\psi}(r)/\sqrt{r}$ is non-increasing, a fact which plays a significant role in the proof of the fast convergence of \hat{r}_k . The price is that r^* is only an upper bound on r' . Although it should be possible to obtain an iterative procedure scheme using the “ L_1 ” localized averages (and not their L_2 version), such results are not yet known.

Though the iterative procedure yields very positive results in many cases, the problem of determining which parameters govern the localized averages associated with a class is far from being fully understood, and its solution will have far reaching implications.

5.3 Geometric interpretation

A question arising from the discussion is the geometric interpretation of \hat{r}^* or its “ L_1 ” analog \hat{r}' . First, we consider F to be class of nonnegative functions, and note that for any set $\sigma = \{x_1, \dots, x_n\}$

$$\mathbb{E}_\varepsilon \sup_{\{f \in F : n^{-1} \sum_{i=1}^n f(x_i) \leq r\}} \left| \sum_{i=1}^n \varepsilon_i f(x_i) \right| \equiv R(V \cap rnB_1^n),$$

where

$$V = \{(f(x_1), \dots, f(x_n)) : f \in F\} \subset [0, 1]^n.$$

Since $R(rnB_1^n) = rn$, the critical radius \hat{r}' defines an intersection body which is extremal - a subset of F (and thus of B_∞^n), whose Rademacher average (or, in a similar fashion, its ℓ -norm) is of the same order of magnitude as the “regular body” containing it. This assumption seems to apply that at that scale, the original body is “large”, e.g., in the sense that a random coordinate projection will not be a good isomorphism. One possible explanation of that fact could be that this intersection body contains a large cube in a high dimensional coordinate projection, and leads to the following

Question 5.2 *Let $K \subset B_\infty^n$ be such that $R(K \cap rn^{1/p}B_p^n) \sim nr$. Does it mean that K has a coordinate projection σ of proportional dimension which contains $crB_\infty^{|\sigma|}$?*

This question is a generalization of Elton’s Theorem [47], which states that if $K \subset B_\infty^n$ such that $R(K) \geq \delta n$, then there is $\sigma \subset \{1, \dots, n\}$, which satisfies that $|\sigma| \geq c_1 \delta^2 n$, and $c_2 \delta B_\infty^{|\sigma|} \subset P_\sigma K$. In our scenario, one has the additional information that $K_{r,p} = K \cap rn^{1/p}B_p^n$ is an intersection body, but in return one hopes for a cube in a projection of K_r whose dimension does not depend on r . Hence, one should think of it as “Elton’s Theorem” for small scales, since r can be as small as $1/n$, and at a scale smaller than $1/\sqrt{n}$, Elton’s Theorem is vacuous. Note that by taking $K' = K_r/r$, Elton’s Theorem implies the case $p = \infty$. Moreover, the case $p \geq 2$ can be verified (up to logarithmic factors in n) in a similar fashion to the proof of Elton’s Theorem in [47]. Of course, for a general class of functions, the constraint which defines the localized averages is not an ℓ_p^n constraint, and thus a different argument might be required altogether to exhibit that the intersection body is “large”.

A question of a similar flavor (again, in the simplified context of classes of nonnegative functions) is how to estimate the functions $H(r) = M^*(V \cap rB_p^n)$ for a given set $V \subset B_\infty^n$, where $M^*(K) = \int_{S^{n-1}} \|x\|_{K^*}$. Even when V is very regular, e.g. $V = B_q^n$ this task is nontrivial, though possible. In fact, in that case the dual to the norm endowed by the intersection body (which is the J -functional of the two spaces) can be estimated up to an absolute constant, which leads to sharp bounds on the integral. The general case seems to be considerably more difficult.

6 Bernstein type of L_p loss classes

The assumption which makes the proof of Theorem 4.3 (and analogous results) possible, is that the class in question has a nontrivial Bernstein type. Thus, one needs to check whether this condition holds for various classes of functions. As stated before, if the target function T belongs to the base class G then $\mathcal{L}_p(G, T)$ is a Bernstein class of type 1 with constant 1. If $T \notin G$ the situation is less obvious. Recall that we are interested in base classes which consist of functions

whose ranges are contained in $[0, 1]$, that the range of T is also contained in the same interval and that the p loss functional associated with the target T is $\ell_p(g) = |T - g|^p - |T - P_G T|^p$, where $P_G T$ is the best approximation of T in G with respect to the $L_p(\mu)$ norm. In general, the geometry of G and the location of T determine whether the loss class has a nontrivial Bernstein type and influences the constants in a manner which will be specified below.

