Jyrki Kivinen
Robert H. Sloan (Eds.)

LNAI 2375

Computational Learning Theory

15th Annual Conference on Computational Learning Theory, COLT 2002
Sydney, Australia, July 2002, Proceedings

Springer

Jyrki Kivinen Robert H. Sloan (Eds.)

Computational Learning Theory

15th Annual Conference on Computational Learning Theory,
COLT 2002
Sydney, Australia, July 8-10, 2002
Proceedings

Springer

Series Editors

Jaime G. Carbonell,Carnegie Mellon University, Pittsburgh, PA, USA
Jörg Siekmann, University of Saarland, Saarbrücken, Germany

Volume Editors

Jyrki Kivinen
Australian National University
Research School of Information Sciences and Engineering
Canberra ACT 0200, Australia
E-mail: Jyrki.Kivinen@anu.edu au

Robert H. Sloan
University of Illinois at Chicago, Computer Science Department
851 S. Morgan St., Chicago, IL 60607, USA
E-mail: sloan@cs.uic.edu

Cataloging-in-Publication Data applied for

Die Deutsche Bibliothek - CIP-Einheitsaufnahme

Computational learning theory : proceedings / 15th Annual Conference on Computational Learning Theory, COLT 2002, Sydney, Australia, July 8 - 10, 2002. Jyrki Kivinen ; Robert H. Sloan (ed.). - Berlin ; Heidelberg ; New York ; Barcelona ; Hong Kong ; London ; Milan ; Paris ; Tokyo : Springer, 2002
(Lecture notes in computer science ; Vol. 2375 : Lecture notes in artificial intelligence)
ISBN 3-540-43836-X

CR Subject Classification (1998): I.2.6, I.2.3, F.4.1, F.1.1, F.2

ISSN 0302-9743
ISBN 3-540-43836-X Springer-Verlag Berlin Heidelberg New York

Springer-Verlag Berlin Heidelberg New York
a member of BertelsmannSpringer Science+Business Media GmbH

http://www.springer.de

Typesetting: Camera-ready by author, data conversion by DA-TeX Gerd Blumenstein
Printed on acid-free paper SPIN 10870415 06/3142 5 4 3 2 1 0

Preface

This volume contains papers presented at the Fifteenth Annual Conference on Computational Learning Theory (COLT 2002) held on the main campus of the University of New South Wales in Sydney, Australia from July 8 to 10, 2002. Naturally, these are papers in the field of computational learning theory, a research field devoted to studying the design and analysis of algorithms for making predictions about the future based on past experiences, with an emphasis on rigorous mathematical analysis.

COLT 2002 was co-located with the Nineteenth International Conference on Machine Learning (ICML 2002) and with the Twelfth International Conference on Inductive Logic Programming (ILP 2002).

Note that COLT 2002 was the first conference to take place after the full merger of the Annual Conference on Computational Learning Theory with the European Conference on Computational Learning Theory. (In 2001 a joint conference consisting of the 5th European Conference on Computational Learning Theory and the 14th Annual Conference on Computational Learning Theory was held; the last independent European Conference on Computational Learning Theory was held in 1999.)

The technical program of COLT 2002 contained 26 papers selected from 55 submissions. In addition, Christos Papadimitriou (University of California at Berkeley) was invited to give a keynote lecture and to contribute an abstract of his lecture to these proceedings.

The Mark Fulk Award is presented annually for the best paper coauthored by a student. This year's award was won by Sandra Zilles for the paper "Merging Uniform Inductive Learners."

April 2002

Jyrki Kivinen
Robert H. Sloan

Thanks and Acknowledgments

We gratefully thank all the individuals and organizations responsible for the success of the conference.

Program Committee

We especially want to thank the program committee: Dana Angluin (Yale), Javed Aslam (Dartmouth), Peter Bartlett (BIOwulf Technologies), Shai Ben-David (Technion), John Case (Univ. of Delaware), Peter Grünwald (CWI), Ralf Herbrich (Microsoft Research), Mark Herbster (University College London), Gábor Lugosi (Pompeu Fabra University), Ron Meir (Technion), Shahar Mendelson (Australian National Univ.), Michael Schmitt (Ruhr-Universität Bochum), Rocco Servedio (Harvard), and Santosh Vempala (MIT).

