Using the Doubling Dimension to Analyze the Generalization of Learning Algorithms

Nader H. Bshouty*

Technion

Yi Li

Genome Institute of Singapore

Philip M. Long

Google

Abstract

Given a set F of classifiers and a probability distribution over their domain, one can define a metric by taking the distance between a pair of classifiers to be the probability that they classify a random item differently. We prove bounds on the sample complexity of PAC learning in terms of the doubling dimension of this metric. These bounds imply known bounds on the sample complexity of learning halfspaces with respect to the uniform distribution that are optimal up to a constant factor.

We then prove a bound that holds for any algorithm that outputs a classifier with zero error whenever this is possible; this bound is in terms of the maximum of the doubling dimension and the VC-dimension of F and

*Part of this research was done while the first author was visiting Google in Mountain View, California.

Email addresses: bshouty@cs.technion.ac.il (Nader H. Bshouty), liy3@gis.a-star.edu.sg (Yi Li), plong@google.com (Philip M. Long)

Preprint submitted to Journal of Computer and System Sciences *January 23, 2009*

strengthens the best known bound in terms of the VC-dimension alone.

Finally, we show that there is no bound on the doubling dimension of halfspaces in $\mathbf{R}^n$ in terms of n that holds independently of the domain distribution. This implies that there is no such a bound in terms of the VC-dimension of F (in contrast with the metric dimension).

Key words: PAC learning, generalization, doubling dimension, doubling metric, learning theory, statistical learning theory, local complexity.
2000 MSC: 68Q32, 68T05

1. Introduction

A set F of classifiers and a probability distribution D over their domain X induce a metric ρ_D in which the distance between classifiers is the probability that they disagree on how to classify a random object. Properties of metrics like this have long been used for analyzing the generalization ability of learning algorithms [10, 33]. This paper is about bounds on the number of examples required for PAC learning in terms of the doubling dimension [3] of this metric space.

The doubling dimension of a metric space is the least d such that any ball can be covered by 2^d balls of half its radius. The doubling dimension has been frequently used lately in the analysis of algorithms [12, 20, 21, 18, 30, 13, 9, 22, 29, 7].

In the PAC-learning model, an algorithm is given examples

$$(x_1, f(x_1)), ..., (x_m, f(x_m))$$

of the behavior of an arbitrary member f of a known class F. The items $x_1, ..., x_m$ are chosen independently at random according to D. The algorithm must, with probability at least $1-\delta$ (w.r.t. to the random choice of $x_1, ..., x_m$), outputs a classifier whose distance from f is at most ϵ.

We show that if (F, ρ_D) has doubling dimension $d(F, D)$, then F can be PAC-learned with respect to D using

$$O\left(\frac{d(F,D)}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right) \tag{1}$$

examples.

The ϵ-doubling dimension of a metric space is d_ϵ such that any ball of radius greater than ϵ can be covered by 2^{d_ϵ} balls of half its radius. We also prove a bound of

$$O\left(\frac{d_{c\epsilon}(F,D)}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right) \tag{2}$$

for an absolute constant c.

If the VC-dimension of F and the doubling dimension of (F, ρ_D) are both at most d, we show that any algorithm that outputs a classifier with zero training error whenever this is possible PAC-learns F w.r.t. D using

$$O\left(\frac{d}{\epsilon}\sqrt{\log\frac{1}{\epsilon}} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right) \tag{3}$$

examples. This compares favorably with the best possible bound of this sort in terms of the VC-dimension alone [33, 8]:

$$O\left(\frac{\mathrm{VC}(F)}{\epsilon}\log\frac{1}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right). \tag{4}$$

(Note, however, that the bound in terms of the VC-dimension alone holds uniformly over all distributions D.)

We then show that if F consists of halfspaces through the origin, and D is the uniform distribution over the unit ball in $\mathbf{R}^n$, then the doubling dimension of (F, ρ_D) is $O(n)$. Thus (1) generalizes the known bound of

$$O\left(\frac{n}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right)$$

for learning halfspaces with respect to the uniform distribution [25], matching a known lower bound for this problem [24] up to a constant factor. The consequences of both (1) and (3) regarding learning halfspaces under the uniform distribution improve on the consequence of (4). Since if there is a halfspace with zero training error, such a halfspace can be found in polynomial-time using linear programming, the bound (3) can be achieved by a polynomial-time algorithm, and (3) is the first improvement over (4) that can be obtained by a polynomial-time algorithm.

Some previous analyses of the sample complexity of learning have made use of the fact that the "metric dimension" [19] is at most the VC-dimension [10, 14]. A bound of

$$d_\epsilon(F, D) \leq \mathrm{VC}(F)\log\frac{1}{\epsilon} + O(\mathrm{VC}(F))$$

follows from these metric dimension bounds. One might have hoped that this could be strengthened, leading to improvements on (4) by applying (2). We show that this is not the case: it is possible to pack $(1/\alpha)^d$ classifiers in a set F of VC-dimension d so that the distance between every pair is in the interval $(\alpha, 2\alpha]$. This implies that there is a class of classifiers F such that

$$d_\epsilon(F, D) \geq \mathrm{VC}(F)\log\frac{1}{\epsilon}.$$

We also establish a separation using halfspaces, a hypothesis class frequently used in practice.

Our analysis, like others, views random examples as a means to eliminate candidate classifiers. When the goal is to obtain a classifier with error rate at most ϵ, candidates with error rates slightly greater than ϵ are the most dangerous, because they are the hardest to discover. Bounding the doubling dimension is useful for analyzing the sample complexity of learning because it limits the richness of a subclass of F near the classifier to be learned, i.e. the most dangerous candidates. For other analyses that exploit bounds on such local richness, please see [32, 31, 4, 25, 26, 36]. Benedek and Itai [6] analyzed learning by first approximating a set of hypotheses with a finite cover, and then choosing the best performer in the cover, as we do in the proof of (1).

2. Preliminaries

2.1. Learning

For some domain X, an *example* consists of a member of X, and its classification in $\{0, 1\}$. A *classifier* is a mapping from X to $\{0, 1\}$. A *training set* is a finite collection of examples. A *learning algorithm* takes as input a training set, and outputs a classifier.

Suppose D is a probability distribution over X. Then define

$$\rho_D(f, g) = \mathbf{Pr}_{x \sim D}[f(x) \neq g(x)]$$

(which is a special case of the L_1 metric between random variables). A learning algorithm A PAC learns F w.r.t. D with accuracy $1 - \epsilon$ and confidence $1 - \delta$ from m examples if, for any $f \in F$, if

- domain elements $x_1, ..., x_m$ are drawn independently at random according to D, and
- $(x_1, f(x_1)), ..., (x_m, f(x_m))$ is passed to A, which outputs h,

then

$$\mathbf{Pr}[\rho_D(f, h) > \epsilon] \leq \delta.$$

If F is a set of classifiers, a learning algorithm is a *consistent hypothesis finder for F* if it outputs an element of F that correctly classifies all of the training data whenever it is possible to do so.