If the base class is convex the situation is simpler to handle, since the best approximation map onto a convex set in a strictly convex and smooth space is well defined and has a geometric interpretation. In particular if $X = L_p(\mu)$ for $1 < p < \infty$, then every target T will have a unique best approximation in G and the loss class $\mathcal{L}_p(G, T)$ has a nontrivial Bernstein type with a constant which depends only on p and not on the underlying measure.

Theorem 6.1 [41] *Let $1 < p < \infty$ and assume that G is a compact convex subset of $L_p(\mu)$. If $2 \leq p < \infty$ then for every $f \in \mathcal{L}_p(G, T)$,*

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

If $1 < p < 2$ then

$$\mathbb{E}_\mu f^2 \leq \frac{4p \cdot 2^p}{p-1} \mathbb{E}_\mu f$$

We shall present a proof in the case $1 < p < 2$. The other case follows a similar path.

We begin by bounding $\mathbb{E}_\mu f^2$ from above. For any $1 < p < \infty$ and any $f \in \mathcal{L}_p(G, T)$, $\mathbb{E}_\mu f^2 \leq p^2 \mathbb{E}_\mu |g - P_G T|^2$, which follows from the fact that $|x|^p$ is a Lipschitz function on $[0, 1]$ with a constant p .

To bound $\mathbb{E}_\mu |g - P_G T|^2$ from above using $\mathbb{E}_\mu f$, we use the uniform convexity of $L_p(\mu)$. Recall that the modulus of convexity of $L_p(\mu)$ is given by $\delta_p(\varepsilon) = 1 - (1 - (\varepsilon/2)^p)^{1/p}$, while for $1 < p < 2$, $\delta_p(\varepsilon) \geq (p-1)\varepsilon^2/2 \cdot 2^p$ [21, 22] (in fact δ_{L_p} is equivalent to $c_p \varepsilon^2$ as $\varepsilon \rightarrow 0$).

The main observation required for the proof is based on a standard separation argument.

Lemma 6.2 *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 $d_g = \|T - g\|$.

Proof (of Theorem 6.1) We estimate $\|g - P_G T\|^2$ from above by $\mathbb{E}_\mu f = d_g^p - d^p$. Applying Lemma 6.2 to $X = L_p$,

$$\|g - P_G T\|^2 \leq \frac{2 \cdot 2^p}{p-1} d_g (d_g - d).$$

Since G consists of functions into $[0, 1]$ and T takes values in $[0, 1]$ then $0 \leq d, d_g \leq 1$. Observe that for any $0 \leq y \leq x \leq 1$, $x(x - y) \leq \frac{2}{p}(x^p - y^p)$, and by selecting $x = d_g$ and $y = d$,

$$\mathbb{E}_\mu f^2 \leq p^2 \|g - P_G T\|^2 \leq \frac{2p^2 \cdot 2^p}{p-1} d_g(d_g - d) \leq \frac{4p \cdot 2^p}{p-1} \mathbb{E}_\mu f.$$

■

Let us mention that our analysis does not pass smoothly to the nonconvex case. To start with, one has a problem with the definition of the loss class, because it is no longer true that every target has a unique best approximation in G . Even if $P_G T$ is a well defined function, the geometric characterization of the nearest point map is no longer valid. However, it is still possible to obtain partial estimates on the Bernstein type of $\mathcal{L}_2(G, T)$. To that end, we require

Lemma 6.3 [49] *Assume that there is some $0 \leq \beta < 1$ such that for every $g \in G$*

$$\langle P_G T - T, P_G T - g \rangle \leq \frac{\beta}{2} \|P_G T - g\|_{L_2(\mu)}^2. \quad (6.1)$$

Then $\mathcal{L}_2(G, T)$ is a Bernstein class of type 1 with a constant $16/(1 - \beta)$.

In cases where T has a unique best approximation in G , one can find the smallest value of β for which (6.1) holds, which leads to a bound on the constant in the Bernstein type of the 2-loss class.