We also acknowledge the creators of the CyberChair software for making a software package that helped the committee do its work.

Local Arrangements, Co-located Conferences Support

Special thanks go to our conference chair Arun Sharma and local arrangements chair Eric Martin (both at Univ. of New South Wales) for setting up COLT 2002 in Sydney. Rochelle McDonald and Sue Lewis provided administrative support. Claude Sammut in his role as conference chair of ICML and program co-chair of ILP ensured smooth coordination with the two co-located conferences.

COLT Community

For keeping the COLT series going, we thank the COLT steering committee, and especially Chair John Shawe-Taylor and Treasurer John Case for all their hard work. We also thank Stephen Kwek for maintaining the COLT web site at `http://www.learningtheory.org`.

Sponsoring Institution

School of Computer Science and Engineering, University of New South Wales, Australia

Referees

Peter Auer
Andrew Barto
Stephane Boucheron
Olivier Bousquet
Nicolò Cesa-Bianchi
Tapio Elomaa
Ran El-Yaniv
Allan Erskine
Henning Fernau
Jürgen Forster
Dean Foster
Claudio Gentile
Judy Goldsmith
Thore Graepel
Lisa Hellerstein
Daniel Herrmann
Colin de la Higuera
Sean Holden
Marcus Hutter
Sanjay Jain
Yuri Kalnishkan
Makoto Kanazawa
Satoshi Kobayashi
Vladimir Koltchinskii
Matti Kääriäinen
Wee Sun Lee
Shie Mannor
Ryan O'Donnell
Alain Pajor
Gunnar Rätsch
Robert Schapire
John Shawe-Taylor
Takeshi Shinohara
David Shmoys
Yoram Singer
Carl Smith
Frank Stephan
György Turán
Paul Vitányi
Manfred Warmuth
Jon A. Wellner
Robert C. Williamson

Table of Contents

Statistical Learning Theory

Online Learning

Inductive Inference

PAC Learning

Boosting

Other Learning Paradigms

Invited Talk

Agnostic Learning Nonconvex Function Classes

Shahar Mendelson and Robert C. Williamson

Research School of Information Sciences and Engineering
Australian National University
Canberra, ACT 0200, Australia
{shahar.mendelson,Bob.Williamson}@anu.edu.au

Abstract. We consider the sample complexity of agnostic learning with respect to squared loss. It is known that if the function class F used for learning is convex then one can obtain better sample complexity bounds than usual. It has been claimed that there is a lower bound that showed there was an essential gap in the rate. In this paper we show that the lower bound proof has a gap in it. Although we do not provide a definitive answer to its validity. More positively, we show one can obtain "fast" sample complexity bounds for nonconvex F for "most" target conditional expectations. The new bounds depend on the detailed geometry of F, in particular the distance in a certain sense of the target's conditional expectation from the set of nonuniqueness points of the class F.

1 Introduction

The agnostic learning model [6] is a generalization of the PAC learning model that does not presume the target function lies within the space of functions (hypotheses) used for learning. There are now a number of results concerning the sample complexity of agnostic learning, especially with respect to the squared loss functional. In particular, in [9] it was shown that if ε is the required accuracy, then the sample complexity (ignoring log factors and the confidence terms) of agnostic learning from a closed class of functions F with squared loss is $O(d/\varepsilon)$ if F is convex, where d is an appropriate complexity parameter (e.g. the empirical metric entropy of the class). This result was extended and improved in [10].

It was claimed in [9] that if F is not convex, there exists a lower bound of $\Omega(1/\varepsilon^2)$ on the sample complexity. Thus, whether or not F is convex seemed important for the sample complexity of agnostic learning with squared loss.

