IV. Truncation Errors

As a byproduct of Theorem 3.1 we can give an estimate of the truncation error which arises if one ignores all the samples outside a finite interval. More precisely, we have the following corollary.

Theorem 4.1: Let us define the truncation error $E_N(t)$ as follows:

$$E_N(t) = F(t) - \sum_{|n|\le N} F(t_n) S_n(t).$$

Then

$$|E_N(t)| \le \frac{2^{2p-q} b A^p C_{\hat{\gamma}}}{(1-\pi\sqrt{D})(p-q-1)N^{p-q-1}}, \quad t \text{ in the compact set } K \qquad (14)$$

where $p > q+1$, and q the polynomial order of growth of $F(t)$.

Proof: We know from Theorem 3.1 that $|F(t_n)| \le b|t_n|^q$ for some $q \ge 0$ and a constant b, and

$$|S_n(t)| \le \frac{2^p A^p C_{\hat{\gamma}}}{(1-\pi\sqrt{D})|t_n|^p} = \frac{C_K}{|t_n|^p}, \quad \text{for all } p \ge 0 \text{ and } t \text{ in the compact set } K.$$

Hence, if we take any $p > q+1$, we will have

$$\begin{aligned}|E_N(t)| &\le \sum_{|n|>N} |F(t_n)|\,|S_n(t)| \\ &\le bC_K \sum_{|n|>N} \frac{1}{|t_n|^{p-q}} \\ &\le bC_K \left(\sum_{n=-\infty}^{-(N+1)} \left(\left| n + \frac{1}{4} \right| \right)^{q-p} + \sum_{n=N+1}^{\infty} \left(n - \frac{1}{4} \right)^{q-p} \right) \\ &\le 2bC_K \int_{N-(1/4)}^{\infty} \frac{1}{(x-\frac{1}{4})^{p-q}}\, dx.\end{aligned}$$

If we use the change of variable, $x - 1/4 = Nt$, and note that $1/2 \le [N-(1/2)]/N, \forall N \ge 1$, we obtain

$$\begin{aligned}|E_N(t)| &\le 2bC_K \int_{N-(1/4)}^{\infty} \frac{1}{(x-\frac{1}{4})^{p-q}}\, dx \le \frac{2bC_K}{N^{p-q-1}} \int_{1/2}^{\infty} \frac{dt}{t^{p-q}} \\ &= \frac{2bC_K}{2^{q-p+1}(p-q-1)N^{p-q-1}}.\end{aligned}$$

Therefore, replacing the constant C_K by its value, we obtain the desired result (14). □

Acknowledgment

This work was performed while the first author was visiting the "Universidad Carlos III de Madrid" in Spain and he wishes to take this opportunity to thank the Mathematics Department for its hospitality and stimulating atmosphere.

References

[1] L. L. Campbell, "Sampling theorem for the Fourier transform of a distribution with bounded support," *SIAM J. Appl. Math.*, vol. 16, no. 3, pp. 626–636, 1968,

[2] H. G. Feichtinger and K. Gröchenig, "Irregular sampling theorems and series expansions of band-limited functions," *J. Math. Anal. Appl.*, vol. 167, pp. 530–556, 1992.

[3] ——, "Iterative reconstructions of multivariate band-limited functions from irregular sampling values," *SIAM J. Math. Anal.*, vol. 26, pp. 244–261, 1992.

[4] R. F. Hoskins and J. De Sousa Pinto, "Sampling expansions for functions band-limited in the distributional sense," *SIAM J. Appl. Math.*, vol. 44, pp. 605–610, 1984.

[5] A. J. Lee, "Characterization of band-limited functions and processes," *Inform. Contr.*, vol. 31, pp. 258–271, 1976.

[6] N. Levinson, *Gap and Density Theorems* (Amer. Math. Soc. Colloq. Pub. Ser.), vol. 26. New York: Amer. Math. Soc., 1940.

[7] R. Paley and N. Wiener, *Fourier Transforms in the Complex Domain* (Amer. Math. Soc. Colloq. Pub. Ser.), vol. 19. Providence, RI: Amer. Math. Soc., 1934.