Without loss of generality, assume that G is a compact subset of $L_2(\mu)$. Denote by $\text{nup}(G, \mu)$ the set of points with more than a unique best approximation in G with respect to the $L_2(\mu)$ norm. Suppose $T \in L_2(\mu) \setminus (G \cup \text{nup}(G, \mu))$, and let

$$r_{G, \mu}(T) := \inf\{\|g - P_G T\| : g \in \{\lambda(T - P_G T) : \lambda > 0\} \cap \text{nup}(G, \mu)\}.$$

Intuitively, $r_{G, \mu}(T) = \|f_{\text{nup}} - P_G T\|$ where f_{nup} is the point in $\text{nup}(G, \mu)$ found by extending a ray from $P_G T$ through T until reaching $\text{nup}(G, \mu)$. One can show that the smallest possible value of β is

$$\beta_{G, \mu}(T) := \frac{\|T - P_G T\|}{r_{G, \mu}(T)} = \frac{\|T - P_G T\|}{\|f_{\text{nup}} - P_G T\|}.$$

Clearly, if G is convex then $\text{nup}(G, \mu)$ is the empty set; hence for all $T \in L_2(\mu)$, $r_{G, \mu}(T) = \infty$ and $\beta_{G, \mu}(T) = 0$, which is exactly the result obtained in Theorem 6.1 - that $\mathcal{L}_2(G, T)$ has Bernstein type 1 with a constant 16.

The analysis of the Bernstein type of loss classes when the target has more than a unique best approximation is incomplete. Moreover, the Bernstein type of p -loss classes in the nonconvex case is far from being understood.

Let us point out that learning is inherently more difficult if the base class G is nonconvex [35], in the sense that the best uniform upper bound on the sample complexity that one can obtain (at least in the agnostic case) is $\Omega(1/\varepsilon^2)$ regardless of the “size” of the base class G .

7 Classes of linear functionals

Here we investigate classes which seemingly have an easy structure - classes which consist of linear functionals on a given domain. The simplest example we consider is when the class G is the dual unit ball of some infinite dimensional Banach space X , and the domain of the functionals is the unit ball B_X . Thus,

$$G = \{x^* : B_X \rightarrow \mathbb{R} \mid \|x^*\| \leq 1\}.$$

It was shown in [19] that G is a uniform Glivenko-Cantelli class of functions if and only if X has a nontrivial type. Actually, if X is infinite dimensional, the exact asymptotic behavior of the combinatorial dimension of G is determined by the Rademacher type of X . This follows from the next observation:

Lemma 7.1 $\{x_1, \dots, x_n\} \subset B_X$ is ε -shattered by B_{X^*} if and only if $(x_i)_{i=1}^n$ are linearly independent and ε -dominate the ℓ_1^n unit-vector basis, that is, for every $(a_i)_{i=1}^n \subset \mathbb{R}$, $\varepsilon \sum_{i=1}^n |a_i| \leq \|\sum_{i=1}^n a_i x_i\|$.

Using this lemma, it is possible to upper bound the combinatorial dimension of B_{X^*} .

Theorem 7.2 [46] *Let X be a Banach space. Then, for every $\varepsilon > 0$,*

$$\text{vc}(B_{X^*}, B_X, \varepsilon) \leq \left(\frac{T_p(X)}{\varepsilon} \right)^{\frac{p}{p-1}} + 1,$$

where $T_p(X)$ is the Rademacher type p constant of X .

If X happens to be infinite dimensional, then this upper bound can be complemented. This is due to the fact that if $p^* = \sup\{p \mid X \text{ has type } p\}$ then $\ell_{p^*}^n$ is finitely represented in X .

Theorem 7.3 [46] *Let X be an infinite dimensional Banach space. Then $\text{vc}(B_{X^*}, B_X, \varepsilon)$ is finite for every $\varepsilon > 0$ if and only if X has a nontrivial type. In addition,*

$$\left(\frac{1}{\varepsilon} \right)^{\frac{p^*}{p^*-1}} - 1 \leq \text{vc}(B_{X^*}, B_X, \varepsilon).$$

A more general problem is when the domain of the class of functionals is not B_X but some other convex symmetric set. Let K and L be two convex symmetric bodies in \mathbb{R}^m and consider elements of the polar body L° as functions on K using the fixed inner product in \mathbb{R}^m . It turns out that computing the combinatorial dimension $\text{vc}(L^\circ, K, \varepsilon)$ is equivalent to a factorization problem. For every two convex, symmetric sets K and L in \mathbb{R}^m and every integer $n \leq m$, let $\Gamma_n(K, L) = \inf \|A\| \|B\|$, where the infimum is taken with respect to all subspaces of $E \subset \mathbb{R}^m$ of dimension n and all operators $B : (E, \|\cdot\|_{L \cap E}) \rightarrow \ell_1^n$, $A : \ell_1^n \rightarrow (E, \|\cdot\|_{K \cap E})$, such that $AB = id : E \rightarrow E$.