However, these are deceptive results. The claimed lower bound relies on a random construction and the fact that for nonconvex F, one can always find a target "function" (actually a target conditional expectation) f^* which has two best approximations in the class F. Unfortunately, as we show here, the random construction has a gap in the proof.

It *is* the case though that sample complexity of agnostic learning does depend on the closeness of f^* to a point with a nonunique best approximation. In this paper we will develop some *nonuniform* results which hold for "most" target conditional expectations in the agnostic learning scenario from a nonconvex class F and obtain sharper sample complexity upper bounds. The proof

J. Kivinen and R. H. Sloan (Eds.): COLT 2002, LNAI 2375, pp. 1–13, 2002.

we present here is based on recently developed methods which can be used for complexity estimates. It was shown in [10] that the complexity of a learning problem can be governed by two properties. The first is the Rademacher complexity of the class, which is a parameter that indicates "how large" the class is (see [11,1]). The other property is the ability to control the mean square value of each loss function using its expectation. We will show that indeed the mean square value can be bounded in terms of the expectation as long as as one knows the distance of the target from the set of points which have more than a unique best approximation in the class.

In the next section we present some basic definitions, notation, and some general complexity estimates. Then, we present our nonuniform upper bound. Finally, we briefly present the proof of the lower bound claimed in [9] and show where there is a gap in the argument.

Thus the present paper does not completely resolve the question of sample complexity for agnostic learning for squared loss. The lower bound proof of [9] may be patchable: $O(1/\varepsilon^2)$ may be the best *uniform* lower bound one can achieve. What is clear from the present paper is the crucial role the set of nonuniqueness points of F plays in the sample complexity of agnostic learning with squared loss.

2 Definitions, Notation and Background Results

If $(\mathcal{X}, d)$ is a metric space, and $U \subseteq \mathcal{X}$, then for $\varepsilon > 0$, we say that $C \subseteq \mathcal{X}$ is an *ε-cover* of U with respect to d if for all $v \in U$, there exists $w \in C$ such that $d(v, w) \leq \varepsilon$. The *ε-covering number* with respect to d, $N(\varepsilon, U, d)$, is the cardinality of the smallest ε-cover of U with respect to d. If the metric d is obvious, we will simply say ε-cover etc.

The closed ball centered at c of radius r is denoted by $B(c,r) := \{x \in \mathcal{X} \colon \|x - c\| \leq r\}$. Its boundary is $\partial B(c,r) := \{x \in \mathcal{X} \colon \|x - c\| = r\}$. If $x \in \mathcal{X}$, and $A \subset \mathcal{X}$, let the distance between A and x be defined as $d_A(x) := \inf\{d(x,a) \colon a \in A\}$. The *metric projection* of x onto A is $P_A(x) := \{a \in A \colon \|x - a\| = d_A(x)\}$. Hence, elements of $P_A(x)$ are all *best approximations* of x in A.

Denote by $L_\infty(\mathcal{X})$ the space of bounded functions on $\mathcal{X}$ with respect to the sup norm and set $B\big(L_\infty(\mathcal{X})\big)$ to be its unit ball. Let μ be a probability measure on $\mathcal{X}$ and put $L_2(\mu)$ to be the Hilbert space of functions from $\mathcal{X}$ to $\mathbb{R}$ with the norm endowed by the inner product $\langle f, g\rangle = \int f(x)g(x)\mathrm{d}\mu(x)$. Let $\mathcal{Y} \subset [-1,1]$, and set F to be a class of functions from $\mathcal{X}$ to $\mathcal{Y}$, and thus a subset of $L_2(\mu)$. Assumptions we will make throughout are that F is a closed subset of $L_2(\mu)$ and that it satisfies a measurability condition called "admissibility" (see [4,5,15]) for details.