2.2. Metrics, Doubling Dimension and VC-Dimension

Let $\Phi = (Z, \rho)$ be a metric space. An *α-cover* for Φ is a set $T \subseteq Z$ such that every element of Z has a counterpart in T that is at a distance at most α (with respect to ρ). An *α-packing* for Φ is a set $T \subseteq Z$ such that every pair of elements of T are at a distance greater than α (again, with respect to ρ). The *α-ball* centered at $z \in Z$, denoted by $B(z, \alpha)$, consists of all $t \in Z$ for which $\rho(z, t) \leq \alpha$.

Denote the size of the largest α-packing by $\mathcal{M}(\alpha; \Phi)$.

Lemma 1 ([19]). *For any metric space $\Phi = (Z, \rho)$, and any $\alpha > 0$, there is an α-packing for Φ that is also an α-cover.*

The *ϵ-doubling dimension* of Φ is the least d such that, for all radii $\alpha > \epsilon$, any α-ball in Φ can be covered by at most 2^d $\alpha/2$-balls. That is, for any $\alpha > \epsilon$ and any $z \in Z$, there is a $C \subseteq Z$ such that

- $|C| \leq 2^d$, and

- $\{t \in Z : \rho(z,t) \leq \alpha\} \subseteq \cup_{c \in C}\{t \in Z : \rho(c,t) \leq \alpha/2\}$.

The *doubling dimension* is the 0-doubling dimension.

The doubling dimension limits the extent to which members of a metric space that are separated from one another can crowd around one element.

Lemma 2 (see [12]). *Suppose $\Phi = (Z, \rho)$ is a metric space with doubling dimension d and $z \in Z$. Then*

$$\mathcal{M}(\alpha; B(z,\beta)) \leq \left(\frac{4\beta}{\alpha}\right)^d.$$

In other words, any α-packing must have at most $(4\beta/\alpha)^d$ elements within distance β of z.

The above bound is also true for $d = d_{\alpha/2}(F, D)$.

The VC-dimension, $\mathrm{VC}(F)$ of a set F of $\{0,1\}$-valued functions with a common domain is the size of the largest set $x_1, ..., x_d$ of domain elements such that

$$\{(f(x_1), ..., f(x_d)) : f \in F\} = \{0,1\}^d.$$

Haussler in [14] gives a bound for the size of largest α-packing in term of the VC-dimension

Lemma 3. *([14]) For any metric space (F, ρ_D) we have*

$$\mathcal{M}(\alpha; F) \leq e(\mathit{VC}(F) + 1)\left(\frac{2e}{\alpha}\right)^{\mathit{VC}(F)}.$$

He also gave a randomized construction of a class of classifiers F that satisfies

$$\mathcal{M}(\alpha; F) \geq \left(\frac{1}{2e\alpha}\right)^{\mathrm{VC}(F)}.$$

In this paper we give a deterministic construction of a class of classifiers F that satisfies

$$\mathcal{M}(\alpha; F) \geq (1-\alpha)\left(\frac{1}{\alpha}\right)^{\mathrm{VC}(F)}$$

while simultaneously satisfying the constraint that $\rho(f,g) \leq 2\alpha$ for all $f, g \in F$.

2.3. Probability

For a function ψ and a probability distribution D, let $\mathbf{E}_{x\sim D}[\psi(x)]$ be the expectation of ψ w.r.t. D. We will shorten this to $\mathbf{E}_D[\psi]$, and if $\mathbf{u} = (u_1, ..., u_m) \in X^m$, then

$$\hat{\mathbf{E}}_{\mathbf{u}}[\psi] = \frac{1}{m}\sum_{i=1}^{m} \psi(u_i).$$

We will use $\mathbf{Pr}_{x\sim D}$, $\mathbf{Pr}_D$, and $\hat{\mathbf{Pr}}_{\mathbf{u}}$ similarly.

In many places in the paper we will use the following Chernoff bounds (see [27]).

Lemma 4. *Let $X_1, \ldots, X_n$ be independent Bernoulli trials, where*

$$\mathbf{Pr}[X_i = 1] \leq p_i.$$

Let $X = \sum_{i=1}^n X_i$ and $\mu = \sum_{i=1}^n p_i$. For any $\eta > 0$ we have

$$\mathbf{Pr}[X > (1+\eta)\mu] < \left(\frac{e^{\eta}}{(1+\eta)^{1+\eta}}\right)^{\mu}.$$

This implies the following: For $\eta \leq 2e-1$

$$\mathbf{Pr}[X > (1+\eta)\mu] < e^{-\mu\eta^2/4}$$

.

Let $X_1, \ldots, X_n$ be independent Bernoulli trials, where

$$\mathbf{Pr}[X_i = 1] \geq p_i.$$

Let $X = \sum_{i=1}^n X_i$ and $\mu = \sum_{i=1}^n p_i$. For any $\eta > 0$ we have

$$\mathbf{Pr}[X < (1 - \eta)\mu] < e^{-\mu\eta^2/2}.$$

3. The strongest upper bound

The proof of our strongest upper bound is an application of the peeling technique [1] (see [31]).

Theorem 5. *Suppose $d(F, D)$ is the doubling dimension of (F, ρ_D). There is an algorithm A that PAC-learns F with respect to D with accuracy $1 - \epsilon$ and confidence $1 - \delta$ from*

$$O\left(\frac{d(F, D)}{\epsilon} + \frac{1}{\epsilon} \log \frac{1}{\delta}\right)$$

examples.

The above statement is also true if $d(F, D)$ is replaced with $d_{\epsilon/4}(F, D)$.

Proof: Let G be an $\epsilon/4$-packing for (F, ρ_D) that is also an $\epsilon/4$-cover (the existence of such a G is implied by Lemma 1).

Let f be an arbitrary target function. For any classifier g, define the *error rate* of g to be $\mathbf{Pr}_{x \sim D}(g(x) \neq f(x))$. Consider a learning algorithm A that takes a random training set S resulting from drawing

$$m = O\left(\frac{d}{\epsilon} + \frac{1}{\epsilon} \log \frac{1}{\delta}\right)$$

examples according to D, where $d = d_{\epsilon/4}(F, D)$, and classifying them using f. The learning algorithm then outputs the element of G with minimum error on the training set, that is

$$\text{argmin}_{g \in G} |\{(x, y) \in S : g(x) \neq y\}|.$$

We wish to show that algorithm A is a PAC-learning algorithm for F with respect to D. First, we observe that some classifier in G has small error rate, which will imply that it is likely that some classifier in G make incorrect classifications on a small fraction of the training data. Finally, the main part of the argument will show that it is likely that any classifier with small training error has a small error rate with respect to the underlying distribution D.

Whatever the target, since G is an $\epsilon/4$-cover of (F, ρ_D), some element of G has error rate at most $\epsilon/4$. Applying Lemma 4, $O((1/\epsilon)\log(1/\delta))$ examples are sufficient that, with probability at least $1 - \delta/2$, this classifier is incorrect on at most a fraction $\epsilon/2$ of the training data. Thus, the training error of the hypothesis output by A is at most $\epsilon/2$ with probability at least $1 - \delta/2$. Thus, the probability that the error rate of the output of A is greater than ϵ is no more than the probability that any classifier with error rate greater than ϵ has training error at most $\epsilon/2$.

