Improved Bounds on the Sample Complexity of Learning*

Yi Li
Department of Computer Science
National University of Singapore
Singapore 117543, Republic of Singapore
liyi@comp.nus.edu.sg

Philip M. Long
Department of Computer Science
National University of Singapore
Singapore 117543, Republic of Singapore
plong@comp.nus.edu.sg

Aravind Srinivasan†
Bell Laboratories
Lucent Technologies
600-700 Mountain Avenue
Murray Hill, NJ 07974-0636, USA
srin@research.bell-labs.com

*A preliminary version of this work appeared in the *Proceedings of the Eleventh Annual ACM-SIAM Symposium on Discrete Algorithms*, 2000.

†Part of this work was done while this author was at the School of Computing of the National University of Singapore.

Proposed running head: Sample complexity.

Address for proofs: Phil Long, Computer Science Department, National University of Singapore, Singapore 117543, Republic of Singapore.

Abstract

We present a new general upper bound on the number of examples required to estimate all of the expectations of a set of random variables uniformly well. The quality of the estimates is measured using a variant of the relative error proposed by Haussler and Pollard. We also show that our bound is within a constant factor of the best possible. Our upper bound implies improved bounds on the sample complexity of learning according to Haussler's decision theoretic model.

Keywords: Sample complexity, machine learning, empirical process theory, PAC learning, agnostic learning.

For a list of typographical symbols used, please see *LaTex*, by Leslie Lamport.

1 Introduction

Haussler [3], building on the work of Valiant [13], Vapnik [14] and others, introduced an abstract model of learning that unified the treatment of a variety of problems. In Haussler's model, "examples" are drawn independently at random according to some probability distribution and given to the learning algorithm, whose goal is to output a "hypothesis" that performs nearly as well as the best hypothesis in some "comparison class". The number of examples which is sufficient to ensure that with high probability a relatively accurate hypothesis can be determined has become known as the *sample complexity* in this context.

Haussler reduced the study of sample complexity to a more basic problem. For a real-valued function f and a probability distribution P over the domain of f, a natural estimate of the expectation of $f(x)$ when x is drawn according to P can be obtained as follows: obtain several samples $x_1, ..., x_m$ independently from P, and use $\frac{1}{m}\sum_{i=1}^{m} f(x_i)$, the sample average, as the estimated expectation. Chernoff-Hoeffding bounds can generally be used to show that accurate estimates are likely to be obtained here if m is large enough. To get good sample complexity bounds in Haussler's model, we need a generalization of this setting: for a domain X, a probability distribution P over X, and a possibly infinite set $\mathcal{F}$ of functions defined on X, one wants to use *one* collection of independent draws from P to simultaneously estimate the expectations of *all* the functions in $\mathcal{F}$ (w.r.t. P).

Let $\nu > 0$ be an adjustable parameter. Haussler proposed using the following measure of distance between two non-negative reals r and s, to determine how far the estimates are from the true expectations:

$$d_\nu(r, s) = \frac{|r - s|}{r + s + \nu}.$$

This can be thought of as a modification of the usual notion of "relative error" to make it well-behaved around 0 (i.e., when both r and s are non-negative reals that are close to 0) and symmetric in its arguments r and s. It can be verified that d_ν is a metric on the set of non-negative reals $\mathcal{R}^+$, and has some good metric properties such as being compatible with the ordering on the reals (if $0 \leq r < s < t$, then $d_\nu(r, s) < d_\nu(r, t)$ and $d_\nu(s, t) < d_\nu(r, t)$) [3]. Also, as seen below, upper bounds on this metric yield upper bounds for other familiar distance metrics.

The *pseudo-dimension* [10] (defined in Section 2) of a class $\mathcal{F}$ of $[0, 1]$-valued functions is a generalization of the Vapnik-Chervonenkis dimension [15], and is a measure of the "richness" of $\mathcal{F}$. Haussler [3] and Pollard [11] showed that, for any class $\mathcal{F}$ whose pseudo-dimension is d, if we

sample

$$O\left(\frac{1}{\alpha^2\nu}\left(d\log\frac{1}{\alpha}+d\log\frac{1}{\nu}+\log\frac{1}{\delta}\right)\right) \tag{1}$$

times, then with probability $1-\delta$, the d_ν distance between the sample average and the true expectation will be at most α, for *all* the functions in $\mathcal{F}$.

In this paper, we prove a bound of

$$O\left(\frac{1}{\alpha^2\nu}\left(d\log\frac{1}{\nu}+\log\frac{1}{\delta}\right)\right)$$

examples, which improves on (1) by a logarithmic factor when α is relatively small. Furthermore, we show that our bound is optimal to within a constant factor.

A line of research culminating in the work of Talagrand [12] studied the analogous problem in which the absolute value of the difference between the sample average and the true expectation was used instead of the d_ν metric: $O(\frac{1}{\alpha^2}(d+\log(1/\delta)))$ examples have been shown to suffice here. A disadvantage of this type of analysis is that, informally, the bottleneck occurs with random variables whose expectation is close to $1/2$. In a learning context, these correspond to hypotheses whose error is close to that obtained through random guessing. If good hypotheses are available, then accurate estimates of the quality of poor hypotheses are unnecessary. The d_ν metric enables one to take advantage of this observation to prove stronger bounds for learning when good hypotheses are available, which is often the case in practice. (See [3] for a more detailed discussion of the advantages of the d_ν metric.) In any case, our upper bound yields a bound within a constant factor of Talagrand's by setting $\alpha=\epsilon$, $\nu=1/2$, and the upper bound for PAC learning by setting $\nu=\epsilon$, $\alpha=1/2$.