[8] E. Pfaffelhuber, "Sampling series for band-limited generalized functions," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 650–654, 1971.

[9] G. G. Walter, "Sampling bandlimited functions of polynomial growth," *SIAM J. Math. Anal.*, vol. 19, no. 5, pp. 1198–1203, 1988.

[10] ——, "Nonuniform sampling of bandlimited functions of polynomial growth," *SIAM J. Math. Anal.*, vol. 23, no. 4, pp. 995–1003, 1992.

[11] M. Zakai, "Band-limited functions and the sampling theorem," *Inform. Contr.*, vol. 8, pp. 143–158, 1965.

[12] A. I. Zayed, *Advances in Shannon's Sampling Theorem*. Boca Raton, FL: CRC Press, 1993.

[13] A. H. Zemanian, *Distribution Theory and Transform Analysis*. New York: MacGraw-Hill, 1965.

[14] ——, *Generalized Integral Transformations*. New York: Dover, 1987.

Covering Numbers for Real-Valued Function Classes

P. L. Bartlett, *Member, IEEE*,
S. R. Kulkarni, *Senior Member, IEEE*,
and S. E. Posner

***Abstract*—We find tight upper and lower bounds on the growth rate for the covering numbers of functions of bounded variation in the $\mathcal{L}_1$ metric in terms of all the relevant constants. We also find upper and lower bounds on covering numbers for general function classes over the family of $\mathcal{L}_1(dP)$ metrics in terms of a scale-sensitive combinatorial dimension of the function class.**

***Index Terms*— Bounded variation, covering numbers, fat-shattering dimension, metric entropy, scale-sensitive dimension, VC dimension.**

I. Introduction

Covering numbers have been studied extensively in a variety of literature dating back to the work of Kolmogorov [10], [12]. They play a central role in a number of areas in information theory and statistics, including density estimation, empirical processes, and machine learning (see, for example, [4], [8], and [16]). Let $\mathcal{F}$ be a subset of a metric space $(\mathcal{X}, \rho)$. For a given $\epsilon > 0$, the metric covering number $\mathcal{N}(\epsilon, \mathcal{F}, \rho)$ is defined as the smallest number of sets of radius ϵ whose union contains $\mathcal{F}$. (We omit ρ if the context is clear.)

Manuscript received February 12, 1996; revised March 31, 1997. This work was supported in part by the NSF under Grant NYI IRI-9457645. This work was done while S. E. Posner was with the Department of Electrical Engineering, Princeton University, Princeton, NJ, and the Department of Statistics, University of Toronto, Toronto, Ont., Canada.

P. L. Bartlett is with the Department of Systems Engineering, Australian National University, Canberra 0200, Australia.

S. R. Kulkarni is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA.

S. E. Posner is with ING Baring Securities, New York, NY 10021 USA.

Publisher Item Identifier S 0018-9448(97)05711-8.

We first find bounds on the covering numbers of functions of bounded variation under the $\mathcal{L}_1$ metric. Specifically, let $\mathcal{F}_1$ be the set of all functions on $[0, T]$ taking values in $[-V/2, V/2]$ with total variation at most $V > 0$. (It is natural to use the same parameter V for the range and variation, since a bound on the variation of a function implies a bound on its range.) We find tight bounds on $\mathcal{N}(\epsilon, \mathcal{F}_1, \mathcal{L}_1)$ in terms of the relevant constants using a rather simple proof. To our knowledge this result has not appeared in the literature, although some related work has been done in the context of density estimation where attention has been given to the problem of finding covering numbers for the classes of densities that are unimodal or nondecreasing (see [4] and [7] for references). These classes are contained in the classes we consider in this correspondence. However, we do not impose a density constraint on the class which accounts for the difference in behavior as a function of the parameters compared to the work of Birgé [4]. In fact, it is the density constraint that accounts for the $\log(VT)$ constant (in [4]) rather than the VT constant that we obtain in Theorem 1. Using a very simple proof we obtain tight upper and lower bounds.