Define $\rho_S(g, h)$ to be the fraction of examples in S on which g and h disagree. We have

$$\begin{aligned}
&\mathbf{Pr}[\exists g \in G,\ \rho_D(g, f) > \epsilon \text{ and } \rho_S(g, f) \leq \epsilon/2] \\
&\leq \sum_{k=0}^{\lfloor \log(1/\epsilon) \rfloor} \mathbf{Pr}[\exists g \in G,\ 2^k \epsilon < \rho_D(g, f) \leq 2^{k+1}\epsilon \text{ and } \rho_S(g, f) \leq \epsilon/2]
\end{aligned}$$

$$\leq \sum_{k=0}^{\lfloor \log(1/\epsilon) \rfloor} |\{g \in G : 2^k \epsilon < \rho_D(g,f) \leq 2^{k+1}\epsilon\}| \times \max_{g \in G: \rho_D(g,f) > 2^k \epsilon} \mathbf{Pr}[\rho_S(g,f) \leq \epsilon/2]$$

$$\leq \sum_{k=0}^{\infty} 2^{(k+4)d} e^{-2^k \epsilon m/8}$$

by Lemma 2 and Lemma 4.

Each of the following steps is a straightforward manipulation: For

$$d \leq \frac{\epsilon m}{64} \text{ and } m \geq \frac{32}{\epsilon} \log \frac{2}{\delta}$$

we have

$$\sum_{k=0}^{\infty} 2^{(k+4)d} e^{-2^k \epsilon m/8} \leq 2^{4d} \sum_{k=0}^{\infty} \left(2^{k-2^{k+3}}\right)^{\frac{\epsilon m}{64}} \leq 2^{4\frac{\epsilon m}{64} - 6\frac{\epsilon m}{64}} \leq 2^{-\frac{\epsilon m}{32}} \leq \frac{\delta}{2}.$$

This completes the proof. □

4. Halfspaces and the uniform distribution

In this section, we illustrate the application of learning results concerning the doubling dimension using the case of learning halfspaces with respect to the uniform distribution. The last paragraph of the proof mirrors the usual proof that a metric with a "doubling measure" has finite doubling dimension (see [16, 21]).

Proposition 6. *If U_n is the uniform distribution over the unit ball in $\mathbf{R}^n$, and H_n is the set of halfspaces that go through the origin, then the doubling dimension of (H_n, ρ_{U_n}) is $O(n)$.*

Proof: Choose $h \in H_n$ and $\alpha > 0$. We will show that the ball of radius α centered at h can be covered by $2^{O(n)}$ balls of radius $\alpha/2$.

Suppose U_{H_n} is the probability distribution over H_n obtained by choosing a normal vector $\mathbf{w}$ uniformly from the unit ball, and outputting the halfspace $\{\mathbf{x} : \mathbf{w} \cdot \mathbf{x} \geq 0\}$. The argument will be a "volume argument" using U_{H_n}.

It is known (see Lemma 4 of [25]) that

$$\mathbf{Pr}_{g \sim U_{H_n}}[\rho_{U_n}(g, h) \leq \alpha/4] \geq (c_1 \alpha)^{n-1}$$

where $c_1 > 0$ is an absolute constant independent of α and n. Furthermore,

$$\mathbf{Pr}_{g \sim U_{H_n}}[\rho_{U_n}(g, h) \leq 5\alpha/4] \leq (c_2 \alpha)^{n-1}$$

where $c_2 > 0$ is another absolute constant.

Suppose we choose arbitrarily $g_1, g_2, ... \in H_n$ that are at a distance at most α from h, but $\alpha/2$ far from one another. By the triangle inequality, $\alpha/4$-balls centered at $g_1, g_2, ...$ are disjoint. Thus, the probability that a random element of H_n is in a ball of radius $\alpha/4$ centered at one of $g_1, ..., g_N$ is at least $N(c_1\alpha)^{n-1}$. On the other hand, since each $g_1, ..., g_N$ has distance at most α from h, any element of an $\alpha/4$ ball centered at one of them is at most $\alpha + \alpha/4$ far from h. Thus, the union of the $\alpha/4$ balls centered at $g_1, ..., g_N$ is contained in the $5\alpha/4$ ball centered at h. Thus $N(c_1\alpha)^{n-1} \leq (c_2\alpha)^{n-1}$, which implies $N \leq (c_2/c_1)^{n-1} = 2^{O(n)}$, completing the proof. □

5. A bound for consistent hypothesis finders

In this section we analyze algorithms that work by finding hypotheses with zero training error. This is one way to achieve computational efficiency, as is the case when F consists of halfspaces.

The following lemma generalizes the Chernoff bound to hold uniformly over a class of random variables; it differs from standard bounds of this type in that it provides especially strong bounds on the probability that an empirical estimate is *much* larger than the true expectation. While the proof uses standard techniques, we have included it because we do not know of a published proof of exactly this statement.

Lemma 7. *Suppose F is a set of $\{0,1\}$-valued functions with a common domain X. Let d be the VC-dimension of F. Let D be a probability distribution over X. Choose $\alpha > 0$ and $K \geq 4$. Then if*

$$m \geq \frac{c\left(d \log \frac{1}{\alpha} + \log \frac{1}{\delta}\right)}{\alpha K \log K},$$

where c is an absolute constant, then

$$\mathbf{Pr}_{\mathbf{u} \sim D^m}[\exists f, g \in F,\ \mathbf{Pr}_D(f \neq g) \leq \alpha \text{ but } \hat{\mathbf{Pr}}_{\mathbf{u}}(f \neq g) > K\alpha] \leq \delta.$$

That is, with probability at least $1 - \delta$, every $f, g \in F$ such that $\mathbf{Pr}_D[f \neq g] \leq \alpha$ satisfies $\hat{\mathbf{Pr}}_{\mathbf{u}}[f \neq g] \leq K\alpha$.

Proof: See Appendix A. □

Now we are ready for the main analysis of this section.

Theorem 8. *Suppose the doubling dimension of (F, ρ_D) is at most d and the VC-dimension of F is at most d. Any consistent hypothesis finder for F PAC learns F with respect to D with accuracy $1 - \epsilon$ and confidence $1 - \delta$ from*

$$m = O\left(\frac{1}{\epsilon}\left(d\sqrt{\log \frac{1}{\epsilon}} + \log \frac{1}{\delta}\right)\right)$$

examples.

Proof: Let

$$\alpha = \epsilon \exp\left(-\sqrt{\ln \frac{1}{\epsilon}}\right).$$

We can assume without loss of generality that ϵ is sufficiently small that $\alpha \leq \epsilon/16$.

Let $f \in F$ be an arbitrary target function. As is often the case (see [33]), we will find it useful to consider a collection of random variables that indicate whether the hypotheses in F make errors or not, because these are the random variables whose probabilities the learning algorithm needs to estimate. For each $h \in F$, define $\ell_h : X \rightarrow \{0, 1\}$ by $\ell_h(x) = 1 \Leftrightarrow h(x) \neq f(x)$. Let $\ell_F = \{\ell_h : h \in F\}$. Notice that $\ell_h = h \oplus f$ where $\oplus$ is the exclusive or. Therefore, $\ell_F = F \oplus f = \{h \oplus f \mid h \in F\}$.