Our upper bound proof makes use of *chaining*, a proof technique due to Kolmogorov, which was first applied to empirical process theory by Dudley.[1] Our analysis is the first application we know of chaining to bound the sample complexity of obtaining small relative error. The proof of our lower bound generalizes an argument of [5] to the case in which estimates are potentially nonzero.

2 Preliminaries

Fix a countably infinite domain X. (We assume X is countable for convenience, but weaker assumptions suffice: see [3].) The pseudo-dimension of a set $\mathcal{F}$ of functions from X to $[0,1]$,

[1] See [10] for a discussion of the history of chaining, and [8] for a simple proof of a bound within a constant factor of Talagrand's using chaining.

denoted by $\mathrm{Pdim}(\mathcal{F})$, is the largest d such that there is a sequence $x_1, ..., x_d$ of domain elements from X and a sequence $r_1, ..., r_d$ of real thresholds such that for each $b_1, ..., b_d \in \{\text{above}, \text{below}\}$, there is an $f \in \mathcal{F}$ such that for all $i = 1, ..., d$, we have $f(x_i) \geq r_i \Leftrightarrow b_i = \text{above}$. For $k \in \mathbf{N}$, the pseudo-dimension of a subset F of $[0,1]^k$ is defined using the above by viewing the elements of F as functions from $\{1, ..., k\}$ to $[0,1]$. The VC-dimension is the restriction of the pseudo-dimension to sets of functions from X to $\{0,1\}$.

We will make use of the usual Hoeffding bound; let $\exp(x)$ denote e^x.

Lemma 1 ([6]) *Let $Y_1, ..., Y_m$ be independent random variables for which each Y_i takes values in $[a_i, b_i]$. Then for any $\eta > 0$, we have*

$\mathbf{Pr}\left(\left|\left(\sum_{i=1}^m Y_i\right) - \left(\sum_{i=1}^m \mathbf{E}(Y_i)\right)\right| > \eta\right) \leq 2\exp\left(\frac{-2\eta^2}{\sum_{i=1}^m (b_i - a_i)^2}\right)$.

The following lower bound is a slight modification of Theorem 5 on page 12 of [1], and is proved similarly.

Lemma 2 *Suppose that $Y_1, \ldots, Y_m$ is a sequence of independent random variables taking only the values 0 and 1, and that for all i, $Pr(Y_i = 1) = p$. Suppose $q = 1 - p$ and $mq \geq 1$. For any integer k such that $k = mp - h \geq 1$, $h > 0$, if we define $\beta = \frac{1}{12k} + \frac{1}{12(m-k)}$, then,*

$$Pr\left(\sum_{i=1}^m Y_i = k\right) \geq \frac{1}{\sqrt{2\pi pqm}} \exp\left(-\frac{h^2}{2pqm} - \frac{h^3}{2p^2m^2} - \frac{h^4}{3q^3m^3} - \frac{h}{2qm} - \beta\right).$$

Proof: Robbins' formula ($m! = \left(\frac{m}{e}\right)^m \sqrt{2\pi m} \cdot e^{\alpha_m}$, where $1/(12m+1) \leq \alpha_m \leq 1/(12m)$) gives

$$\begin{aligned} Pr\left(\sum_{i=1}^m Y_i = k\right) &\geq \left(\frac{pm}{k}\right)^k \left(\frac{qm}{m-k}\right)^{m-k} \left(\frac{m}{2\pi k(m-k)}\right)^{1/2} e^{-\beta} \\ &= (2\pi pqm)^{-1/2} e^{-\beta} \left(\frac{pm}{k}\right)^{k+1/2} \left(\frac{qm}{m-k}\right)^{m-k+1/2} \\ &= (2\pi pqm)^{-1/2} e^{-\beta} \left(1 - \frac{h}{pm}\right)^{-k-1/2} \left(1 + \frac{h}{qm}\right)^{-m+k-1/2}. \end{aligned}$$

Since, for $t > 0$, $\ln(1+t) < t - \frac{1}{2}t^2 + \frac{1}{3}t^3$ and $\ln(1-t) < -t - \frac{1}{2}t^2 - \frac{1}{3}t^3$, we have

$$\begin{aligned} Pr\left(\sum_{i=1}^m Y_i = k\right) \geq (2\pi pqm)^{-1/2} e^{-\beta} \exp\Bigg((pm - h + 1/2) \cdot \left(\frac{h}{pm} + \frac{h^2}{2p^2m^2} + \frac{h^3}{3p^3m^3}\right) \\ -(qm + h + 1/2) \cdot \left(\frac{h}{qm} - \frac{h^2}{2q^2m^2} + \frac{h^3}{3q^3m^3}\right)\Bigg). \end{aligned}$$

Expanding this expression and noting that $h < mp$ and $mq \geq 1$ completes the proof. □

The following correlational result involving a "balls and bins" experiment will be useful.

Lemma 3 ([9, 2]) *Suppose we throw m balls independently at random into n bins, each ball having an arbitrary distribution. Let B_i be the random variable denoting the number of balls in the ith bin. Then for any $t_1, ..., t_n$, $\mathbf{Pr}\left(\bigwedge_{i=1}^{n} B_i \geq t_i\right) \leq \prod_{i=1}^{n} \mathbf{Pr}(B_i \geq t_i)$.*

We will also use the following, which has been previously used (see [7]).