We also investigate the metric covering numbers for general classes of real-valued functions under the family of $\mathcal{L}_1(dP)$ metrics, where P is a probability distribution. Upper bounds in terms of the Vapnik–Chervonenkis and/or pseudodimension of the function class were first provided by Dudley [6] and improved by Haussler [8], [9], and Pollard [16]. Various lower bounds have also been obtained (e.g., see [13]). Recent work has shown the importance of scale-sensitive versions of the various combinatorial dimensions in learning problems (see, for example, [1] and [3]). Using techniques due to Haussler and results from Alon *et al.* [1], Lee *et al.* [14] proved an upper bound on $\max_P \log_2 \mathcal{N}(\epsilon, \mathcal{F}, \mathcal{L}_1(dP))$ in terms of a scale-sensitive dimension of the function class. We improve on this result, and provide lower bounds. As will be shown, in general our bounds cannot be significantly improved.

II. Functions of Bounded Variation

Our first result is to obtain tight upper and lower bounds on the covering numbers for the class $\mathcal{F}_1$ of functions of bounded variation.

Theorem 1: For all $\epsilon \leq VT/12$

$$(\log_2 e)\frac{VT}{54\epsilon} \leq \log_2 \mathcal{N}(\epsilon, \mathcal{F}_1, \mathcal{L}_1) \leq \frac{13VT}{\epsilon}.$$

Comments:

- Certainly under the $\mathcal{L}_\infty$ metric, the class of functions of bounded variation (or even the subset of functions in this class that are also continuous) is not precompact, and hence $\mathcal{N}(\epsilon, \mathcal{F}_1, \mathcal{L}_\infty) = \infty$. However, it has been established (see [15]) that the class of functions of bounded variation that are also Lipschitz-smooth have covering numbers of order $(1/\epsilon)^{1/\epsilon}$ in the $\mathcal{L}_\infty$ metric.
- Let $\mathcal{F}_2$ be all functions of variation at most V that map from $[0, T]$ to $[-B, B]$, for some $B > V/2$. The theorem can be extended to $\mathcal{F}_2$ as
$$\frac{VT}{54\epsilon} + \log_2 \frac{eBT}{6\epsilon} \leq \log_2 \mathcal{N}(\epsilon, \mathcal{F}_2, \mathcal{L}_1) \leq \frac{39VT}{2\epsilon} + \log_2 \frac{3(2B - V)T}{8\epsilon}.$$
Both the upper and lower bounds are obtained by considering vertical shifts of $\mathcal{F}_1$ using the proof below.
- The upper bound (with T set to 1) in the theorem can be extended to give the upper bound
$$\max_P \log_2 \mathcal{N}(\epsilon, \mathcal{F}_1, \mathcal{L}_1(dP)) \leq \frac{13V}{\epsilon}$$
i.e., a uniform bound on the covering numbers for all weighted $\mathcal{L}_1$ norms where P is an arbitrary probability measure on $[0, T]$. The proof is modified by taking an equiprobable partition of $[0, T]$ rather than a partition of equal size. In the next section, we consider upper and lower bounds on this quantity for general function classes.

Proof of Theorem 1:

Upper Bound: Define $\mathcal{I}$ as the set of all nondecreasing functions on $[0, T]$ taking values in $[0, V]$. Let $\mathcal{G}(\epsilon/2)$ be any $\epsilon/2$-covering of $\mathcal{I}$. It is well known that for each $f \in \mathcal{F}_1$ there exists $g, h \in \mathcal{I}$ such that $f = g - h$ (see, e.g., [11, Theorem 4, p. 331]). More precisely, if $v(x)$ denotes the total variation over the interval $[0, x]$, then g and h can be taken to be

$$g(x) = [v(x) + f(x)]/2 + V/4$$

and

$$h(x) = [v(x) - f(x)]/2 + V/4.$$

It is easy to show that both g and h are nondecreasing. Also, if f takes values in $[-V/2, V/2]$ and has total variation bounded by V, then both g and h take values only in the range $[0, V]$.