Lemma 7.4 [46] For every integer n and any convex and symmetric sets K and L ,

$$\frac{1}{\Gamma_n(K, L)} = \sup\{\varepsilon | \exists \{x_1, \dots, x_n\} \subset K, \varepsilon(L \cap E) \subset \text{absconv}(x_1, \dots, x_n)\} \quad (7.1)$$

where $E = \text{span}\{x_1, \dots, x_n\}$ and $\text{absconv}(x_1, \dots, x_n)$ is the symmetric convex hull of $\{x_1, \dots, x_n\}$.

Thus, $1/\Gamma_n(K, L)$ is the discrete inverse of $\text{vc}(L^\circ, K, \varepsilon)$.

In [46] the factorization constants $\Gamma_n(B_p^m, B_q^m)$ were investigated. However, from the Machine Learning perspective, the significant point is that one can estimate the combinatorial dimension of classes of functions which can be viewed as linear functionals on some domain. One such case is when the base class G is the unit ball of some Sobolev space X which is embedded in $C(\Omega)$. Hence, for every $g \in G$ and every $x \in \Omega$, $g(x) = \delta_x(g)$. If the point evaluation functionals are uniformly bounded in X^* (say by 1) then

$$\text{vc}(G, \Omega, \varepsilon) \leq \text{vc}(B_{X^{**}}, B_{X^*}, \varepsilon) \leq \left(\frac{T_p(X)}{\varepsilon}\right)^{\frac{p}{p-1}},$$

which is often tighter than results obtained by more direct approaches [40].

The next example, which plays a central role in modern Machine Learning, has a similar flavor. Let Ω be a compact set and set $K : \Omega \times \Omega \rightarrow \mathbb{R}$ to be a positive definite, continuous function. Given a probability measure ν on Ω , set $T_K : L_2(\nu) \rightarrow L_2(\nu)$ by $(T_K f)(x) = \int K(x, y)f(y)d\nu(y)$. According to the spectral Theorem, T_K has a diagonal representation, and denote by $(\phi_n(x))_{n=1}^\infty$ the orthonormal basis and by $(\lambda_i)_{i=1}^\infty$ the non-increasing sequence of eigenvalues corresponding to this diagonal representation. Mercer's Theorem implies that $\nu \times \nu$ almost surely,

$$K(x, y) = \sum_{n=1}^{\infty} \lambda_n \phi_n(x) \phi_n(y),$$

and since K is continuous, then under mild assumptions on ν this representation of K holds for every $x, y \in \Omega$.

Let F_K be the class consisting of all the functions of the form $\sum_{i=1}^{\infty} a_i K(x_i, \cdot)$ where $x_i \in \Omega$ and $a_i \in \mathbb{R}$ satisfy that $\sum_{i,j=1}^{\infty} a_i a_j K(x_i, x_j) \leq 1$.

There is extensive literature on learning algorithms based on the class F_K and on similar classes (in which $(a_i)_{i=1}^\infty$ satisfy slightly different constraints), all called kernel classes [55, 13]. These are probably the most frequently used classes in real-life applications such as optical character recognition, DNA sequence analysis, Time Series Prediction and many more.

It is easy to see that F_K is the unit ball of the Hilbert space associated with the kernel, called the *reproducing kernel Hilbert space* and denoted by \mathcal{H} . An alternative way to define the reproducing kernel Hilbert space which makes the interpretation of F_K as a class of linear functionals clearer, is via the *feature map*. Define $\Phi : \Omega \rightarrow \ell_2$ by $\Phi(x) = (\sqrt{\lambda_i} \phi_i(x))_{i=1}^\infty$. Then,

$$F_K = \{f(\cdot) = \langle \beta, \Phi(\cdot) \rangle_{\ell_2} \mid \|\beta\|_2 \leq 1\}.$$