Lemma 4 *For all $x \in [0,1]$ and all real $a \leq 1$,*

$$(1-a)^x \leq 1 - ax.$$

Proof: The LHS is convex in x, and the RHS is linear in x; the LHS and RHS are equal when x is 0 or 1. □

For $\vec{x} = (x_1, ..., x_m) \in X^m$, and $f : X \to [0,1]$, define $\hat{\mathbf{E}}_{\vec{x}}(f) = \frac{1}{m}\sum_{i=1}^{m} f(x_i)$ to be the sample average of f w.r.t. $\vec{x}$. For a probability distribution P over X, and a function f defined on X, let $\mathbf{E}_P(f)$ denote the expectation of $f(x)$ when x is drawn according to P.

Recall from the introduction that for $\nu > 0$ and $r, s \geq 0$, $d_\nu(r,s) = \frac{|r-s|}{\nu+r+s}$. We will find it useful in our analysis to extend the domain of d_ν to pairs r, s for which $r + s > -\nu$.

For a family $\mathcal{F}$ of $[0,1]$-valued functions defined on X, define $\mathrm{opt}(\mathcal{F}, \nu, \alpha, \delta)$ to be the least M such that for all $m \geq M$, for any probability distribution P over X, if m examples $\vec{x} = (x_1, ..., x_m)$ are drawn independently at random according to P, with probability at least $1 - \delta$, for all $f \in \mathcal{F}$, $d_\nu(\hat{\mathbf{E}}_{\vec{x}}(f), \mathbf{E}_P(f)) \leq \alpha$. Let $\mathrm{opt}(d, \nu, \alpha, \delta)$ be the maximum of $\mathrm{opt}(\mathcal{F}, \nu, \alpha, \delta)$ over all choices of $\mathcal{F}$ for which $\mathrm{Pdim}(\mathcal{F}) = d$. In other words, $\mathrm{opt}(d, \nu, \alpha, \delta)$ is the best possible bound on $\mathrm{opt}(\mathcal{F}, \nu, \alpha, \delta)$ in terms of $\mathrm{Pdim}(\mathcal{F})$, ν, α, and δ. The following is the main result of this paper.

Theorem 5 $\mathrm{opt}(d, \nu, \alpha, \delta) = \Theta\left(\frac{1}{\alpha^2 \nu}\left(d \log \frac{1}{\nu} + \log \frac{1}{\delta}\right)\right)$.

3 Upper bound

For each positive integer m, let Γ_m denote the set of all permutations of $\{1, \ldots, 2m\}$ that, for each $i \leq m$, either swap i and $m+i$, or leave both i and $m+i$ fixed. For any $g \in \mathbf{R}^{2m}$, and $\sigma \in \Gamma_m$, let $\mu_1(g, \sigma) = (1/m)\sum_{i=1}^{m} g_{\sigma(i)}$, and $\mu_2(g, \sigma) = (1/m)\sum_{i=1}^{m} g_{\sigma(m+i)}$.

We will make use of the following known lemma, which is proved by first bounding the probability that a sample gives rise to an inaccurate estimate in terms of the probability that two samples give rise to dissimilar estimates, and then applying the fact that any permutation that swaps corresponding elements of the two samples is equally likely.

Lemma 6 ([15, 10, 3]) *Choose a set $\mathcal{F}$ of functions from X to $[0,1]$, a probability distribution P over X, and $\nu > 0$, $0 < \alpha < 1$, and $m \geq 2/(\alpha^2\nu)$. Suppose U is the uniform distribution over Γ_m. Then,*

$$P^m\{\vec{x} : \exists f \in \mathcal{F},\ d_\nu(\hat{\mathbf{E}}_{\vec{x}}(f), \mathbf{E}_P(f)) > \alpha\}$$
$$\leq 2 \cdot \sup_{\vec{x} \in X^{2m}} U\left\{\sigma : \exists f \in \mathcal{F},\ d_\nu\left(\frac{1}{m}\sum_{i=1}^{m} f(x_{\sigma(i)}), \frac{1}{m}\sum_{i=1}^{m} f(x_{\sigma(m+i)})\right) > \alpha/2\right\}.$$

We will use the following lemma due to Haussler.

Lemma 7 ([3]) *Choose $m \in \mathbf{N}$. Let $g \in [0,1]^{2m}$, $\nu > 0$, and $0 < \alpha < 1$, and let U be the uniform distribution over Γ_m. Then $U\{\sigma : d_\nu(\mu_1(g,\sigma), \mu_2(g,\sigma)) > \alpha\} \leq 2e^{-2\alpha^2\nu m}$.*

Lemma 8 shows that when the L_1 norm of g is relatively small, one can get something stronger.