Definition 1. *Let $F \subset L_2(\mu)$. A point $f \in L_2(\mu)$ is said to be a nup point (nonunique projection) of F with respect to (w.r.t.) $L_2(\mu)$ if it has two or more best approximations in F with respect to the $L_2(\mu)$ norm. Define*

$$\mathrm{nup}(F,\mu) := \{f \in L_2(\mu) : f \textit{ is a nup point of } F \textit{ w.r.t. } L_2(\mu)\}.$$

It is possible to show that in order to solve the agnostic learning problem of approximating a random variable Y with values in $\mathcal{Y}$ by elements in F, it suffices to learn the function $f^* = \mathbb{E}(Y|X=x)$. Indeed, for every $f \in F$,

$$\begin{aligned}\mathbb{E}\big(f(X)-Y\big)^2 &= \mathbb{E}\big(\mathbb{E}(Y|X)-f(X)\big)^2 + \mathbb{E}\big(\mathbb{E}(Y|X)-Y\big)^2 \\ &= \mathbb{E}\big(f^*(X)-f(X)\big)^2 + \mathbb{E}\big(f^*(X)-Y\big)^2.\end{aligned}$$

Thus, a minimizer of the distance between $f(X)$ and Y will depend only on finding a minimizer for $\mathbb{E}\big(f^*(X)-f(X)\big)^2$, that is, solving the function learning problem of approximating f^* by members of F with respect to the $L_2(\mu)$ norm.

Assume that we have fixed the target f^*. We denote by f_a its best approximation in F with respect to the given $L_2(\mu)$ norm. (Of course f_a is unique only if $f^* \notin \mathrm{nup}(F,\mu)$.) For any function $f \in F$, let the squared loss function associated with f^* and f be

$$g_{f,f^*} : x \mapsto (f(x)-f^*(x))^2 - (f_a(x)-f^*(x))^2,$$

and set $\mathcal{L}(f^*) = \mathcal{L} := \{g_{f,f^*} : f \in F\}$.

Interestingly, although a "randomly chosen" $f^* \in L_2(\mu)$ is very unlikely[1] to be in $\mathrm{nup}(F,\mu)$, as we shall see below, $\mathrm{nup}(F,\mu)$ nevertheless controls the sample complexity of learning f^* for all $f^* \in L_2(\mu) \setminus F$.

Definition 2. *For any set $\{x_1,\ldots,x_n\} \subset \mathcal{X}$, let μ_n be the empirical measure supported on the set; i.e. $\mu_n = \frac{1}{n}\sum_{i=1}^n \delta_{x_i}$. Given a class of functions F, a random variable Y taking values in $\mathcal{Y}$, and parameters $0 < \varepsilon, \delta < 1$ let $C_F(\varepsilon,\delta,Y)$ be the smallest integer such that for any probability measure μ*

$$\Pr\{\exists g_{f,f^*} \in \mathcal{L}(f^*) : \mathbb{E}_{\mu_n} g_{f,f^*} < \varepsilon, \mathbb{E}_\mu g_{f,f^*} \geq 2\varepsilon\} < \delta, \tag{1}$$

where $f^* = \mathbb{E}(Y|X=x)$.

The quantity $C_F(\varepsilon,\delta,Y)$ is known as the *sample complexity of learning a target Y with the function class F*. The definition means that if one draws a sample of size greater than $C_F(\varepsilon,\delta,Y)$ then with probability greater than $1-\delta$, if one "almost minimizes" the empirical loss (less than ε) then the expected loss will not be greater than 2ε. Typically, the sample complexity of a class is defined as the "worst" sample complexity when going over all possible selections of Y.

Recent results have yielded good estimates on the probability of the set in (1). These estimates are based on the Rademacher averages as a way of measuring the complexity of a class of functions. The averages are better suited to proving sample complexity results than classical techniques using the union bound over an ε-cover, mainly because of the "functional Bennett inequality" due to Talagrand [14].

[1] Since Hilbert spaces are uniformly convex it follows from a theorem of Stechkin [13] (see [17, page 9] or [3, page 29]) that $L_2(\mu) \setminus \mathrm{nup}(F,\mu)$ is a countable intersection of open dense sets. This implies that if one puts a reasonable probability measure ν on $L_2(\mu)$, then $\nu(\{f \in L_2(\mu) : f \notin \mathrm{nup}(F,\mu)\}) = 1$.