By the definition of a cover of a set, there exists $\phi_1, \phi_2 \in \mathcal{G}(\epsilon/2)$ such that

$$\|g - \phi_1\|_{\mathcal{L}_1} \leq \epsilon/2 \qquad \|h - \phi_2\|_{\mathcal{L}_1} \leq \epsilon/2.$$

This gives

$$\begin{aligned}\|f - (\phi_1 - \phi_2)\|_{\mathcal{L}_1} &= \|g - h - (\phi_1 - \phi_2)\|_{\mathcal{L}_1} \\ &\leq \|g - \phi_1\|_{\mathcal{L}_1} + \|h - \phi_2\|_{\mathcal{L}_1} \leq \epsilon.\end{aligned}$$

Thus we can produce an ϵ-covering of $\mathcal{F}_1$ by taking all pairs from $\mathcal{G}(\epsilon/2) \times \mathcal{G}(\epsilon/2)$. Hence

$$\mathcal{N}(\epsilon, \mathcal{F}_1, \mathcal{L}_1) \leq (\mathcal{N}(\epsilon/2, \mathcal{I}, \mathcal{L}_1))^2. \tag{1}$$

We now find an upper bound on $\mathcal{N}(\epsilon, \mathcal{I}, \mathcal{L}_1)$ by producing an ϵ-covering of $\mathcal{I}$. Partition $[0, T]$ into $n_1 \geq T/h_1 \geq 1$ subintervals of length no more than h_1, i.e.,

$$[0, h_1), [h_1, 2h_1), \cdots, [(n_1 - 1)h_1, T].$$

Let $\Phi(n_1, n_2)$ be the set of all functions that are constant on these subintervals, nondecreasing, and taking values only in the set $\{(j - 1/2)h_2 \,|\, j = 1, \cdots, n_2\}$ where $h_2 = V/n_2$. It is easy to see that the cardinality of $\Phi(n_1, n_2)$, denoted $|\Phi(n_1, n_2)|$, satisfies

$$|\Phi(n_1, n_2)| = \binom{n_1 + n_2}{n_1} < 2^{n_1 + n_2}.$$

Next, note that for any $f \in \mathcal{I}$ there exists a $\phi \in \Phi(n_1, n_2)$ such that

$$\|f - \phi\|_{\mathcal{L}_1} \leq h_1 V + h_2 T/2.$$

(To see this, consider the error of the best constant approximation to f on a subinterval, and the additional error introduced by quantizing the range.) Choosing $h_1 = \epsilon/2V$, $h_2 = \epsilon/T$, $n_1 = 2\lceil VT/\epsilon \rceil$, and $n_2 = \lceil VT/\epsilon \rceil$ gives $\|f - \phi\|_{\mathcal{L}_1} \leq \epsilon$. Hence, with this choice we see that

$$\mathcal{N}(\epsilon, \mathcal{I}, \mathcal{L}_1) \leq |\Phi(n_1, n_2)| \leq 2^{n_1 + n_2} = 2^{3\lceil VT/\epsilon \rceil}$$

for $\epsilon < VT$. We then have from (1) and the fact that $\epsilon \leq VT/12$ that

$$\mathcal{N}(\epsilon, \mathcal{F}_1, \mathcal{L}_1) \leq (\mathcal{N}(\epsilon/2, \mathcal{I}, \mathcal{L}_1))^2 \leq 2^{12\lceil VT/\epsilon \rceil} \leq 2^{13VT/\epsilon}.$$

Lower Bound: Partition $[0, T]$ into n segments. Take the set Γ of all binary $\{0, h\}$-valued functions constant over each segment. If $n \geq 2$, the bounded variation and boundedness constraints impose that

$$h \leq V/n. \tag{2}$$

There are 2^n functions in this set. For any two functions $\gamma_i, \gamma_j \in \Gamma$ define $d(i, j)$ as the number of segments on which the two functions have different values. It is then easy to see that

$$\|\gamma_i - \gamma_j\|_{\mathcal{L}_1} = d(i, j)hT/n.$$

Hence, $\gamma_i, \gamma_j \in \Gamma$ are ϵ-close iff $d(i, j) \leq \epsilon n/hT$. For an arbitrary $\gamma \in \Gamma$, let $C(\epsilon)$ be the number of functions in Γ that are ϵ-close to γ. Then