Lemma 8 *Choose $\nu, \alpha > 0$. Choose $m \in \mathbf{N}$ and $g \in [-1,1]^{2m}$ for which $\sum_{i=1}^{2m} |g_i| \leq c\nu m$ for some $c \leq 2/3$. Then if U is the uniform distribution over Γ_m, $U\{\sigma : d_\nu(\mu_1(g,\sigma), \mu_2(g,\sigma)) > \alpha\} \leq 2e^{-\alpha^2\nu m/36c}$.*

Proof: Expanding the definition of d_ν and simplifying, we get

$$d_\nu(\mu_1(g,\sigma), \mu_2(g,\sigma)) = \frac{\left|\sum_{i=1}^{m}\left(g_{\sigma(i)} - g_{\sigma(m+i)}\right)\right|}{\nu m + \sum_{i=1}^{2m} g_i}.$$

Also, note that $\sum_{i=1}^{m}(g_i - g_{m+i})^2 \leq \sum_{i=1}^{m} 2|g_i - g_{m+i}| \leq 2c\nu m$. One can sample uniformly from Γ_m by independently deciding whether σ swaps i and $m+i$ for $i = 1, ..., m$. Thus, applying the Hoeffding bound (Lemma 1) with the fact that $-|g_{\sigma(i)} - g_{\sigma(m+i)}| \leq g_{\sigma(i)} - g_{\sigma(m+i)} \leq |g_{\sigma(i)} - g_{\sigma(m+i)}|$, we get

$$\begin{aligned} U\{\sigma : d_\nu(\mu_1(g,\sigma), \mu_2(g,\sigma)) > \alpha\} &= U\left\{\sigma : \left|\sum_{i=1}^{m}(g_{\sigma(i)} - g_{\sigma(m+i)})\right| > \alpha\left(\nu m + \sum_{i=1}^{2m} g_i\right)\right\} \\ &\leq 2\exp\left(\frac{-\alpha^2(\nu m + \sum_{i=1}^{2m} g_i)^2}{4c\nu m}\right). \end{aligned}$$

Since $-c\nu m \leq \sum_{i=1}^{2m} g_i \leq c\nu m$ and $c \leq 2/3$, the term $(\nu m + \sum_{i=1}^{2m} g_i)^2$ takes its minimal value at $\sum_{i=1}^{2m} g_i = -c\nu m$. Therefore

$$U\{\sigma : d_\nu(\mu_1(g,\sigma), \mu_2(g,\sigma)) > \alpha\} \leq 2\exp\left(\tfrac{-\alpha^2(\nu m - c\nu m)^2}{4c\nu m}\right).$$

Since $c \leq 2/3$, the lemma follows. □

For $\vec{v}, \vec{w} \in \mathbf{R}^k$, let $\ell_1(\vec{v}, \vec{w}) = \frac{1}{k}\sum_{i=1}^{k} |v_i - w_i|$. For $F \subseteq \mathbf{R}^k$, $\vec{v} \in \mathbf{R}^k$, define $\ell_1(\vec{v}, F) = \min\{\ell(\vec{v}, \vec{f}) : \vec{f} \in F\}$; if $F = \emptyset$, then $\ell_1(\vec{v}, F) = \infty$. The following result of [4] bounds the size of a "well-separated" set of a certain pseudo-dimension:

Lemma 9 ([4]) *For all $k \in \mathbf{N}$, for all $0 < \epsilon \leq 1$, if each pair f, g of distinct elements of some $F \subseteq [0,1]^k$ has $\ell_1(f,g) > \epsilon$, then $|F| \leq (41/\epsilon)^{\mathrm{Pdim}(F)}$.*

The following is the key lemma in our analysis, and is a new application of chaining.

Lemma 10 *Choose $d \in \mathbf{N}$. Choose an integer $m \geq \frac{125(2d+1)}{\alpha^2 \nu}$ and $F \subseteq [0,1]^{2m}$ for which $\mathrm{Pdim}(F) = d$. Then if U is the uniform distribution over Γ_m, for any $\alpha > 0$, $\nu > 0$,*

$$U\left\{\sigma : \exists f \in F,\ d_\nu\left(\mu_1(f,\sigma), \mu_2(f,\sigma)\right) > \alpha\right\} \leq 6 \cdot (2624/\nu)^d e^{-\alpha^2 \nu m/90}.$$

Proof: Let $F_{-1} = \emptyset$. For each nonnegative integer j, construct F_j by initializing it to F_{j-1}, and as long as there is a $f \in F$ for which $\ell_1(f, F_j) > \nu/2^{2j+4}$, choosing such an f and adding it to F_j. For each $f \in F$ and each $j \geq 0$ choose an element $\psi_j(f)$ of F_j such that $\ell_1(f, \psi_j(f))$ is minimized. (Since $\ell_1(\vec{v}, \vec{w}) > \nu/2^{2j+4}$ for distinct $\vec{v}, \vec{w} \in F_j$, F_j is finite by Lemma 9. So this minimum is well-defined.) We have $\ell_1(f, \psi_j(f)) \leq \nu/2^{2j+4}$, as otherwise f would have been added to F_j. Let $G_0 = F_0$, and for each $j > 0$, define G_j to be $\{f - \psi_{j-1}(f) : f \in F_j\}$. Since $\ell_1(f, \psi_{j-1}(f)) \leq \nu/2^{2j+2}$, we have for all $g \in G_j$ that $\sum_{i=1}^{2m} |g_i| \leq \nu m/2^{2j+1}$. By induction, for each k, each $f \in F_k$ has $g_{f,0} \in G_0, \cdots, g_{f,k} \in G_k$ such that $f = \sum_{j=0}^{k} g_{f,j}$. Let $F_* = \cup_k F_k$. Since for all $f \in F$, for all k we have $\ell_1(f, \psi_k(f)) \leq \nu/2^{2k+4}$, F_* is dense in F w.r.t. ℓ_1. Define

$$p = U\left\{\sigma : \exists f \in F,\ d_\nu\left(\mu_1(f,\sigma), \mu_2(f,\sigma)\right) > \alpha\right\}.$$