Definition 3. *Let μ be a probability measure on $\mathcal{X}$ and suppose F is a class of uniformly bounded functions. For every integer n, set*

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

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.

Various relationships between Rademacher averages and classical measures of complexity are shown in [11,12]. It turns out that the best sample complexity bounds to date are in terms of *local* Rademacher averages. Before presenting these bounds, we require the next definition.

Definition 4. *We say that $F \subset L_2(\mu)$ is star-shaped with centre f if for every $g \in F$, the interval $[f,g] = \{tf + (1-t)g : 0 \leq t \leq 1\} \subset F$. Given F and f, let*

$$\mathrm{star}(F, f) := \bigcup_{g \in F} [f, g].$$

Theorem 1. *Let $F \subset B\big(L_\infty(\mathcal{X})\big)$, fix some f^* bounded by 1 and set $\mathcal{L}(f^*)$ to be the squared loss class associated with F and f^*. Assume that there is a constant B such that for every $g \in \mathcal{L}(f^*)$, $\mathbb{E}g^2 \leq B\mathbb{E}g$.*

Let $\mathcal{G} := \mathrm{star}(\mathcal{L}, 0)$ and for every $\varepsilon > 0$ set $\mathcal{G}_\varepsilon = \mathcal{G} \cap \{h : \mathbb{E}h^2 \leq \varepsilon\}$. Then for every $0 < \varepsilon, \delta < 1$,

$$\Pr\{\exists g \in \mathcal{L}, \mathbb{E}_{\mu_n} g \leq \varepsilon/2, \mathbb{E}g \geq \varepsilon\} \leq \delta$$

provided that

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

where C is an absolute constant.

Using this result one can determine an upper bound on the sample complexity in various cases. The one we present here is a bound in terms of the metric entropy of the class.

Theorem 2 ([12]). *Let Y be a random variable on $\mathcal{Y}$ and put $f^* = \mathbb{E}(Y|X = x)$. Let F,$\mathcal{L}$, $\mathcal{G}$ and B be as in theorem 1.*

1. If there are $\gamma, p, d \geq 1$ such that for every $\varepsilon > 0$,

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

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

$$C(\varepsilon, \delta, Y) \leq \frac{C_{p,\gamma}}{\varepsilon} \max\Big\{ d \log^p \frac{1}{\varepsilon}, B \log \frac{2}{\delta} \Big\},$$

where $C_{p,\gamma}$ depends only on p and γ.

2. If there are $0 < p < 2$ and $\gamma \geq 1$ such that for every $\varepsilon > 0$,

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

then

$$C(\varepsilon, \delta, Y) \leq C_{p,\gamma} \max\Big\{ \big(\frac{1}{\varepsilon}\big)^{1+\frac{p}{2}}, B \log \frac{2}{\delta} \Big\},$$

where $C_{p,\gamma}$ depends only on p and γ.

From this result it follows that if the original class F is "small enough", one can establish good generalization bounds, if, of course, the mean-square value of each member of the loss class can be uniformly controlled by its expectation. This is trivially the case in the proper learning scenario, since each loss function is nonnegative. It was known to be true if F is convex in the squared loss case [9] and was later extended in the more general case of p-loss classes for $2 \leq p < \infty$ [10].

Our aim is to investigate this condition and to see what assumptions must be imposed on f^* to ensure such a uniform control of the mean square value in terms of the expectation.

3 Nonuniform Agnostic Learnability of Nonconvex Classes

We will now study agnostic learning using nonconvex hypothesis classes. The key observation is that whilst in the absence of convexity one can not control $\mathbb{E}[g^2_{f,f^*}]$ in terms of $\mathbb{E}[g_{f,f^*}]$ *uniformly* in f^*, one can control it *nonuniformly* in f^* by exploiting the geometry of F. The main result is corollary 1.