Since $\ell_g(x) \neq \ell_h(x)$ exactly when $g(x) \neq h(x)$, the doubling dimension of ℓ_F is the same as the doubling dimension of F, that is $d(\ell_F, D) = d(F, D)$; the VC-dimension of ℓ_F is also known to be the same as the VC-dimension of F (see [33]).

Let G be an α-packing in ℓ_F that is also an α-cover, as in Lemma 1.

We will show that with probability at least $1 - \delta$ any hypothesis $g \in F$ that is consistent with a training set $\mathbf{u}$ of size m has error at most ϵ. This is equivalent to

$$\mathbf{Pr}_{\mathbf{u} \sim D^m}[(\exists g \in \ell_F)\ \mathbf{E}_D[g] > \epsilon \text{ and } \hat{\mathbf{E}}_{\mathbf{u}}(g) = 0] \leq \delta.$$

We want to bound the probability of this event, which concerns all candidate hypotheses, in terms of an event that is determined only by the effect of the random sample on the elements in the cover G.

The first step is to argue if some classifier has a large error rate with respect to the underlying distribution, then so must its nearest neighbor in the cover. For each $g \in \ell_F$, let $\phi(g)$ be its nearest neighbor in G. Since $\alpha \leq \epsilon/8$, by the triangle inequality,

$$\mathbf{E}_D(g) > \epsilon \text{ and } \hat{\mathbf{E}}_{\mathbf{u}}(g) = 0 \;\Rightarrow\; \mathbf{E}_D(\phi(g)) > 7\epsilon/8 \text{ and } \hat{\mathbf{E}}_{\mathbf{u}}(g) = 0. \qquad (5)$$

This statement still includes the condition $\hat{\mathbf{E}}_{\mathbf{u}}(g) = 0$ which concerns a classifier that is not in the cover. We can remedy this by observing that either $\phi(g)$ had small error rate on the training data, or $\phi(g)$'s training error was much larger than g's, and therefore $\phi(g)$ often disagreed with g on the training data. That is,

$$\hat{\mathbf{E}}_{\mathbf{u}}(g) = 0 \;\Rightarrow\; (\hat{\mathbf{E}}_{\mathbf{u}}(\phi(g)) \leq \epsilon/4 \text{ or } \hat{\mathbf{Pr}}_{\mathbf{u}}(\phi(g) \neq g) > \epsilon/4).$$

Combining this with (5), we have

$$\begin{aligned} &\mathbf{Pr}_{\mathbf{u}\in D^m}[\exists g \in \ell_F, \mathbf{E}_D(g) > \epsilon \text{ but } \hat{\mathbf{E}}_{\mathbf{u}}(g) = 0] \\ &\leq \mathbf{Pr}_{\mathbf{u}\in D^m}[\exists g \in \ell_F, \mathbf{E}_D(\phi(g)) > 7\epsilon/8 \text{ but } \hat{\mathbf{E}}_{\mathbf{u}}(\phi(g)) \leq \epsilon/4] \\ &\qquad\qquad + \mathbf{Pr}_{\mathbf{u}\in D^m}[\exists g \in \ell_F, \hat{\mathbf{Pr}}_{\mathbf{u}}[\phi(g) \neq g] > \epsilon/4]. \qquad (6) \end{aligned}$$

Now, we will bound the two terms in (6) one at a time. Let us begin with the first part. We have

$$\begin{aligned} &\mathbf{Pr}_{\mathbf{u}\in D^m}[\exists g \in \ell_F, \mathbf{E}_D(\phi(g)) > 7\epsilon/8 \text{ but } \hat{\mathbf{E}}_{\mathbf{u}}(\phi(g)) \leq \epsilon/4] \\ &= \mathbf{Pr}_{\mathbf{u}\in D^m}[\exists g \in G, \;\; \mathbf{E}_D(g) > 7\epsilon/8 \text{ but } \hat{\mathbf{E}}_{\mathbf{u}}(g) \leq \epsilon/4] \\ &\leq \sum_{k=0}^{\lfloor \log(8/(7\epsilon)) \rfloor} \mathbf{Pr}[\exists g \in G, \;\; 2^k(7\epsilon/8) < \rho_D(g, \ell_f) \leq 2^{k+1}(7\epsilon/8) \\ &\qquad\qquad \text{and } \hat{\mathbf{Pr}}_{\mathbf{u}}[f \neq g] \leq \epsilon/4] \\ &\leq \sum_{k=0}^{\infty} \left(\frac{7\epsilon 2^{k+2}}{\alpha} \right)^d e^{-(7/64)2^k \epsilon m}, \end{aligned}$$

by Lemma 2 and Lemma 4.

Computing a geometric sum exactly as in the proof of Theorem 5, we have that $m = O(d/\epsilon)$ suffices for

$$\mathbf{Pr}_{\mathbf{u} \in D^m}[\exists g \in \ell_F, \mathbf{E}_D(\phi(g)) > 7\epsilon/8 \text{ but } \hat{\mathbf{E}}_{\mathbf{u}}(\phi(g)) \leq \epsilon/4] \leq \left(\frac{c_1 \epsilon}{\alpha}\right)^d e^{-c_2 \epsilon m},$$

for absolute constants $c_1, c_2 > 0$.

By plugging in the value of α and solving, we can see that

$$m = O\left(\frac{1}{\epsilon}\left(d\sqrt{\log\frac{1}{\epsilon}} + \log\frac{1}{\delta}\right)\right)$$

suffices for

$$\mathbf{Pr}_{\mathbf{u} \in D^m}[\exists g \in \ell_F, \mathbf{E}_D(\phi(g)) > 7\epsilon/8 \text{ but } \hat{\mathbf{E}}_{\mathbf{u}}(\phi(g)) \leq \epsilon/4] \leq \delta/2. \tag{7}$$

Now, we turn to bounding the second term of (6). That is, we want to show that it is unlikely that the training error of $\phi(g)$ is much worse than that of g. Recall that G is an α-cover of F. Since $\mathbf{Pr}_D[\phi(g) \neq g] \leq \alpha \leq \epsilon/8$ for all $g \in \ell_F$, applying Lemma 7 with $K = \epsilon/(4\alpha)$ (recall that $\alpha \leq \epsilon/16$, so that $K \geq 4$), we get that there is an absolute constant $c > 0$ such that

$$m \geq \frac{c\left(d \log\frac{1}{\alpha} + \log\frac{1}{\delta}\right)}{\left(\frac{\epsilon}{4} - \alpha\right)\log(\frac{\epsilon}{4\alpha})} \tag{8}$$

also suffices for

$$\mathbf{Pr}_{\mathbf{u} \in D^m}[\exists g \in \ell_F, \hat{\mathbf{Pr}}_{\mathbf{u}}(\phi(g) \neq g) > \epsilon/4] \leq \delta/2.$$

Substituting the value α into (8), it is sufficient that

$$m \geq \frac{c\left(d\left(\log\frac{1}{\epsilon} + \sqrt{\log\frac{1}{\epsilon}}\right) + \log\frac{1}{\delta}\right)}{\frac{\epsilon}{8}\left(\sqrt{\log\frac{1}{\epsilon}} - \log 4\right)} = O\left(\frac{1}{\epsilon}\left(d\sqrt{\log\frac{1}{\epsilon}} + \log\frac{1}{\delta}\right)\right).$$

Putting this together with (7) and (6) completes the proof. □

6. The Doubling Dimension versus the VC-dimension

In this section we give bounds on $d_\epsilon(F, D)$ for different classes of F. We provide two lower bounds that imply that there is no general bound on the doubling dimension in terms of the VC-dimension.