Since F_* is dense in F,

$$p = U\left\{\sigma : \exists f \in F_*, d_\nu\left(\mu_1(f,\sigma), \mu_2(f,\sigma)\right) > \alpha\right\}$$

and thus,

$$p = U\left\{\sigma : \exists f \in F_*,\ \left|\sum_{i=1}^{m}\left(f_{\sigma(i)} - f_{\sigma(m+i)}\right)\right| > \alpha\left(\nu m + \sum_{i=1}^{2m} f_i\right)\right\}.$$

For each $f \in F_*$, there are $g_{f,0} \in G_0, g_{f,1} \in G_1, \ldots$ such that $f = \sum_{j=0}^{\infty} g_{f,j}$ (only a finite number of the $g_{f,j}$'s are nonzero). Applying the triangle inequality, we see that

$$p \leq U\left\{\sigma : \exists f \in F_*,\ \sum_{j=0}^{\infty}\left|\sum_{i=1}^{m}\left((g_{f,j})_{\sigma(i)} - (g_{f,j})_{\sigma(m+i)}\right)\right| > \alpha\left(\nu m + \sum_{j=0}^{\infty}\sum_{i=1}^{2m}(g_{f,j})_i\right)\right\}.$$

Let $\nu_0 = \nu/3$, and for each $j \in \mathbf{N}$, let $\nu_j = \nu\sqrt{j+1}/(3 \cdot 2^j)$. Then $\sum_{j=0}^{\infty} \nu_j \leq \nu$, and hence,

$$p \leq U\left\{\sigma : \exists f \in F_*,\ \sum_{j=0}^{\infty}\left|\sum_{i=1}^{m}\left((g_{f,j})_{\sigma(i)} - (g_{f,j})_{\sigma(m+i)}\right)\right| > \sum_{j=0}^{\infty}\alpha\left(\nu_j m + \sum_{i=1}^{2m}(g_{f,j})_i\right)\right\}$$

$$\leq \sum_{j=0}^{\infty} U\left\{\sigma : \exists f \in F_*,\ \left|\sum_{i=1}^{m}\left((g_{f,j})_{\sigma(i)} - (g_{f,j})_{\sigma(m+i)}\right)\right| > \alpha\left(\nu_j m + \sum_{i=1}^{2m}(g_{f,j})_i\right)\right\}$$

$$\leq \sum_{j=0}^{\infty} U\left\{\sigma : \exists g \in G_j,\ \left|\sum_{i=1}^{m}\left(g_{\sigma(i)} - g_{\sigma(m+i)}\right)\right| > \alpha\left(\nu_j m + \sum_{i=1}^{2m} g_i\right)\right\}, \tag{2}$$

since each $g_{f,j} \in G_j$.

Choose $j > 0$. For each $g \in G_j$, $\sum_{i=1}^{2m} |g_i| \leq \nu m/2^{2j+1}$, and $\nu_j m = \frac{\nu m}{3}\sqrt{j+1}/2^j$. By applying Lemma 8 with $c_j = 3/(\sqrt{j+1}2^{j+1})$, we see that

$$U\left\{\sigma : \exists g \in G_j, \quad \left|\sum_{i=1}^{m}\left(g_{\sigma(i)} - g_{\sigma(m+i)}\right)\right| > \alpha\left(\nu_j m + \sum_{i=1}^{2m} g_i\right)\right\} \leq 2|G_j| \exp\left(\frac{-\alpha^2 \nu_j m}{36 c_j}\right).$$

Plugging in the values of ν_j and c_j, we get

$$U\left\{\sigma : \exists g \in G_j,\ \left|\sum_{i=1}^{m}\left(g_{\sigma(i)} - g_{\sigma(m+i)}\right)\right| > \alpha\left(\nu_j m + \sum_{i=1}^{2m} g_i\right)\right\} \leq 2|G_j| \exp\left(-\alpha^2 \nu m (j+1)/180\right).$$

Distinct elements $\vec{v}$ and $\vec{w}$ of F_j have $\ell_1(\vec{v}, \vec{w}) > \nu/2^{2j+4}$; so by Lemma 9, $|G_j| \leq (164 \cdot 4^{j+1}/\nu)^d$. Thus

$$\begin{aligned}
&\sum_{j=1}^{\infty} U\left\{\sigma : \exists g \in G_j,\ \left|\sum_{i=1}^{m}\left(g_{\sigma(i)} - g_{\sigma(m+i)}\right)\right| > \alpha\left(\nu_j m + \sum_{i=1}^{2m} g_i\right)\right\} \\
&\leq \sum_{j=1}^{\infty} 2(164 \cdot 4^{j+1}/\nu)^d e^{-\alpha^2 \nu m (j+1)/180} \\
&\leq 4(164 \cdot 16/\nu)^d e^{-\alpha^2 \nu m/90}.
\end{aligned} \tag{3}$$

Note that $G_0 = F_0$, and therefore the elements of G_0 are in $[0,1]^{2m}$. Thus, we can apply Lemma 7 to get

$$U\left\{\sigma : \exists g \in G_0,\ \left|\sum_{i=1}^{m}\left(g_{\sigma(i)} - g_{\sigma(m+i)}\right)\right| > \alpha\left(\nu_0 m + \sum_{i=1}^{2m} g_i\right)\right\} \leq 2|G_0| \exp(-2\alpha^2 \nu_0 m).$$

Substituting the value of ν_0, upper bounding the size of $|G_0|$ using Lemma 9, and combining with (2) and (3) completes the proof. □

Combining Lemma 10 with Lemma 6, and solving for m proves the upper bound of Theorem 5.