The following result is a generalization of [8, lemma 14] (cf. [7, lemma A.12]).

Lemma 1. *Let F be a class of functions from $\mathcal{X}$ to $\mathcal{Y}$. Put $\alpha \in [0,1)$, set $f^* \in L_2(\mu)$ and suppose f^* has range contained in $[0,1]$. If for every $f \in F$*

$$\langle f_a - f^*, f_a - f \rangle \leq \frac{\alpha}{2} \|f_a - f\|^2, \tag{2}$$

then for every $g_{f,f^*} \in \mathcal{L}(f^*)$,

$$\mathbb{E}[g^2_{f,f^*}] \leq \frac{16}{1-\alpha} \mathbb{E}[g_{f,f^*}].$$

Proof. For the sake of simplicity, we denote each loss function by g_f. Observe that

$$\begin{aligned} \mathbb{E}[g_f^2] &= \mathbb{E}[((f^*(X) - f(X))^2 - (f^*(X) - f_a(X))^2)^2] \\ &= \mathbb{E}[((2f^*(X) - f(X) - f_a(X))(f_a(X) - f(X)))^2] \\ &\leq 16\mathbb{E}[(f(X) - f_a(X))^2] \\ &= 16\|f_a - f\|^2. \end{aligned} \tag{3}$$

Furthermore,

$$\begin{aligned}
\mathbb{E}[g_f] &= \mathbb{E}[(f^*(X) - f(X))^2 - (f^*(X) - f_a(X))^2] \\
&= \mathbb{E}[(f^*(X) - f_a(X))^2 + (f_a(X) - f(X))^2 \\
&\qquad + 2(f^*(X) - f_a(X))(f_a(X) - f(X)) - (f^*(X) - f_a(X))^2] \\
&= \mathbb{E}[(f_a(X) - f(X))^2 + 2(f^*(X) - f_a(X))(f_a(X) - f(X))] \\
&= \mathbb{E}[(f_a(X) - f(X))^2] + 2\mathbb{E}[(f^*(X) - f_a(X))(f_a(X) - f(X))] \\
&= \|f_a - f\|^2 + 2\langle f^* - f_a, f_a - f\rangle \\
&= \|f_a - f\|^2 - 2\langle f_a - f^*, f_a - f\rangle \\
&\geq \|f_a - f\|^2 - \alpha\|f_a - f\|^2 \\
&= (1-\alpha)\|f_a - f\|^2 \\
&= \frac{1-\alpha}{16}\mathbb{E}[g_f^2].
\end{aligned}$$

□

Lemma 2. *Fix $f^* \in L_2(\mu)$. Then, $f \in L_2(\mu)$ satisfies (2) if and only if f is not contained in*

$$B^{(\alpha)} := B\left(\frac{1}{\alpha}(f^* - f_a) + f_a, \frac{1}{\alpha}\|f^* - f_a\|\right),$$

which is the closed ball in $L_2(\mu)$ centered at $\frac{1}{\alpha}(f^* - f_a) + f_a$ with radius $\frac{1}{\alpha}\|f^* - f_a\|$.

Proof. Note that $\langle f^* - f_a, f - f_a\rangle \leq \frac{\alpha}{2}\|f_a - f\|^2$ if and only if $\|f_a - f\|^2 - \frac{2}{\alpha}\langle f^* - f_a, f - f_a\rangle \geq 0$. Clearly, the latter is equivalent to

$$\begin{aligned}
\langle f_a - f, f_a - f\rangle + \langle\frac{1}{\alpha}(f^* - f_a), \frac{1}{\alpha}(f^* - f_a)\rangle - \langle\frac{1}{\alpha}(f^* - f_a), \frac{1}{\alpha}(f^* - f_a)\rangle \\
+ \frac{2}{\alpha}\langle f^* - f_a, f_a - f\rangle \geq 0.
\end{aligned}$$