$$C(\epsilon) = \sum_{l=0}^{\lfloor \epsilon n/hT \rfloor} \binom{n}{l} \tag{3}$$

where $\lfloor \alpha \rfloor$ denotes the largest integer no bigger than α. The Chernoff–Okamoto inequality (see, e.g., [6]) is

$$\sum_{l=0}^{m} \binom{n}{l} p^l (1-p)^{n-l} \leq e^{-(np-m)^2/[2np(1-p)]}$$

for $p \leq 1/2$ and $m \leq np$. Letting $p = 1/2$ we get that

$$\sum_{l=0}^{\lfloor \epsilon n/hT \rfloor} \binom{n}{l} \leq 2^n e^{-2(n/2 - \lfloor \epsilon n/hT \rfloor)^2/n} \leq 2^n e^{-(n/2)(1-2\epsilon/hT)^2}.$$

(The same result can be obtained by using Hoeffding's inequality.) Since the cardinality of Γ is 2^n, it is clear that we need at least $2^n/C(\epsilon)$ functions for an ϵ-cover, i.e.,

$$\mathcal{N}(\epsilon, \mathcal{F}_1, \mathcal{L}_1) \geq 2^n/C(\epsilon) \geq e^{(n/2)(1-2\epsilon/hT)^2}. \tag{4}$$

To obtain a large lower bound we want to maximize this expression subject to (2). Equivalently

$$\max_{n \geq 2,\, h \leq V/n} n\left(1 - \frac{2\epsilon}{hT}\right)^2 \geq \max_{n \geq 2} n\left(1 - \frac{2\epsilon n}{VT}\right)^2.$$

If $\epsilon \leq VT/12$, we may choose $n = \lfloor VT/(6\epsilon) \rfloor$ to show that this maximum is at least

$$\frac{2VT}{27\epsilon} - \frac{4}{9} \geq \frac{VT}{27\epsilon}.$$

Hence from (4), a lower bound on the covering number is

$$\mathcal{N}(\epsilon, \mathcal{F}_1, \mathcal{L}_1) \geq e^{VT/54\epsilon}, \qquad \forall \epsilon \leq VT/12. \qquad \square$$

III. General Function Classes

In this section, we give upper and lower bounds on the quantity

$$\max_P \mathcal{N}(\epsilon, \mathcal{F}, \mathcal{L}_1(dP))$$

for rather general classes $\mathcal{F}$ of $[0, 1]$-valued functions defined on a set X, where the max is taken over all probability distributions P on X. These bounds are given in terms of the following scale-sensitive dimension of $\mathcal{F}$ (see, for example, [1]). Define

$$\operatorname{fat}_{\mathcal{F}}(\epsilon) = \max\{n\text{: some } x \in X^n \text{ is } \epsilon\text{-shattered by } \mathcal{F}\}$$

where a sequence $x \in X^n$ is ϵ-shattered by $\mathcal{F}$ if there is a sequence $r \in [0, 1]^n$ such that, for all $b \in \{0, 1\}^n$, there is an $f \in \mathcal{F}$ with

$$f(x_i) \begin{cases} \geq r_i + \epsilon, & \text{if } b_i = 1 \\ \leq r_i - \epsilon, & \text{otherwise.} \end{cases}$$

For example, it is easy to show that bounded variation functions $\mathcal{F}_1$ have $\operatorname{fat}_{\mathcal{F}_1}(\epsilon) = \lfloor V/2\epsilon \rfloor$. (To see the upper bound, consider an ϵ-shattered sequence and, for a suitable sequence r, find the sequence b for which the corresponding function has maximum variation. An easy construction with $r = 0$ gives the lower bound.)