4 Lower bound

In this section, we establish the lower bound side of Theorem 5. For positive integers d and n, we define $X_{d,n}$ to be an arbitrary set of nd elements of X. We view $X_{d,n}$ as the union of d disjoint

subsets, which we will call *types*; there will be n elements of each type. We refer to the jth element in type i as $a_{i,j}$. Let $P_{d,n}$ be the uniform distribution on $X_{d,n}$. The function class $\mathcal{F}_d$ consists of all functions mapping X to $\{0,1\}$ that take the value 1 on at most one point in each type, and take the value 0 outside of $X_{d,n}$. It is easy to check that the pseudo-dimension of $\mathcal{F}_d$ is d.

We begin by establishing the first term of the lower bound.

Theorem 11 *For any real* $0 < \nu \leq 1/100$, $0 < \alpha \leq 1/100$, $0 < \delta \leq 1/5$ *and any integer* $d \geq 1$, $\mathrm{opt}(\mathcal{F}_d, \nu, \alpha, \delta) > \frac{d}{30\alpha^2\nu} \ln \frac{1}{3\nu}$.

Proof: Suppose d, α and ν are given. Set $n = \lfloor \frac{1}{\nu} \rfloor$ and $P = P_{d,n}$. We will show that if $m = \lfloor d \ln(1/(3\nu))/(30\alpha^2\nu) \rfloor$, then

$$P^m \left\{ \vec{x} : \exists f \in \mathcal{F}_d,\ d_\nu \left(\hat{\mathbf{E}}_{\vec{x}}(f), \mathbf{E}_P(f) \right) > \alpha \right\} > 1/5,$$

proving the theorem. For a sample $\vec{x}$, for each type i, and each $j \in \{1, ..., n\}$, let $B_{i,j}(\vec{x})$ denote the number of times that $a_{i,j}$ appears in $\vec{x}$. If

$$p = P^m \left\{ \vec{x} : \exists f \in \mathcal{F}_d,\ d_\nu \left(\hat{\mathbf{E}}_{\vec{x}}(f), \mathbf{E}_P(f) \right) > \alpha \right\}$$

we have

$$\begin{aligned}
p &\geq P^m \left\{ \vec{x} : \exists f \in \mathcal{F}_d,\ d_\nu \left(\hat{\mathbf{E}}_{\vec{x}}(f), \mathbf{E}_P(f) \right) > \alpha,\ \ \hat{\mathbf{E}}_{\vec{x}}(f) < \mathbf{E}_P(f) \right\} \\
&= P^m \left\{ \vec{x} : \exists f \in \mathcal{F}_d,\ \frac{\mathbf{E}_P(f) - \hat{\mathbf{E}}_{\vec{x}}(f)}{\mathbf{E}_P(f) + \hat{\mathbf{E}}_{\vec{x}}(f) + \nu} > \alpha \right\} \\
&= P^m \left\{ \vec{x} : \exists f \in \mathcal{F}_d,\ \ \hat{\mathbf{E}}_{\vec{x}}(f) < \mathbf{E}_P(f) \frac{1-\alpha}{1+\alpha} - \frac{\alpha\nu}{1+\alpha} \right\} \\
&\geq P^m \left\{ \vec{x} : \exists f \in \mathcal{F}_d,\ \ \hat{\mathbf{E}}_{\vec{x}}(f) < \mathbf{E}_P(f) \frac{1-\alpha}{1+\alpha} - \frac{\alpha\nu}{1+\alpha}, \mathbf{E}_P(f) \geq \frac{1}{2n} \right\} \\
&\geq P^m \left\{ \vec{x} : \exists f \in \mathcal{F}_d,\ \hat{\mathbf{E}}_{\vec{x}}(f) < \mathbf{E}_P(f) \frac{1-3\alpha}{1+\alpha},\ \ \mathbf{E}_P(f) \geq \frac{1}{2n} \right\} \\
&\geq P^m \left\{ \vec{x} : \exists f \in \mathcal{F}_d,\ \hat{\mathbf{E}}_{\vec{x}}(f) < \mathbf{E}_P(f)(1-4\alpha),\ \ \mathbf{E}_P(f) \geq \frac{1}{2n} \right\} \\
&\geq P^m \left\{ \vec{x} : \exists I \subseteq \{1, ..., d\},\ |I| = \lceil d/2 \rceil,\ \ \forall i \in I,\ \exists j,\ B_{i,j}(\vec{x}) < \frac{m}{nd}(1-4\alpha) \right\}.
\end{aligned}$$

Let ϕ_i be the indicator function for the event that there exists $j \in \{1, ..., n\}$ for which $B_{i,j}(\vec{x}) < \frac{m}{nd}(1-4\alpha)$. Then,

$$p \geq P^m \left\{ \vec{x} : \sum_{i=1}^{d} \phi_i(\vec{x}) \geq \lceil d/2 \rceil \right\}. \tag{4}$$