Haussler in [14] gave a randomized construction of a class of classifiers F of VC-dimension d and a distribution D such that

$$|F| \geq \left(\frac{1}{2e\alpha}\right)^d$$

and where for every two classifiers $f, g \in F$, $\mathbf{Pr}_D[f \neq g] \geq \alpha$. This construction doesn't seem to give a bound on the doubling dimension.

Our first construction is deterministic. We construct a class of classifiers F of VC-dimension d and a distribution D such that

$$|F| \geq \left(\frac{1}{\alpha}\right)^d$$

where for every two classifiers $f, g \in F$, $2\alpha > \mathbf{Pr}_D[f \neq g] \geq \alpha$. This improves Haussler's bound and implies

$$d_{2\epsilon}(F, D) \geq \mathrm{VC}(F) \log \frac{1}{\epsilon}.$$

Then we build another deterministic construction that holds for the class H of halfspaces under a non-uniform distribution D. The class H will be a subset of halfspaces over a space of dimension $2d$ and has VC-dimension d. We show

$$d_{2\epsilon}(H, D) \geq \frac{\mathrm{VC}(H)}{2} \log \frac{1}{\epsilon}.$$

Both of our constructions make use of finite fields [23]. We will use a few facts about finite fields.

Lemma 9 (see [23]). *For any prime power $q = p^m$ and any positive integer k:*

- *There is a finite field of size q.*
- *Any two finite fields of size q are isomorphic (that is, there is essentially only one field of size of q, called $\mathrm{GF}(q)$).*
- *Tuples of n members of $\mathrm{GF}(q)$ form a vector space $\mathrm{GF}(q)^n$ over $\mathrm{GF}(q)$ of dimension n.*
- *For any linearly independent*

  $$\mathbf{x}_1, ..., \mathbf{x}_k \in \mathrm{GF}(q)^n,$$

  the subspace of $\mathrm{GF}(q)^n$ spanned by $\mathbf{x}_1, ..., \mathbf{x}_k$ has size q^k.

6.1. A relatively tight lower bound using finite fields

Now we're ready for the lower bound.

Theorem 10. *For any prime power q and positive integer d and $\alpha = 1/q$ there is a set F of classifiers and a probability distribution D over their common domain with the following properties:*

- *the VC-dimension of F is at most d*
- *for each $f, g \in F$, $\alpha < \rho_D(f, g) \leq 2\alpha$.*
- $|F| \geq (1 - \alpha) \left(\frac{1}{\alpha}\right)^d$
- *the doubling dimension $d(F, D)$ (and $d_{2\alpha}(F, D)$) is at least $d \log \frac{1}{\alpha}$.*

Proof: Let $X = \mathrm{GF}(q)^{d+1}$, and let F consist of indicator functions for all subspaces of X of dimension d. In other words, F is the set of indicator functions for $\{\mathbf{x} : \mathbf{x} \cdot \mathbf{a} = 0\}$, for all nonzero $\mathbf{a} \in X$. Let D be the uniform distribution over X.

First, let us prove that F has VC-dimension at most d. Choose distinct $\mathbf{x}_1, ..., \mathbf{x}_{d+1} \in X$. If they are linearly independent, then, by definition, they do not lie in a common proper subspace of X, and therefore they cannot all be labeled 1 by a function in F. If they are linearly dependent, then one of them lies in the subspace spanned by the others; say $\mathbf{x}_{d+1}$ lies in subspace spanned by $\mathbf{x}_1, ..., \mathbf{x}_d$. This means that any $f \in F$ for which $f(\mathbf{x}_1) = ... = f(\mathbf{x}_d) = 1$ also has $f(\mathbf{x}_{d+1}) = 1$.

Next, Lemma 9 implies that for any $f \in F$

$$\mathbf{Pr}_{\mathbf{x} \sim D}[f(\mathbf{x}) = 1] = \frac{q^d}{q^{d+1}} = 1/q = \alpha.$$

This immediately implies that for any $f, g \in F$,

$$\begin{aligned} \rho_D(f,g) &= \mathbf{Pr}_{\mathbf{x} \sim D}[f(\mathbf{x}) = 1 \text{ and } g(\mathbf{x}) = 0] \\ &\qquad + \mathbf{Pr}_{\mathbf{x} \sim D}[f(\mathbf{x}) = 0 \text{ and } g(\mathbf{x}) = 1] \\ &\leq 2\alpha. \end{aligned}$$

Now for the lower bound on ρ_D; suppose f and g are distinct members of F and the S and T be the subspaces corresponding to f and g respectively. Since S and T are distinct, the subspace $S \cap T$ must have dimension less than d. Thus Lemma 9 implies

$$\begin{aligned} \rho_D(f,g) &= \mathbf{Pr}_{\mathbf{x} \sim D}[f(\mathbf{x}) = 1] + \mathbf{Pr}_{\mathbf{x} \sim D}[g(\mathbf{x}) = 1] \\ &\qquad - 2\mathbf{Pr}_{\mathbf{x} \sim D}[f(\mathbf{x}) = 1 \text{ and } g(\mathbf{x}) = 1] \end{aligned}$$

$$
\begin{aligned}
&= 2/q - 2 \times \frac{q^{\dim(S \cap T)}}{q^{d+1}} \\
&\geq 2/q - 2/q^2 \\
&> 1/q = \alpha,
\end{aligned}
$$

completing the proof of the second bullet point.

Finally, let us lower bound the size of F. For any nonzero $\mathbf{a}$, F has an indicator function for $S = \{\mathbf{x} : \mathbf{x} \cdot \mathbf{a} = 0\}$. The set of all $\mathbf{a}$ which yield S this way consists exactly of nonzero elements of the orthogonal complement of S. Since S is d dimensional, its orthogonal complement is one-dimensional, and, by Lemma 9, it has q elements. Thus,

$$|F| = \frac{q^{d+1} - 1}{q - 1} \geq \left(\frac{1}{\alpha} \right)^d .$$

Therefore the doubling dimension and the 2α-doubling dimension is at least

$$\log |F| = d \log \frac{1}{\alpha}.$$

This completes the proof. □

Theorem 10 implies that there is no bound on the doubling dimension of (G, ρ_D) in terms of the VC-dimension of G. For any constraint on the VC-dimension, a set G satisfying the constraint can have arbitrarily large doubling dimension by setting the value of α in Theorem 10 arbitrarily small.

Theorem 10 matches an upper bound of Haussler [14] on $\mathcal{M}(\alpha; (\rho_D, F))$ in terms of the VC-dimension of F up to a constant, despite the fact that Haussler's upper bound did not require that $\rho_D(f, g) \leq 2\alpha$ for all $f, g \in F$.

6.2. A lower bound using halfspaces

Theorem 10 still leaves open the possibility of a bound on the doubling dimension of halfspaces in $\mathbf{R}^n$ that holds independent of D. We show in this section that this is not possible.