In other words, the feature map is a way of embedding the space Ω in ℓ_2 and F_K can be represented as the unit ball in the space. Each $\beta \in \ell_2$ acts as a functional on the image of the Ω via the feature map, denoted by $\Phi(\Omega)$, and in particular, if B_2 denotes the unit ball in ℓ_2 then $\text{vc}(F, \Omega, \varepsilon) = \text{vc}(B_2, \Phi(\Omega), \varepsilon)$ for any $\varepsilon > 0$. Using the volumetric methods of [46] it is possible to estimate the combinatorial dimension of F_K and of other families of kernel classes. For F_K one can show that if, for some measure with a strictly positive density, the eigenfunctions of the integral operator T_K are uniformly bounded, then the combinatorial dimension of the class can be bounded above in terms of the eigenvalues of T_K . In general, it is possible to connect the fact that $\{x_1, \dots, x_n\}$ is ε -shattered to the eigenvalues of the Gram matrix $(K(x_i, x_j))_{i,j=1}^n$. However, spectral information is not enough to determine the largest ε for which $\{x_1, \dots, x_n\}$ is ε -shattered by F_K , and for that one needs to analyze the geometry of TB_1^n , where T is the linear operator which maps the unit vectors e_i to $\Phi(x_i)$.

Because of the relatively simple geometry of F_K , the L_2 -localized averages associated with F_K can be computed exactly. Observe that the index set $H_r = F_K \cap \{f \mid \mathbb{E}_\mu f^2 \leq r\}$ with respect to which the supremum is taken, is an intersection of a ball and an ellipsoid, which makes the computation of the L_2 norm of $\sup_{f \in H_r} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right|$ possible. Indeed, if \mathcal{E} and \mathcal{E}' are ellipsoids with the same principal directions and axes $(a_i)_{i=1}^\infty$ and $(b_i)_{i=1}^\infty$ respectively, then the ellipsoid \mathcal{B} whose axes are $(\min\{a_i, b_i\})_{i=1}^\infty$ satisfies that $\mathcal{B} \subset \mathcal{E} \cap \mathcal{E}' \subset \sqrt{2}\mathcal{B}$. Therefore, one can replace the set $\mathcal{E} \cap \mathcal{E}'$ indexing the Rademacher process by \mathcal{B} , losing only a multiplicative factor, and it is a straightforward exercise to compute the L_2 norm of the supremum of the process indexed by \mathcal{B} . Next, one shows that the L_2 norm of $\sup_{f \in H_r} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right|$ is equivalent to its L_1 norm, which follows from concentration.

Theorem 7.5 [42] *For every $1 < p < \infty$ there is a constant c_p for which the following holds. Let $F \subset B(L_\infty(\Omega))$, set μ to be a probability measure on Ω and put $\sigma_F^2 = \sup_{f \in F} \mathbb{E}_\mu f^2$. If n satisfies that $\sigma_F^2 \geq 1/n$ then*

$$c_p \left(\mathbb{E} \sup_{f \in F} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right|^p \right)^{\frac{1}{p}} \leq \mathbb{E} \sup_{f \in F} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right| \leq \left(\mathbb{E} \sup_{f \in F} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right|^p \right)^{\frac{1}{p}},$$

where $(X_i)_{i=1}^n$ are independent random variables distributed according to μ .

The upper bound is just Hölder's inequality. The condition on the σ_F is needed only for the lower bound, which is more delicate than the Kahane-Khintchine inequality because of the additional randomness due to $(X_i)_{i=1}^n$.

Combining these facts one can prove the following:

Corollary 7.6 [42] *There are absolute constants C and c such that if $(\lambda_i)_{i=1}^\infty$ is the nondecreasing sequence of eigenvalues of the integral operator $(T_K f)(x) = \int K(x, y) f(y) d\mu(y)$ and if $\lambda_1 \geq 1/n$, then for every $r \geq 1/n$,*

$$\frac{1}{\sqrt{n}} \mathbb{E} \sup_{\{f \in F: \mathbb{E}_\mu f^2 \leq r\}} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right| \sim \left(\sum_{j=1}^\infty \min\{\lambda_j, r\} \right)^{\frac{1}{2}},$$

where $(X_i)_{i=1}^\infty$ are independent, distributed according to μ .

It is evident that the localized averages depend on the underlying measure μ , since the spectrum of the integral operator may change dramatically when the measure changes. Moreover, since the underlying measure is *unknown* it is impossible to estimate the eigenvalues of the integral operator. This leads to another question: is it possible to replace the eigenvalues of the integral operator with those of the normalized Gram matrix $(n^{-1}K(X_i, X_j))_{i,j=1}^n$? In other words, is there a concentration of measure phenomenon for the eigenvalues of $n^{-1}(K(X_i, X_j))_{i,j=1}^n$ around those of T_K ? A positive answer would enable one to estimate the L_2 -localized averages from empirical (and thus computable) data.