Fix i, j, and let $r = 1/(nd)$. Since $\mathbf{E}_{P^m}[B_{i,j}(\vec{x})] = mr$, a simple calculation with binomial coefficients shows that

$$P^m\{\vec{x} : B_{i,j}(\vec{x}) = y\} \leq P^m\{\vec{x} : B_{i,j}(\vec{x}) = y+1\}$$

for $y = 0, 1, \ldots, \lfloor mr \rfloor - 1$. Thus,

$$P^m\{\vec{x} : B_{i,j}(\vec{x}) < mr(1-4\alpha)\} \geq \lceil\sqrt{mr}\rceil \cdot P^m\{\vec{x} : B_{i,j}(\vec{x}) = \lceil mr(1-4\alpha)\rceil - \lceil\sqrt{mr}\rceil\}.$$

If we put $h = mr - (\lceil mr(1-4\alpha)\rceil - \lceil\sqrt{mr}\rceil)$, $p = r$, $q = 1-r$ and apply Lemma 2, we obtain

$$\begin{aligned} P^m\left\{\vec{x} : B_{i,j}(\vec{x}) < mr(1-4\alpha)\right\} \quad &\geq \quad \sqrt{mr}\sqrt{\frac{1}{2\pi mrq}} \times \exp\left(-\frac{h^2}{2rqm} - \frac{h^3}{2r^2m^2} - \frac{h^4}{3q^3m^3} - \frac{h}{2qm}\right) \\ &\quad \times \exp\left(-\frac{1}{12}\left(\frac{1}{mr-h} + \frac{1}{mq+h}\right)\right). \end{aligned}$$

Because $r \leq 1/100$, $q = 1-r \geq 1-1/100$, $\alpha \leq 1/100$, $m \geq \lfloor \frac{\ln(100/3)}{30\alpha^2 r} \rfloor$ and $mr - h \geq 1$, a simple calculation shows that

$$P^m\{\vec{x} : B_{i,j}(\vec{x}) < mr(1-4\alpha)\} > (e^{-29\alpha^2 rm})/3.$$

Crucially, by Lemma 3,

$$\begin{aligned} P^m\{\vec{x} : \phi_i(\vec{x}) = 1\} \quad &= \quad 1 - P^m\{\vec{x} : \forall j, B_{i,j}(\vec{x}) \geq mr(1-4\alpha)\} \\ &\geq \quad 1 - (1 - P^m\{\vec{x} : B_{i,j}(\vec{x}) < mr(1-4\alpha)\})^n \\ &> \quad 1 - (1 - \frac{1}{3}e^{-29\alpha^2 rm})^n \\ &> \quad 1 - e^{-0.99}. \end{aligned}$$

Let $z = P^m\{\vec{x} : \sum_{i=1}^d \phi_i(\vec{x}) \leq \lceil d/2\rceil - 1\}$. Then

$$\left(1 - e^{-0.99}\right) d < \mathbf{E}(\sum_{i=1}^d \phi_i(\vec{x})) \leq dz/2 + (1-z)d.$$

Solving the above inequality, we have $z < 2/e^{0.99} < 4/5$. This implies that $P^m\{\vec{x} : \sum_{i=1}^d \phi_i(\vec{x}) \geq \lceil d/2\rceil\} > 1/5$, which, since $p \geq P^m\{\vec{x} : \sum_{i=1}^d \phi_i(\vec{x}) \geq \lceil d/2\rceil\}$ by (4), completes the proof. □

A similar proof establishes the second term in the lower bound:

Theorem 12 *For any real $0 < \nu \leq 1/100$, $0 < \alpha \leq 1/100$, $0 < \delta \leq 1/5$ and any integer $d \geq 1$,* $\mathrm{opt}(\mathcal{F}_d, \nu, \alpha, \delta) > \frac{1}{30\alpha^2\nu} \ln \frac{1}{6\delta\nu}$.

Proof: Choose $d \in \mathbf{N}$ and $0 < \alpha, \nu \leq 1/100$. Set $n = \lfloor \frac{1}{\nu} \rfloor$. Let P be the distribution that allocates probability $1/n$ to each of $a_{1,1}, ..., a_{1,n}$.

Here, we will show that if $m = \left\lfloor \frac{1}{30\alpha^2\nu} \ln \frac{1}{6\delta\nu} \right\rfloor$, then

$$P^m \left\{ \vec{x} : \exists f \in \mathcal{F}_d,\ d_\nu \left(\hat{\mathbf{E}}_{\vec{x}}(f), \mathbf{E}_P(f) \right) > \alpha \right\} > \delta$$

which will prove the theorem.

For some sample $\vec{x}$, for each $j \in \{1, ..., n\}$, let $B_j(\vec{x})$ denote the number of times that $a_{1,j}$ appears in $\vec{x}$. If

$$p = P^m \left\{ \vec{x} : \exists f \in \mathcal{F}_d,\ d_\nu \left(\hat{\mathbf{E}}_{\vec{x}}(f), \mathbf{E}_P(f) \right) > \alpha \right\}$$

is the quantity we wish to lower bound, arguing as in the proof of Theorem 11, we have

$$\begin{aligned} p &\geq P^m \left\{ \vec{x} : \exists f \in \mathcal{F}_d,\ \ \hat{\mathbf{E}}_{\vec{x}}(f) < \mathbf{E}_P(f) \frac{1-2\alpha}{1+\alpha}, \mathbf{E}_P(f) = \frac{1}{n} \right\} \\ &\geq P^m \left\{ \vec{x} : \exists f \in \mathcal{F}_d,\ \ \hat{\mathbf{E}}_{\vec{x}}(f) < \mathbf{E}_P(f)(1-3\alpha),\ \mathbf{E}_P(f) = \frac{1}{n} \right\}. \end{aligned}$$