Thus,

$$\langle f_a - f + \frac{1}{\alpha}(f^* - f_a), f_a - f + \frac{1}{\alpha}(f^* - f_a)\rangle \geq \langle\frac{1}{\alpha}(f^* - f_a), \frac{1}{\alpha}(f^* - f_a)\rangle. \quad (4)$$

Clearly, f satisfies (4) if and only if $\|f - (f_a + \frac{1}{\alpha}(f^* - f_a))\| \geq \frac{1}{\alpha}\|f^* - f_a\|$; hence it belongs to the region outside of $B^{(\alpha)}$. □

In the limit as $\alpha \to 0$, ∂B_α approaches a hyperplane. Then by the unique supporting hyperplane characterization of convex sets [16, theorem 4.1] this implies F is convex.

We will use lemma 2 as indicated in figure 1. The key factor in bounding B is the closeness of f^* to $\mathrm{nup}(F, \mu)$ in a particular sense. Suppose $f^* \in L_2(\mu) \setminus (F \cup \mathrm{nup}(F, \mu))$, and let

$$r_{F,\mu}(f^*) := \inf\{\|f - P_F(f^*)\| : f \in \{\lambda(f^* - P_F(f^*)) : \lambda > 0\} \cap \mathrm{nup}(F, \mu)\}.$$

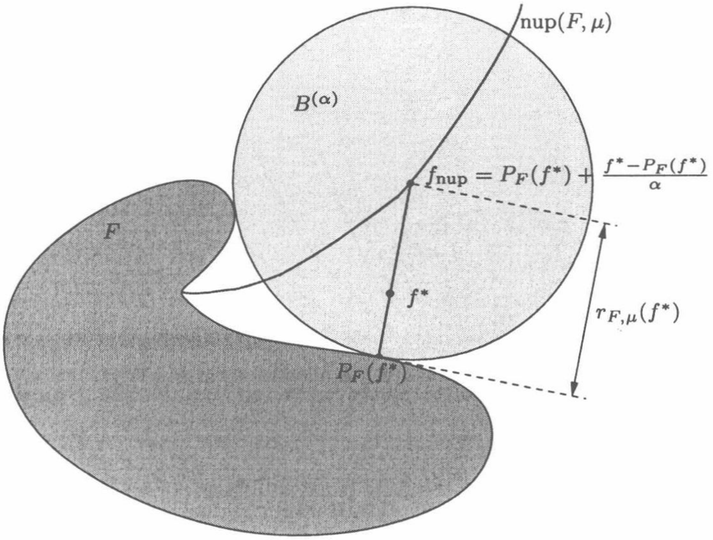
nup(F, μ)
$B^{(\alpha)}$
$f_{\mathrm{nup}} = P_F(f^*) + \frac{f^* - P_F(f^*)}{\alpha}$
F
f^*
$r_{F,\mu}(f^*)$
$P_F(f^*)$

Fig. 1. Illustration of lemma 2 and the definition of $r_{F,\mu}(f^*)$

Observe that $r_{F,\mu}(f^*) = \|f_{\mathrm{nup}} - P_F(f^*)\|$ where f_{nup} is the point in $\mathrm{nup}(F, \mu)$ found by extending a ray ρ from $P_F(f^*)$ through f^* until reaching $\mathrm{nup}(F, \mu)$ (see Figure 1). Let

$$\alpha_{F,\mu}(f^*) := \frac{\|f^* - P_F(f^*)\|}{r_{F,\mu}(f^*)} = \frac{\|f^* - P_F(f^*)\|}{\|f_{\mathrm{nup}} - P_F(f^*)\|}$$

and observe that $\alpha_{F,\mu}(f^*) \in [0, 1]$ is the largest α such that $B_F^{(\alpha)}(f^*)$ only intersects F at $P_F(f^*)$ and as f^* "approaches" $\mathrm{nup}(F, \mu)$ along ρ, $\alpha_{F,\mu}(f^*) \to 1$.

Note that if F is convex then $\mathrm{nup}(F, \mu)$ is the empty set; hence for all $f^* \in L_2(\mu)$, $r_{F,\mu}(f^*) = \infty$ and $\alpha_{F,\mu}(f^*) = 0$.