Theorem 11. *For any prime p and positive integer n and $\alpha = 1/p$ there is a probability distribution D over $\mathbf{R}^n$ and a set F_n of halfspaces in $\mathbf{R}^n$ with the following properties:*

- *for each pair $f, g \in F_n$, $\alpha < \rho_D(f, g) \leq 2\alpha$.*
- $|F_n| \geq \left(\frac{1}{\alpha}\right)^{\lceil n/4 \rceil}$.

Proof: Our lower bound for halfspaces will proceed by proving a lower bound for another concept class G, and then embedding G into halfspaces. (A lower bound for a different learning model was proved by embedding a class containing G into halfspaces in [17, Corollary 45].)

Let $p > d$ be two integers such that p is a prime number and let $\alpha = 1/p$. The domain X will consist of d copies of GF(p); formally $X = \{1, ..., d\} \times$ GF(p). The elements of $\{i\} \times$ GF(p) will be called the *ith block* of X. Each classifier in G will be the indicator function for a set of d elements of X, one element from each block. The elements will be chosen by evaluating polynomials over GF(p); let Z be the set of all such polynomials of degree at most $\lceil d/2 \rceil - 1$. For each polynomial $\phi \in Z$, let f_ϕ be the indicator function for $\{(i, \phi(i)) : i \in \{1, ..., d\}\}$. Then let

$$G = \{f_\phi : \phi \in Z\}.$$

Since each classifier f_ϕ has $|f_\phi^{-1}(1)| \leq d$, the VC-dimension of G is at most d and

$$\mathbf{Pr}[f_\phi = 1] = \frac{d}{pd} = \alpha.$$

Each $a_0, ..., a_{\lceil d/2 \rceil - 1} \in \mathrm{GF}(p)$ leads to a distinct polynomial $a_0 + a_1 x + ... + a_{\lceil d/2 \rceil - 1} x^{\lceil d/2 \rceil - 1}$, so

$$|G| = |Z| = p^{\lceil d/2 \rceil} = \left(\frac{1}{\alpha} \right)^{\lceil d/2 \rceil}.$$

Now for two classifiers f_{ϕ_1} and f_{ϕ_2} we have

$$\mathbf{Pr}[f_{\phi_1} \neq f_{\phi_2}] \leq \frac{2d}{pd} = 2\alpha$$

and

$$\begin{aligned} \mathbf{Pr}[f_{\phi_1} \neq f_{\phi_2}] &= \frac{2 \sum_{i=1}^{d} \mathbf{I}[\phi_1(i) \neq \phi_2(i)]}{pd} \\ &= \frac{2 \sum_{i=1}^{d} (1 - \mathbf{I}[\phi_1(i) = \phi_2(i)])}{pd} \\ &= \frac{2d - 2|\{1 \leq i \leq d \mid \phi_1(i) = \phi_2(i)\}|}{pd}. \end{aligned}$$

The number of zeroes of a polynomial is bounded by its degree, so

$$\mathbf{Pr}[f_{\phi_1} \neq f_{\phi_2}] \geq \frac{2d - 2 \cdot deg(\phi_1 - \phi_2)}{pd} > \alpha.$$

This shows that

$$d_{2\alpha}(G, D) \geq \frac{d}{2} \log \frac{1}{\alpha}.$$

We now embed the class into halfspaces (a different embedding was employed in [17]). We will assign each element of $X = \{1, ..., d\} \times \mathrm{GF}(p)$ a vector in the $2d$-dimensional space. Then we show that each classifier f_ϕ has a halfspace representation over those vectors. (The distribution D is the image of the uniform distribution on X after this embedding.)

Consider the 2-vectors

$$u_j = \left(\cos \frac{2\pi j}{p}, \sin \frac{2\pi j}{p} \right), j = 0, 1, \ldots, p-1.$$

We assign to the instance $(i,j) \in X$ a $2d$-vector $v^{(i,j)}$. To do this, we will think of the indices $\{1,...,2d\}$ as being divided into d blocks of size 2, corresponding to the d blocks of X. The vector $v^{(i,j)}$ corresponding to $(i,j) \in X$ is equal to u_j in block i and $(0,0)$ in the other blocks.

Let $A = (A_{1,1}, A_{1,2}, A_{2,1}, A_{2,2}, \ldots, A_{d,1}, A_{d,2})$ be a vector of $2d$ coefficients. Notice that

$$\begin{aligned} A \cdot v^{(i,j)} &= (A_{1,1}, A_{1,2}, A_{2,1}, A_{2,2}, \ldots, A_{d,1}, A_{d,2}) \cdot v^{(i,j)} \\ &= A_{i,1} \cos \frac{2\pi j}{p} + A_{i,2} \sin \frac{2\pi j}{p} \\ &= (A_{i,1}, A_{i,2}) \cdot u_j \qquad (9) \end{aligned}$$

depends only the coefficients in block i.

For $A_1^{(j)} = \cos \frac{2\pi j}{p}$ and $A_2^{(j)} = \sin \frac{2\pi j}{p}$, we have

$$(A_1^{(j)}, A_2^{(j)}) \cdot u_j = 1 \text{ and } (A_1^{(j)}, A_2^{(j)}) \cdot u_k < 1 \text{ for all } k \neq j. \qquad (10)$$

Consider now a classifier f_ϕ. This classifier is 1 for the vectors $(i, \phi(i))$ and zero elsewhere. Consider the halfspace

$$A_1^{(\phi(1))} X_1 + A_2^{(\phi(1))} X_2 + A_1^{(\phi(2))} X_3 + A_2^{(\phi(2))} X_4 + \cdots + A_1^{(\phi(d))} X_{2d-1} + A_2^{(\phi(d))} X_{2d} \geq 1.$$

By (9) and (10) assigning $(i, \phi(i))$ in the halfspace we get

$$(A_1^{(\phi(i))}, A_2^{(\phi(i))}) \cdot u_{\phi(i)} \geq 1$$

and therefore $v^{(i,\phi(i))}$ is classified as 1 in the halfspace. Assigning $v^{(i,j)}$ for $j \neq \phi(i)$ in the halfspace we get

$$(A_1^{(\phi(i))}, A_2^{(\phi(i))}) \cdot u_j < 1$$

and therefore $v^{(i,j)}$, $j \neq \phi(i)$ is classified as 0 in the halfspace. This completes the proof. □

7. Conclusion

The doubling dimension is a clean and intuitive way to identify cases in which the local complexity of families of classifiers is limited. A number of natural questions remain regarding the relationship between the doubling dimension and learning.

One compelling problem is to extend the analysis of this paper to the case in which no classifier in F has zero error. This case is complicated by the fact that there is no single target whose neighborhood we should consider. This might be addressed using regularization.

It also is not clear whether a bound on the VC-dimension is necessary to obtain the sample complexity bound of Theorem 8, or whether that bound can be improved even given a bound on the VC-dimension.

Proving meaningful lower bounds in terms of the doubling dimension appears problematic – a pair $d(F, D)$ can be arbitrarily large even if all the classifiers in F are enclosed within an arbitrarily small ρ_D ball.

Theorem 10 shows that the doubling dimension can be much larger than the VC-dimension. It also can be much smaller, for example when F shatters a large set with zero probability under D.