Thus

$$p \geq P^m \left\{ \vec{x} : \exists j \in \{1, ..., n\}\ B_j(\vec{x}) < \frac{m}{n}(1-3\alpha) \right\}. \tag{5}$$

Choose $j \in \{1, ..., n\}$. If $r = 1/n$, we have

$$\begin{aligned} P^m \left\{ \vec{x} : B_j(\vec{x}) < mr(1-3\alpha) \right\} &= P^m \left\{ \vec{x} : B_j(\vec{x}) \leq \lceil mr(1-3\alpha) \rceil - 1 \right\} \\ &\geq \lceil \sqrt{mr} \rceil P^m \left\{ \vec{x} : B_j(\vec{x}) = \lceil mr(1-3\alpha) \rceil - \lceil \sqrt{mr} \rceil \right\}. \end{aligned}$$

Let $h = mr - (\lceil mr(1-3\alpha) \rceil - \lceil \sqrt{mr} \rceil)$, $p = r$, and $q = 1 - r$. As before, application of Lemma 2 yields

$$\begin{aligned} P^m \left\{ \vec{x} : B_j(\vec{x}) < mr(1-3\alpha) \right\} &\geq \sqrt{mr} \sqrt{\frac{1}{2\pi mrq}} \times \exp\left(-\frac{h^2}{2rqm} - \frac{h^3}{2r^2m^2} - \frac{h^4}{3q^3m^3} - \frac{h}{2qm} \right) \\ &\quad \times \exp\left(-\frac{1}{12} \left(\frac{1}{mr-h} + \frac{1}{mq+h} \right) \right) \end{aligned}$$

Because $r \leq 0.01$, $q = 1 - r \geq 0.99$, $\alpha \leq 0.01$, and $m \geq \lfloor \frac{\ln(500/6)}{30\alpha^2 r} \rfloor$, $mr - h \geq 1$, a simple calculation shows that

$$P^m \left\{ \vec{x} : B_j(\vec{x}) < mr(1-3\alpha) \right\} > \frac{1}{3} e^{-29\alpha^2 rm}.$$

Applying (5),

$$p \geq P^m \left\{ \vec{x} : \exists j, B_j(\vec{x}) < mr(1-3\alpha) \right\}$$

$$
\begin{aligned}
&= 1 - P^m\{\vec{x} : \forall j, B_j(\vec{x}) \geq mr(1-3\alpha)\} \\
&\geq 1 - \prod_{j=1}^{n} P^m\{\vec{x} : B_j(\vec{x}) \geq mr(1-3\alpha)\} \qquad \text{(by Lemma 3)} \\
&= 1 - (1 - P^m\{\vec{x} : B_j(\vec{x}) < mr(1-3\alpha)\})^n \\
&> 1 - (1 - \frac{1}{3}e^{-29\alpha^2 rm})^n \\
&\geq 1 - (1 - \frac{1}{3}e^{-\frac{29}{30}\frac{1}{n\nu}\ln\frac{1}{6\delta\nu}})^n \qquad \text{(since} \quad m = \left\lfloor \tfrac{1}{30\alpha^2\nu}\ln\tfrac{1}{6\delta\nu} \right\rfloor \text{)} \\
&> 1 - (1 - \frac{1}{3}e^{-\ln\frac{1}{6\delta\nu}})^n \qquad \text{(since} \quad \tfrac{29}{30}\tfrac{1}{n\nu} < \tfrac{29}{30}\tfrac{1}{1-\nu} < 1 \text{)} \\
&= 1 - (1 - 2\delta\nu)^n \\
&> 1 - (1 - 2\delta\frac{1}{n+1})^n \\
&> \delta,
\end{aligned}
$$

by Lemma 4, since $n \geq 100$. □

Acknowledgements. We thank the reviewers of this article for their helpful comments.

References

[1] B. Bollobas. *Random Graphs.* Academic Press, 1985.

[2] Devdatt Dubhashi and Desh Ranjan. Balls and bins: A study in negative dependence. *Random Structures & Algorithms*, 13(2):99–124, Sept 1998.

[3] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

[4] 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.

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

[6] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *American Statistical Association Journal*, 58:13–30, 1963.

[7] N. Littlestone. *Mistake Bounds and Logarithmic Linear-threshold Learning Algorithms.* PhD thesis, UC Santa Cruz, 1989.

[8] P. M. Long. The complexity of learning according to two models of a drifting environment. *Proceedings of the 1998 Conference on Computational Learning Theory*, 1998.

[9] C. L. Mallows. An inequality involving multinomial probabilities. *Biometrika*, 55:422–424, 1968.

[10] D. Pollard. *Convergence of Stochastic Processes*. Springer Verlag, 1984.

[11] D. Pollard. Rates of uniform almost-sure convergence for empirical processes indexed by unbounded classes of functions, 1986. Manuscript.

[12] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Annals of Probability*, 22:28–76, 1994.

[13] L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

[14] V. N. Vapnik. Inductive principles of the search for empirical dependences (methods based on weak convergence of probability measures). *Proceedings of the 1989 Workshop on Computational Learning Theory*, 1989.

[15] 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.