Combining theorem 2 with lemmas 1 and 2 leads to our main positive result:

Corollary 1. *Let $F \subset L_2(\mu)$ be a class of functions into $[0, 1]$, set Y to be a random variable taking its values in $[0, 1]$, and assume that $f^* = \mathbb{E}(Y|X) \notin \mathrm{nup}(F, \mu)$. Assume further that there are constants $d, \gamma, p \geq 1$ such that for every empirical measure μ_n, $\log N(\varepsilon, F, L_2(\mu_n)) \leq d \log^p(\gamma/\varepsilon)$. Then, there exists a constant $C_{p,\gamma}$, which depends only on p and γ, such that for every $0 < \varepsilon, \delta < 1$,*

$$C_F(\varepsilon, \delta, Y) \leq \frac{C_{p,\gamma}}{\varepsilon} \max\left\{ d \log^p \frac{1}{\varepsilon}, \frac{\log \frac{2}{\delta}}{1 - \alpha_{F,\mu}(f^*)} \right\}.$$

Note that this result is *non-uniform* in the target Y because some functions f^* are harder to learn than others. For all $f^* \in F^\alpha := \{f \in H \colon \alpha_{F,\mu}(f) \geq \alpha\}$, one

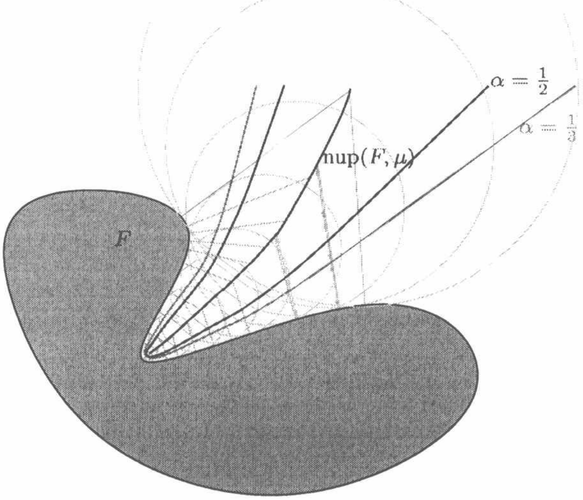

Fig. 2. Illustration of the sets F^α. The lines marked $\alpha = \frac{1}{3}$ and $\alpha = \frac{1}{2}$ are the boundaries of $F^{1/2}$ and $F^{1/3}$ respectively

obtains a uniform bound in terms of α. Figure 2 illustrates the boundaries of F^α for a given F and two different values of α. If F is convex, then $\alpha_{F,\mu}(f^*) = 0$ always and one recovers a completely uniform result.

4 The Lower Bound

In this section we present the geometric construction which led to the claimed lower bound presented in [9]. We then show where there is a gap in the proof and the bound can not be true for function learning. In our discussion we shall use several geometric results, the first of which is the following standard lemma, whose proof we omit.

Lemma 3. *Let X be a Hilbert space and set $x \in X$ and $r > 0$. Put $B_1 = B(x,r)$ and let $y \in \partial B_1$. For any $0 < t < 1$ let $z_t = tx + (1-t)y$ and set $B_2 = B(z_t, \|z_t - y\|)$. Then $B_2 \subset B_1$ and $\partial B_1 \cap \partial B_2 = \{x\}$.*

Using lemma 3 it is possible to show that even if x has several best approximations in G, then any point on the interval connecting x and any one of the best approximations of x has a unique best approximation.

Corollary 2. *Let $x \in X$, set $y \in P_G(x)$ and for every $0 < t < 1$ let $z_t = tx + (1-t)y$. Then, $P_G(z_t) = y$.*

Proof. We begin by showing that $P_G(z_t) \subset P_G(x)$. To that end, note that $d_G(z_t) = \|z_t - y\| = (1-t)\|x - y\|$. Indeed, $d_G(z_t) \le \|z_t - y\|$. If there is