In the context of learning, the doubling dimension bounds the maximum extent to which candidate classifiers can crowd around a target. It may be useful instead to consider bounds the average crowding, given a prior over target functions. This may provide a way around the results of Section 6, and allow a bound in terms of the VC-dimension. (A result like this would be analogous to the bound on the density of the one-inclusion graph [15].) Bounding the average crowding by the VC-dimension for a "least favorable prior" may provide a way to obtain improved general bounds on the sample complexity of PAC learning, making progress on an open problem posed by Ehrenfeuch, Haussler, Kearns and Valiant [11].

Acknowledgements

We thank Gábor Lugosi and Tong Zhang for their help, Sanjoy Dasgupta for asking about distribution-free bounds on the doubling dimension for half-spaces, and anonymous reviewers for their comments.

References

[1] K. Alexander. Rates of growth for weighted empirical processes. In *Proc. of Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, volume 2, pages 475–493, 1985.

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

[3] P. Assouad. Plongements lipschitziens dans. *R . Bull. Soc. Math. France*, 111(4):429–448, 1983.

[4] P. L. Bartlett, O. Bousquet, and S. Mendelson. Local Rademacher complexities. *Annals of Statistics*, 33(4):1497–1537, 2005.

[5] P. L. Bartlett and P. M. Long. Prediction, learning, uniform convergence, and scale-sensitive dimensions. *Journal of Computer and System Sciences*, 56(2):174–190, 1998.

[6] G. Benedek and A. Itai. Learnability with respect to fixed distributions. *Theoretical Computer Science*, 86(2):377–389, 1991.

[7] A. Beygelzimer, S. Kakade, and J. Langford. Cover trees for nearest neighbor. *ICML*, 2006.

[8] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *JACM*, 36(4):929–965, 1989.

[9] H. T. H. Chan, A. Gupta, B. M. Maggs, and S. Zhou. On hierarchical routing in doubling metrics. *SODA*, 2005.

[10] R. M. Dudley. Central limit theorems for empirical measures. *Annals of Probability*, 6(6):899–929, 1978.

[11] A. Ehrenfeucht, D. Haussler, M. Kearns, and L. G. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–251, 1989.

[12] A. Gupta, R. Krauthgamer, and J. R. Lee. Bounded geometries, fractals, and low-distortion embeddings. *FOCS*, 2003.

[13] S. Har-Peled and M. Mendel. Fast construction of nets in low dimensional metrics, and their applications. *SICOMP*, 35(5):1148–1184, 2006.

[14] D. Haussler. Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

[15] D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting $\{0,1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):129–161, 1994.

[16] J. Heinonen. *Lectures on analysis on metric spaces.* Springer-Verlag, 2001.

[17] D. P. Helmbold, N. Littlestone, and P. M. Long. Apple tasting. *Information and Computation*, 161(2):85–139, 2000. Preliminary version in FOCS'92.

[18] J. Kleinberg, A. Slivkins, and T. Wexler. Triangulation and embedding using small sets of beacons. *FOCS*, 2004.

[19] A. N. Kolmogorov and V. M. Tihomirov. ϵ-entropy and ϵ-capacity of sets in functional spaces. *American Mathematical Society Translations (Ser. 2)*, 17:277–364, 1961.

[20] R. Krauthgamer and J. R. Lee. The black-box complexity of nearest neighbor search. *ICALP*, 2004.

[21] R. Krauthgamer and J. R. Lee. Navigating nets: simple algorithms for proximity search. *SODA*, 2004.

[22] F. Kuhn, T. Moscibroda, and R. Wattenhofer. On the locality of bounded growth. *PODC*, 2005.

[23] R. Lidl and H. Niederreiter. *Finite fields.* Cambridge University Press, 1997.

[24] P. M. Long. On the sample complexity of PAC learning halfspaces against the uniform distribution. *IEEE Transactions on Neural Networks*, 6(6):1556–1559, 1995.

[25] P. M. Long. An upper bound on the sample complexity of PAC learning halfspaces with respect to the uniform distribution. *Information Processing Letters*, 87(5):229–234, 2003.

[26] S. Mendelson. Estimating the performance of kernel classes. *Journal of Machine Learning Research*, 4:759–771, 2003.

[27] R. Motwani and P. Raghavan. *Randomized Algorithms.* Cambridge University Press, 1995.

[28] J. Shawe-Taylor, M. Anthony, and N. Biggs. Bounding the sample size with the Vapnik-Chervonenkis dimension. *Discrete Applied Mathematics*, 42:65–73, 1993.

[29] A. Slivkins. Distance estimation and object location via rings of neighbors. *PODC*, 2005.

[30] K. Talwar. Bypassing the embedding: Approximation schemes and compact representations for low dimensional metrics. *STOC*, 2004.

[31] S. van de Geer. *Empirical processes in M-estimation.* Cambridge Series in Statistical and Probabilistic Methods, 2000.

[32] A. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes With Applications to Statistics*. Springer, 1996.

[33] V. N. Vapnik. *Estimation of Dependencies based on Empirical Data.* Springer Verlag, 1982.

[34] V. N. Vapnik. *Statistical Learning Theory.* New York, 1998.

[35] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

[36] T. Zhang. Information theoretical upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, 2006. to appear.

A. Proof of Lemma 7

We want to bound the probability that any error rate on the sample is much worse than the corresponding error rate with respect to the underlying distribution. As usual [35, 34, 2, 31], we begin with a symmetrization step, in which the underlying distribution is replaced with another sample. As in [28], we will find it useful for this "ghost sample" to be much bigger than the sample given to the learning algorithm.

Claim 12. *There is a constant c_0 such that, for $m \geq (c_0/\alpha)$, and for any positive integer k,*

$$\begin{aligned}
&\mathbf{Pr}_{\mathbf{u}\sim D^m}(\exists f, g \in F,\ \mathbf{Pr}_D(f \neq g) \leq \alpha \text{ but } \hat{\mathbf{Pr}}_{\mathbf{u}}(f \neq g) > K\alpha) \\
&\leq 2\mathbf{Pr}_{\mathbf{u}\sim D^{(k+1)m}}(\exists f, g \in F,\ \frac{1}{km}\sum_{i=1}^{km} 1_{f\neq g}(u_i) \leq 2\alpha \\
&\qquad\qquad \text{but } \frac{1}{m}\sum_{i=1}^{m} 1_{f\neq g}(u_{km+i}) > K\alpha)).
\end{aligned}$$

Proof: This proof closely follows the usual outline (see [8]). Let

$$\begin{aligned} J &= \{\mathbf{u} \in X^{(k+1)m} : \exists f, g \in F,\ \frac{1}{km}\sum_{i=1}^{km} 1_{f \neq g}(u_i) \leq 2\alpha \\ &\qquad\qquad\qquad\qquad \text{but } \frac{1}{m}\sum_{i=1}^{m} 1_{f \neq g}(u_{km+i}) > K\alpha\} \\ Q &= \left\{\mathbf{u} \in X^m : \exists f, g \in F,\ \mathbf{Pr}_D(f \neq g) \leq \alpha \text{ but } \hat{\mathbf{Pr}}_{\mathbf{u}}(f \neq g) > K\alpha\right\}. \end{aligned}$$