Theorem 2: There are constants c_1 and c_2 such that, for any permissible[1] class $\mathcal{F}$ of $[0, 1]$-valued functions defined on a set X

$$\operatorname{fat}_{\mathcal{F}}(4\epsilon)/32 \leq \max_P \log_2 \mathcal{N}(\epsilon, \mathcal{F}, \mathcal{L}_1(dP)) \leq c_1 \operatorname{fat}_{\mathcal{F}}(c_2\epsilon)[\log_2(1/\epsilon)]^2.$$

Comments:

- There is a gap between the upper and lower bounds. The class $\mathcal{F}_1$ of bounded variation functions shows that the lower bound is tight within a constant factor. The following example shows that in general some gap between the upper and lower bounds is essential. For any positive integers d, n, let $\mathcal{F}_{d,n}$ be the class of all functions from $\{1, 2, \cdots, d\}$ to $\{0, 1/n, \cdots, 1\}$ and let P be the uniform distribution on $\{1, 2, \cdots, d\}$. Then $\operatorname{fat}_{\mathcal{F}_{d,n}}(\epsilon) = d$ for any $\epsilon \leq 1/2$, and yet for $\epsilon = 1/(2nd)$ we have
$$\log_2 \mathcal{N}(\epsilon, \mathcal{F}_{d,n}, \mathcal{L}_1(dP)) > d \log_2(1/(2d\epsilon)).$$
- Clearly, Theorem 2 can trivially be extended to classes of functions that map to an arbitrary interval $[a, b]$, by scaling ϵ by a factor $(b - a)$.
- Recently, Cesa-Bianchi and Haussler [5] have proved a related result in a considerably more general setting. Their result gives bounds on covering numbers of a class of functions that take values in an arbitrary totally bounded metric space, in terms of a scale-sensitive dimension of the class. In the special case of real-valued functions, their scale-sensitive dimension is different from that considered here, although [2, Lemmas 8 and 9] show that the two quantities are within log factors of each other.

To prove the upper bound we need three lemmas. The first shows that a bound on $\operatorname{fat}_{\mathcal{F}}$ implies that, for any finite sequence $x = (x_1, \cdots, x_m)$ from X, there is a small subset of $\mathcal{F}$ that is a cover of the restriction of $\mathcal{F}$ to x, denoted

$$\mathcal{F}_{|x} = \{(f(x_1), \cdots, f(x_m))\text{: } f \in \mathcal{F}\} \subseteq \Re^m.$$

(In defining the cover and covering numbers $\mathcal{N}(\epsilon, \mathcal{F}_{|x})$, we use the scaled ℓ_1 metric on $\Re^m$ defined by

$$\rho(a, b) = (1/m) \sum_{i=1}^{m} |a_i - b_i|.$$

We can also consider $\mathcal{F}_{|x}$ as a class of functions mapping from $\{1, \cdots, m\}$ to $\Re$; this is how we define $\operatorname{fat}_{F_{|x}}$.) The second lemma shows that this implies there is a small cover for the set of absolute differences between functions in $\mathcal{F}$. The third lemma shows that this implies a uniform convergence result for this set. We can use this result to show that, for some set of sequences of positive probability under P, the cover for the restriction of $\mathcal{F}$ to the sequence induces a cover for $\mathcal{F}$.

The first lemma is implicit in [3, proof of Theorem 9]. (This gives a slightly better bound—by a factor of $\log d$—than that given in [1].)

Lemma 1: Suppose $x \in X^m$ and $\mathcal{F}$ is a set of $[0, 1]$-valued functions defined on X. Let $d = \operatorname{fat}_{\mathcal{F}_{|x}}(\epsilon/4)$, where $\epsilon > 0$. If $n \geq 2d \log_2(64e^2/(\epsilon \ln 2))$, there is a subset T of $\mathcal{F}$ for which $T_{|x}$ is an ϵ-cover of $\mathcal{F}_{|x}$, and

$$|T| \leq 2\left(\frac{16}{\epsilon}\right)^{6d \log_2(32en/(d\epsilon))}.$$

[1] This is a benign measurability condition—see, for example, [16].