There are some partial asymptotic results in that direction due to Koltchinskii and Giné [31, 32]. In our case, T_K is trace class (because K is continuous), and for such operators they proved that the spectrum of $(n^{-1}K(X_i, X_j))_{i,j=1}^n$ converges almost surely (as sets) to the spectrum of the integral operator T_K . There are few known results on the rate of convergence, and those are less than satisfactory.

In a similar way one can bound the empirical local Rademacher averages, $\frac{1}{\sqrt{n}}\mathbb{E}_\varepsilon \sup_{\{f \in F: \mathbb{E}_{\mu_n} f^2 \leq r\}} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right|$ where μ_n is the (random) empirical measure supported on (X_1, \dots, X_n) , in which case the eigenvalues of the integral operator in Corollary 7.6 are replaced by the normalized eigenvalues of the Gram matrix - making the iterative procedure presented in the previous section feasible.

8 Concluding Remarks

Although this article focused in its entirety on the sample complexity problem, we would like to end it by briefly mentioning several other questions which are significant in Learning Theory and have a mathematical flavor.

Firstly, there is the question of the learning algorithm. Empirical minimization is not the only possibility to attack the learning problem. Even if one chooses empirical minimization, there is an algorithmic problem of finding an almost minimizer, and the *computational complexity* of the algorithm has to be estimated. Secondly, there is an issue of the degree of approximation; a solution of the learning problem is a function which is almost the optimal in the class. However, even the optimal can be “very far” from the target. This is a classical tradeoff between large classes for which the approximation error is small for many targets, but the sample complexity is huge, and small classes for which a small sample is needed for the probabilistic construction but the approximation error is very large. Balancing the statistical error and the approximation error simultaneously can be achieved via *model selection*, in which a very large class G is divided into an increasing family of classes G_i such that $\bigcup_{i=1}^\infty G_i = G$, and one wants to find the “best class” from which the target T can be approximated. A formulation similar in nature involves regularization. For example, if G is the

entire reproducing kernel Hilbert space with norm $\|\cdot\|_K$ then one can consider $\ell(g) = (g - T)^2 + \gamma\|g\|_K$. This loss functional balances the way g approximates T on the sample and the “smoothness” of g , captured by $\|g\|_K$.

All of the issues mentioned above and others that were not mentioned are intriguing mathematically. Our hope is that this article would encourage more mathematicians to investigate questions arising in this rapidly developing field, boosting the theoretical side of Machine Learning.

References

- [1] N. Alon, S. Ben–David, N. Cesa–Bianchi, D. Haussler: Scale sensitive dimensions, uniform convergence and learnability, *J. of ACM* 44 (4) 615–631, 1997.
- [2] M. Anthony, P.L. Bartlett: *Neural Network Learning*, Cambridge University Press, 1999.
- [3] P.L. Bartlett, O. Bousquet, S. Mendelson: Localized Rademacher Complexities, *Ann. Stat.* to appear.
- [4] P.L. Bartlett, P.M. Long: Prediction, learning, uniform convergence, and scale-sensitive dimensions, *J. Comput. System Sci.* 56, 174–190, 1998.
- [5] P.L. Bartlett, S. Mendelson: Empirical risk minimization, preprint.
- [6] A. Barron, L. Birgé, P. Massart: Risk bounds for model selection via penalization, *Probab. Theory Related Fields* 113 (3), 301-413, 1999.
- [7] L. Birgé, P. Massart: From model selection to adaptive estimation, in *Festschrift for Lucien Le Cam*, 55-87, Springer, New York, 1997.
- [8] S. Boucheron, G. Lugosi, P. Massart: A sharp concentration inequality with applications in random combinatorics and learning, *Random Struct. Algor.* 16, 277–292, 2000.
- [9] S. Boucheron, G. Lugosi, P. Massart: Concentration inequalities using the entropy method, *Ann. Probab.* 31(3) 1583-1614, 2003.
- [10] O. Bousquet: A Bennett concentration inequality and its application to suprema of empirical processes, *C. R. Acad. Sci. Paris, Ser. I*, 334, 495-500, 2002.
- [11] O. Bousquet, V. Koltchinskii, D. Panchenko: Some local measures of complexity of convex hulls and generalization bounds, in *Proceedings of the 15th annual conference on Computational Learning Theory COLT02*, Jyrki Kivinen and Robert H. Sloan(Eds.), Lecture Notes in Computer Sciences 2375, Springer, 59-73, 2002.
- [12] B. Carl, I. Kyrezi, A. Pajor: Metric entropy of convex hulls in Banach spaces, *J. London Math. Soc.* 60(2), 871–896, 1999.
- [13] P. Cucker, S. Smale: On the mathematical foundations of learning, *B. Am. Math. Soc.* 39(1) 1–49, 2002.
- [14] L. Devroye, L. Györfi, G. Lugosi: *A probabilistic theory of pattern recognition*, Springer, 1996.
- [15] L. Devroye, G. Lugosi: *Combinatorial methods in density estimation*, Springer, 2000.