By Fubini's Theorem,

$$\begin{aligned} \mathbf{Pr}_{D^{(k+1)m}}(J) &= \mathbf{E}_{\mathbf{u} \sim D^m}(\mathbf{Pr}_{\mathbf{v} \sim D^{km}}((v_1, ..., v_{km}, u_1, ..., u_m) \in J)) \\ &\geq \mathbf{E}_{\mathbf{u} \sim D^m}(\mathbf{Pr}_{\mathbf{v} \sim D^{km}}((v_1, ..., v_{km}, u_1, ..., u_m) \in J) | \mathbf{u} \in Q) \\ &\qquad \times \mathbf{Pr}_{D^m}(Q) \qquad (11) \end{aligned}$$

Suppose $\mathbf{u} \in Q$, and let $f_0, g_0 \in F$ witness this membership. Then $\mathbf{Pr}_D(f_0 \neq g_0) \leq \alpha$, and Lemma 4 implies that there is a constant independent of α such that $m \geq (c_0/\alpha)$ is sufficient for the disagreement rate between f_0 and g_0 on the "ghost sample" $\mathbf{v}$ to be at least 2α with probability at most 1/2. Therefore,

$$\mathbf{Pr}_{\mathbf{v} \sim D^{km}}((v_1, ..., v_{km}, u_1, ..., u_m) \in J) \leq 1/2.$$

Since this is true for any $\mathbf{u} \in Q$, (11) implies that

$$\mathbf{Pr}_{D^{(k+1)m}}(J) \geq \mathbf{E}_{(u_1, ..., u_{km}) \sim D^{km}}(1/2)\mathbf{Pr}_{D^m}(Q) = \mathbf{Pr}_{D^m}(Q)/2,$$

completing the proof of this claim. □

Continuing with the proof of Lemma 7, let $k = \lceil 1/\alpha \rceil$. Suppose Γ is the set of permutations π on $\{1, ..., (k+1)m\}$ such that $\pi(\{i, m+i, ..., km+i\}) = \{i, m+i, ..., km+i\}$ for all $i \in \{1, ..., m\}$. That is π separately permutes

$\{1, m+1, ..., km+1\}$, $\{2, m+2, ..., km+2\}$, and so on. Let U be the uniform distribution over Γ. Then, because product distributions are unaffected by permutations,

$$\begin{aligned}
&\mathbf{Pr}_{\mathbf{u} \sim D^{(k+1)m}}(\exists f, g \in F,\ \frac{1}{km}\sum_{i=1}^{km} 1_{f \neq g}(u_i) \leq 2\alpha \\
&\qquad\qquad\qquad\qquad \text{but } \frac{1}{m}\sum_{i=1}^{m} 1_{f \neq g}(u_{km+i}) > K\alpha) \\
&= \mathbf{Pr}_{\mathbf{u} \sim D^{(k+1)m}, \pi \sim U}(\exists f, g \in F,\ \frac{1}{km}\sum_{i=1}^{km} 1_{f \neq g}(u_{\pi(i)}) \leq 2\alpha \\
&\qquad\qquad\qquad\qquad \text{but } \frac{1}{m}\sum_{i=1}^{m} 1_{f \neq g}(u_{\pi(km+i)}) > K\alpha) \\
&\leq \max_{\mathbf{u} \in X^{(k+1)m}} \mathbf{Pr}_{\pi \sim U}(\exists f, g \in F,\ \frac{1}{km}\sum_{i=1}^{km} 1_{f \neq g}(u_{\pi(i)}) \leq 2\alpha \\
&\qquad\qquad\qquad\qquad \text{but } \frac{1}{m}\sum_{i=1}^{m} 1_{f \neq g}(u_{\pi(km+i)}) > K\alpha).
\end{aligned}$$

For the time being, fix $f, g \in F$.

Choose $\mathbf{u} \in X^{(k+1)m}$ and $\pi \in \Gamma$ such that $\frac{1}{km}\sum_{i=1}^{km} 1_{f \neq g}(u_{\pi(i)}) \leq 2\alpha$. Then

$$\begin{aligned}
&\frac{1}{(k+1)m} \sum_{i=1}^{(k+1)m} 1_{f \neq g}(u_{\pi(i)}) \\
&= \frac{1}{(k+1)m}\left(\left(\sum_{i=1}^{km} 1_{f \neq g}(u_{\pi(i)})\right) + \sum_{i=1}^{m} 1_{f \neq g}(u_{\pi(km+i)})\right) \\
&\leq \frac{1}{(k+1)m}(2\alpha km + m) \\
&\leq 3\alpha, \qquad (12)
\end{aligned}$$

since $k \geq 1/\alpha$. Now, suppose π is chosen uniformly at random according to U. Then

$$\begin{aligned}
&\mathbf{E}\left(\frac{1}{m}\sum_{i=1}^{m} 1_{f\neq g}(u_{\pi(km+i)})\right) \\
&= \frac{1}{m}\sum_{i=1}^{m} \mathbf{E}(1_{f\neq g}(u_{\pi(km+i)})) \\
&= \frac{1}{m}\sum_{i=1}^{m} \frac{1}{k+1}\sum_{j=1}^{k+1} 1_{f\neq g}(u_{(j-1)m+i}) \\
&\leq 3\alpha,
\end{aligned}$$

by (12). Applying Lemma 4, this implies that, for our fixed f and g,

$$\begin{aligned}
&\mathbf{Pr}_{\pi\sim U}\left(\frac{1}{km}\sum_{i=1}^{km} 1_{f\neq g}(u_{\pi(i)}) \leq 2\alpha \text{ but } \frac{1}{m}\sum_{i=1}^{m} 1_{f\neq g}(u_{\pi(km+i)}) > K\alpha)\right) \\
&\leq \left(\frac{\exp\left(\frac{K}{3}-1\right)}{\left(\frac{K}{3}\right)^{\frac{K}{3}}}\right)^{3\alpha m} \\
&\leq e^{-c_1 K \log K \alpha m},
\end{aligned}$$

for an absolute constant c_1, for all $K \geq 4$.

We have

$$\begin{aligned}
&|\{(z_1, ..., z_{m+k}) \;:\; \exists f, g \in F,\; \forall i,\; z_i = 1 \Leftrightarrow f(u_i) \neq g(u_i)\}| \\
&\leq |\{(f(z_1), ..., f(z_{m+k}), g(z_1), ..., g(z_{m+k})) \;:\; f, g \in F\}| \\
&\leq ((e(k+1)m/d)^d)^2,
\end{aligned}$$

by the Sauer-Shelah lemma. This means that

$$\max_{\mathbf{u}\in X^{(k+1)m}} \mathbf{Pr}_{\pi\sim U}(\exists f, g \in F,\; \frac{1}{km}\sum_{i=1}^{km} 1_{f\neq g}(u_{\pi(i)}) \leq 2\alpha$$

$$\text{but } \frac{1}{m}\sum_{i=1}^{m} 1_{f\neq g}(u_{\pi(km+i)}) > K\alpha))$$

$$\leq \left(\frac{e(k+1)m}{d}\right)^{2d} e^{-c_1(K\log K)\alpha m}.$$

From here, the usual manipulations (see Lemma 18 of [5]) complete the proof.