Lemma 2: For a class $\mathcal{F}$ of $[0, 1]$-valued functions, let

$$|\mathcal{F} - \mathcal{F}| = \{|f_1 - f_2|: f_1, f_2 \in \mathcal{F}\}.$$

Then, with the metric $\mathcal{L}_1(dP)$ on $\mathcal{F}$

$$\mathcal{N}(\epsilon, |\mathcal{F} - \mathcal{F}|, \mathcal{L}_1(dP)) \leq \mathcal{N}(\epsilon/2, \mathcal{F}, \mathcal{L}_1(dP))^2.$$

Proof: Take an $\epsilon/2$-cover T for $\mathcal{F}$. Then for all $f_1, f_2 \in \mathcal{F}$, pick $t_1, t_2 \in T$ within $\epsilon/2$ of f_1 and f_2, respectively. We have

$$\| |f_1 - f_2| - |t_1 - t_2| \|_{\mathcal{L}_1(dP)} \leq \| |f_1 - t_1| + |f_2 - t_2| \|_{\mathcal{L}_1(dP)} \leq \epsilon.$$

It follows that $\{|t_1 - t_2|: t_1, t_2 \in T\}$ is an ϵ-cover for $\mathcal{F}$. □

The third lemma, which gives a uniform convergence property, is due to Pollard [16]. (Haussler [8] has a related result with improved constants.) The class $\mathcal{G}$ that we will consider is $|\mathcal{F} - \mathcal{F}|$.

Lemma 3: For a permissible class $\mathcal{G}$ of $[0, 1]$-valued functions defined on a set X, a probability distribution P on X, and $\epsilon > 0$

$$P^m\left\{x \in X^m: \exists g \in \mathcal{G}, \left|\frac{1}{m}\sum_{i=1}^m g(x_i) - Eg\right| > \epsilon\right\} \leq 4 \max_{x \in X^m} \mathcal{N}(\epsilon/16, \mathcal{G}_{|x}) e^{-\epsilon^2 m/128}.$$

Proof of Theorem 2:

Upper Bound: We start by establishing the following chain of inequalities:

$$P^m\left\{x \in X^m: \exists f_1, f_2 \in \mathcal{F}, \left|\frac{1}{m}\sum_{i=1}^m |f_1(x_i) - f_2(x_i)| - E|f_1 - f_2|\right| > \epsilon/2\right\} \quad (5)$$

$$\leq 4 \max_{x \in X^m} \mathcal{N}(\epsilon/32, |\mathcal{F} - \mathcal{F}|_{|x}) e^{-\epsilon^2 m/(128 \cdot 4)} \quad (6)$$

$$\leq 4 \max_{x \in X^m} \mathcal{N}(\epsilon/64, \mathcal{F}_{|x})^2 e^{-\epsilon^2 m/(128 \cdot 4)} \quad (7)$$

$$\leq 16\left(\frac{16 \cdot 64}{\epsilon}\right)^{12d \log_2 (32 \cdot 64 e m/(d\epsilon))} e^{-\epsilon^2 m/(128 \cdot 4)} \quad (8)$$

where (6) holds by Lemma 3, (7) holds by Lemma 2, and (8) holds by Lemma 1 for $d = \operatorname{fat}_{\mathcal{F}}(\epsilon/(64 \cdot 4))$ and

$$m \geq 2d \log_2 (64^2 e^2/(\epsilon \ln 2)).$$

It is easy to show that the probability in (5) is less than 1 for $m \geq k_1 d/\epsilon^3$, where k_1 is a constant. (Actually, a more tedious calculation shows that $m \geq (k_1/\epsilon^2 d) \log^2 (1/\epsilon)$ will suffice.) In this case, it follows that there is an $x \in X^m$ such that any $T \subseteq \mathcal{F}$ for which $T_{|x}$ is an $(\epsilon/2)$-cover for $\mathcal{F}_{|x}$ is an ϵ-cover for $\mathcal{F}$. To see this, notice that for all $f \in \mathcal{F}$ there is a $t \in T$ with

$$1/m \sum_{i=1}^m |t(x_i) - f(x_i)| \leq \epsilon/2$$

and if x is chosen in the complement of the set described in (5), then

$$\left|1/m \sum_{i=1}^m |t(x_i) - f(x_i)| - E|t - f|\right| \leq \epsilon/2$$

so $E|t - f| \leq \epsilon$. (In fact, for sufficiently large m, a proper cover for the restriction of $\mathcal{F}$ to almost any m-sequence induces a cover for $\mathcal{F}$.) Together with Lemma 1, this shows that some $T \subseteq \mathcal{F}$ satisfies