- [16] R.M. Dudley: Central limit theorems for empirical measures, *Ann. Probab.* 6(6), 899-929, 1978.
- [17] R.M. Dudley: Universal Donsker classes and metric entropy, *Ann. Probab.* 15, 1306–1326, 1987.
- [18] R.M. Dudley: *Uniform Central Limit Theorems*, Cambridge Studies in Advanced Mathematics 63, Cambridge University Press 1999.
- [19] R.M. Dudley, E. Giné, J. Zinn: Uniform and universal Glivenko–Cantelli classes, *J. Theoret. Prob.* 4, 485–510, 1991.
- [20] A. Ehrenfeucht, D. Haussler, M.J. Kearns, L.G. Valiant: A general lower bound on the number of examples needed for learning, *Inform. Comput.* 82, 247–261, 1989.
- [21] P. Habala, P. Haĵek, V. Zizler: *Introduction to Banach spaces* vol I and II, matfyzpress, Univ. Karlovy, Prague, 1996.
- [22] O. Hanner: On the uniform convexity of L^p and l^p , *Ark. Math.* 3, 239-244, 1956.
- [23] D. Haussler: Sphere packing numbers for subsets of Boolean n -cube with bounded Vapnik-Chervonenkis dimension, *J. Comb. Theory A* 69, 217-232, 1995.
- [24] J. Hoffman-Jørgensen: Sums of independent Banach space valued random variables, Aarhus University preprint series 1972/1973, 15, 1973.
- [25] J. Hoffman-Jørgensen: Probability in Banach spaces, Lecture notes in Mathematics 598, 164-229, Springer-Verlag, 1976.
- [26] F. Gao: Metric entropy of convex hulls, *Isr. J. Math.* 123, 359–364, 2001.
- [27] E. Giné and J. Zinn: Some limit theorems for empirical processes, *Ann. Probab.* 12(4), 929–989, 1984.
- [28] E. Giné, J. Zinn: Gaussian characterization of uniform Donsker classes of functions, *Ann. Probab.* 19(2), 758–782, 1991.
- [29] J.P. Kahane: *Some random series of functions*, Heath-Lexington, 1968 (second edition: Cambridge University Press, 1985).
- [30] M.G. Karpovsky, V.D. Milman: Coordinate density of sets of vectors, *Discrete Math.* 24, 177-184, 1978.
- [31] V. Koltchinskii: Asymptotics of spectral projections of some random matrices approximating integral operators, in *Progress in Probability* Vol 43, Birkhauser 1998.
- [32] V. Koltchinskii, E. Giné: Random matrix approximation of spectra of integral operators, *Bernoulli* 6(1), 113-167, 2000.

- [33] V. Koltchinskii, D. Panchenko: Rademacher processes and bounding the risk of function learning, *High dimensional probability, II (Seattle, WA, 1999)*, 443–457, in *Progress in Probability* 47, Birkhauser.
- [34] M. Ledoux: *The concentration of measure phenomenon*, Mathematical Surveys and Monographs, Vol 89, AMS, 2001.
- [35] W.S. Lee, P.L. Bartlett, R.C. Williamson: The Importance of Convexity in Learning with Squared Loss, *IEEE Trans. Info. Theory* 44 (5), 1974-1980, 1998.
- [36] G. Lugosi: Pattern classification and learning theory, in L. Györfi (editor), *Principles of Nonparametric Learning* Springer, 1–56, 2002.
- [37] P. Massart: About the constants in Talagrand’s concentration inequality for empirical processes, *Ann. Probab.* 28(2) 863-884, 2000.
- [38] S. Mendelson: On the size of convex hulls of small sets, *J. Mach. Learn. Res.* 2, 1-18, 2001.
- [39] S. Mendelson: Rademacher averages and phase transitions in Glivenko-Cantelli class, *IEEE Trans. Info. Theory*, 48(1), 251-263, 2002.
- [40] S. Mendelson: Learnability in Hilbert spaces with Reproducing Kernels, *J. Complexity*, 18(1), 152-170, 2002.
- [41] S. Mendelson, Improving the sample complexity using global data, *IEEE Trans. Info. Theory* 48(7), 1977-1991, 2002.
- [42] S. Mendelson: Geometric parameters of kernel machines, in *Proceedings of the 15th annual conference on Computational Learning Theory COLT02*, Jyrki Kivinen and Robert H. Sloan(Eds.), Lecture Notes in Computer Sciences 2375, Springer, 29-43, 2002.
- [43] S. Mendelson: A few notes on Statistical Learning Theory in *Proceedings of the Machine Learning Summer School, Canberra 2002*, S. Mendelson and A.J. Smola (Eds.), in Lecture notes in computer sciences 2600, Springer 2003.
- [44] S. Mendelson, A. Pajor, M. Rudelson: work in progress.
- [45] S. Mendelson, M. Rudelson: Concentration of covering numbers, unpublished notes.
- [46] S. Mendelson, G. Schechtman: The shattering dimension of sets of linear functionals, *Ann. Probab.* to appear.
- [47] S. Mendelson, R. Vershynin: Entropy and the combinatorial dimension, *Invent. Math.* 152(1), 37-55, 2003.
- [48] S. Mendelson, R. Vershynin: Remarks on the Geometry of Coordinate Projections in \mathbb{R}^n , *Isr. J. Math.* to appear.

- [49] S. Mendelson, R.C. Williamson: Agnostic learning of non-convex classes of functions, in *Proceedings of the 15th annual conference on Computational Learning Theory COLT02*, Jyrki Kivinen and Robert H. Sloan(Eds.), Lecture Notes in Computer Sciences 2375, Springer, 1-13, 2002.
- [50] V.D. Milman, N. Tomczak-Jaegermann: Sudakov type inequalities for convex bodies in \mathbb{R}^n , in *Geometric aspects in Functional Analysis 85-85*, Lecture Notes in Mathematics 1267, 113–121, Springer 1987.
- [51] A. Pajor: *Sous espaces ℓ_1^n des espaces de Banach*, Hermann, Paris, 1985.
- [52] E. Rio: Une inegalite de Bennett pour les maxima de processus empiriques. (French) [A Bennett-type inequality for maxima of empirical processes]. Ann. Inst. H. Poincar Probab. Statist. 38(6) 1053–1057, 2002.
- [53] M. Rudelson, R. Vershynin: preprint.
- [54] N. Sauer: On the density of families of sets, J. Comb. Theory A, 13, 145-147, 1972.
- [55] B. Schölkopf, A.J. Smola: *Learning with kernels*, MIT press, 2001.
- [56] S. Shelah: A combinatorial problem: stability and orders for models and theories in infinitary languages, Pac. J. Math. 41, 247-261, 1972.
- [57] N. Tomczak–Jaegermann: *Banach–Mazur distance and finite–dimensional operator Ideals*, Pitman monographs and surveys in pure and applied Mathematics 38, 1989.
- [58] M. Talagrand: Type, infratype, and the Elton-Pajor Theorem, Invent. Math. 107, 41–59, 1992.
- [59] M. Talagrand: Sharper bounds for Gaussian and empirical processes, Ann. Probab. 22(1), 28-76, 1994.
- [60] M. Talagrand: Vapnik-Chervonenkis type conditions and uniform Donsker classes of functions, Ann. Probab. to appear.
- [61] A.W. Van der Vaart, J.A. Wellner: *Weak convergence and Empirical Processes*, Springer-Verlag, 1996.
- [62] V. Vapnik: *Statistical Learning Theory*, Wiley 1998.
- [63] V.N. Vapnik, A.Ya Chervonenkis: Necessary and sufficient conditions for uniform convergence of means to mathematical expectations, Theory Prob. Applic. 26(3), 532–553, 1971.
- [64] A. Vidyasagar: *The Theory of learning and generalization* Springer-Verlag, 1996.
- [65] D.X. Zhou: The covering number in Learning Theory, J. Complexity, 18(3) 739–767, 2002.

- [66] D.X. Zhou and S. Smale: Estimating the approximation error in Learning Theory, *Anal. Appl. (Singap.)* 1(1) 17–41, 2003.