$$\log_2 \mathcal{N}(\epsilon, \mathcal{F}, \mathcal{L}_1(dP)) \leq \log_2 |T| \leq 1 + c_1 \operatorname{fat}_{\mathcal{F}}(c_2\epsilon)\left[\log_2 \frac{1}{\epsilon}\right]^2$$

from which the result follows.

Lower Bound: Suppose $\operatorname{fat}_{\mathcal{F}}(4\epsilon) \geq d$. Then consider the uniform distribution on a 4ϵ-shattered set of size d, and consider the restriction of the class $\mathcal{F}$ to this set. The same argument as the proof of the lower bound in Theorem 1, gives

$$\mathcal{N}(\epsilon, \mathcal{F}, \mathcal{L}_1(dP)) \geq \exp(d/32).$$ □

ACKNOWLEDGMENT

The authors wish to thank two anonymous reviewers for providing helpful comments and suggestions.

REFERENCES

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, "Scale-sensitive dimensions, uniform convergence, and learnability," in *Proc. ACM Symp. on Foundations of Computer Science*, 1993.

[2] M. Anthony and P. L. Bartlett, "Function learning from interpolation," in *Computational Learning Theory: Proc. 2nd European Conf., EuroCOLT'95*, 1995, pp. 211–221.

[3] P. L. Bartlett and P. Long, "More theorems about scale-sensitive dimensions and learnability," in *Proc. 8th Annu. Conf. on Computational Learning Theory*,, 1995, pp. 392–401.

[4] L. Birgé, "Estimating a density under order restrictions: Nonasymptotic minimax risk," *Ann. Statist.*, vol. 15, pp. 995–1012, 1987.

[5] N. Cesa-Bianchi and D. Haussler, "A graph-theoretic generalization of the Sauer–Shelah lemma," DSI, Universitá di Milano, Milano. Italy, Internal Rep. 170 96, 1996.

[6] R. M. Dudley, "Central limit theorems for empirical measures," *Ann. Probab.*, vol. 6, no. 6, pp. 899–929, 1978.

[7] P. Groeneboom, "Some current developments in density estimation," in *CWI Monographs.* Amsterdam. The Netherlands: North-Holland, 1986.

[8] D. Haussler, "Decision theoretic generalizations of the PAC model for neural net and other learning applications," *Inform. Comput.*, vol. 100, pp. 78–150, 1992.

[9] ——, "Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik–Chervonenkis dimension," *J. Comb. Theory*, ser. A, vol. 69, no. 2, p. 217, 1995.

[10] A. N. Kolmogorov, "Asymptotic characteristics of some completely bounded metric spaces," *Dokl. Akad. Nauk. SSSR*, vol. 108, pp. 585–589, 1956.

[11] A. N. Kolmogorov and S. V. Fomin, *Introductory Real Analysis.* New York: Dover, 1970.

[12] A. N. Kolmogorov and V. M. Tihomirov, "ϵ-entropy and ϵ-capacity of sets in function spaces," *Amer. Math. Soc. Transl. (2)*, vol. 17, pp. 277–364, 1961.

[13] S. R. Kulkarni, S. K. Mitter, and J. N. Tsitsiklis, "Active learning using arbitrary binary-valued queries," *Mach. Learning*, vol. 11, pp. 23–35, 1993.

[14] W. S. Lee, P. L. Bartlett, and R. C. Williamson, "On efficient agnostic learning of linear combinations of basis functions," in *Proc. 8th Annu. Conf. on Computational Learning Theory*, 1995, pp. 369–376.

[15] G. G. Lorentz, "Metric entropy and approximation," *Bull. Amer. Math. Soc.*, vol. 72, pp. 903–937, 1966.

[16] D. Pollard, *Convergence of Stochastic Processes.* New York: Springer, 1984.