JOURNAL OF COMBINATORIAL THEORY, Series A **69**, 217–232 (1995)

Sphere Packing Numbers for Subsets of the Boolean n-Cube with Bounded Vapnik–Chervonenkis Dimension

DAVID HAUSSLER*,†

Department of Computer and Information Sciences, University of California, Santa Cruz, California 95064 and Mathematical Sciences Research Institute, Berkeley, California

Communicated by the Managing Editors

Received November 16, 1991

Let $V \subseteq \{0,1\}^n$ have Vapnik–Chervonenkis dimension d. Let $\mathcal{M}(k/n, V)$ denote the cardinality of the largest $W \subseteq V$ such that any two distinct vectors in W differ on at least k indices. We show that $\mathcal{M}(k/n, V) \leq (cn/(k+d))^d$ for some constant c. This improves on the previous best result of $((cn/k)\log(n/k))^d$. This new result has applications in the theory of empirical processes. © 1995 Academic Press, Inc.

1. STATEMENT OF RESULTS

Let n be a natural number greater than zero. Let $V \subseteq \{0,1\}^n$. For a sequence of indices $I = (i_1, \ldots, i_k)$, with $1 \leq i_j \leq n$, let $V_{|_I}$ denote the projection of V onto I, i.e.,

$$V_{|_I} = \{(v_{i_1}, \ldots, v_{i_k}) : (v_1, \ldots, v_n) \in V\}.$$

If $V_{|_I} = \{0,1\}^k$ then we say that V *shatters* the index sequence I. The *Vapnik–Chervonenkis dimension* of V is the size of the longest index sequence I that is shattered by V [VC71] (this terminology comes from [HW87]). We will denote this number by d. Hence

$$d = \max\left\{k : \exists I = (i_1, \ldots, i_k), 1 \leq i_j \leq n, \text{with } V_{|_I} = \{0,1\}^k\right\}.$$

This quantity plays a important role in certain areas of statistics, in

*The author gratefully acknowledges the support of ONR Grant N00014-91-J-1162 and the Mathematical Sciences Research Institute at UC Berkeley, supported under NSF Grant NSF-DMS 8505550.

†E-mail: haussler@cse.ucsc.edu.

0097-3165/95 $6.00

particular in the theory of empirical processes [Dud78, Vap82, GZ84, Dud84, Pol84, Tal87a, Tal87b, Tal88, Pol90]. It has also been used recently in the fields of computational geometry [HW87, Wel88, MSW90, EGS88, CF88, CW89] and machine learning [BEHW89, HP88, RHW89, FC90, VW91].

Let $|V|$ denote the cardinality of V. The following result is well known, and was independently discovered by several people, including Sauer [Sau72] and Vapnik and Chervonenkis (see [Ass83] for a review, and also [Dud84]).

LEMMA 1 (Sauer/VC). *If the Vapnik–Chervonenkis dimension of V is d, then*

$$|V| \leq \sum_{i=0}^{d} \binom{n}{i} \leq (en/d)^d,$$

where e is the base of the natural logarithm.

For vectors $\mathbf{u}, \mathbf{v} \in \{0,1\}^n$, let

$$\rho(\mathbf{u}, \mathbf{v}) = \frac{1}{n} \sum_{i=1}^{n} |u_i - v_i|.$$

For any $\varepsilon > 0$, a set of vectors $W \subset \{0,1\}^n$ is *ε-separated* if for all distinct $\mathbf{u}, \mathbf{v} \in W$, $\rho(\mathbf{u}, \mathbf{v}) \geq \varepsilon$. The *$\varepsilon$ packing number* for a set $V \subseteq \{0,1\}^n$, denoted $\mathcal{M}(\varepsilon, V)$, is the cardinality of the largest ε-separated subset W of V. Thus for integer r, $\mathcal{M}((2r+1)/n, V)$ is the cardinality of the largest set of disjoint L_1 balls of radius r/n with centers in V, or equivalently, the size of the largest r-bit error correcting code contained in V. In this paper we demonstrate the following result.

THEOREM 1. *If the Vapnik–Chervonenkis dimension of V is d and $\varepsilon = k/n$ for integer k, $1 \leq k \leq n$, then*

$$\mathcal{M}(\varepsilon, V) \leq e(d+1)\left(\frac{2e(n+1)}{k+2d+2}\right)^d \leq e(d+1)\left(\frac{2e}{\varepsilon}\right)^d.$$

This shows that the L_1 sphere packing numbers for arbitrary sets with Vapnik–Chervonenkis dimension d behave something like L_1 sphere packing numbers for bounded regions of d-dimensional Euclidean space.

Note that for $k = 1$ (i.e., $\varepsilon = 1/n$), any two distinct vectors in V are ε-separated, thus the first bound gives a result similar to the Sauer/VC bound (Lemma 1) in this case, although not as tight. However, for larger

values of ε, Theorem 1 improves on the best previous result, which is

$$\mathcal{M}(\varepsilon, V) \leq \left(\frac{c_0}{\varepsilon} \log \frac{1}{\varepsilon}\right)^d,$$

where c_0 is some constant, obtained using the method of Dudley [Dud78] (see [Hau92] for a bound on the constants in Dudley's result). By eliminating of the extra log factor in Dudley's result, certain key bounds in the theory of empirical processes can also be improved by a logarithmic factor [Tal94] (see Remark 2 at the end of this paper). This result also has applications in geometry [W92] and set discrepancy [M94].

It is likely that the constant $2e$ in our result can be further improved. (It certainly can be improved for small d by using more precise upper estimates of $\sum_{i=0}^{d}\binom{n}{i}$ than that given in Lemma 1.) However, to within some multiplicative constant, the general form of the first bound of Theorem 1 is as tight as possible. This follows from the following lower bound.

THEOREM 2. *For all natural numbers d, $s \geq 1$ there exists a subset $V \subset \{0,1\}^n$, where $n = sd$, with Vapnik–Chervonenkis dimension d such that for each k, $1 \leq k \leq n$,*

$$\mathcal{M}(k/n, V) \geq \left(\frac{n}{2e(k+d)}\right)^d.$$

This leaves a gap from $1/2e$ to $2e$ for the best universal value of the key constant in the bound of Theorem 1. Again, it is likely that the lower bound of Theorem 2 can be improved as well. However, at this time we do not have a good guess as to what the best possible constant is. This remains an intriguing open problem.

In remarks at the end of this paper we consider some consequences of Theorem 1 for some more general kinds of packing numbers associated with the Vapnik–Chervonenkis dimension and the *pseudodimension*, a related notion used in the theory of empirical processes [Pol90, Hau92]. However, it remains open whether similar results hold for some of the other generalizations of the Vapnik–Chervonenkis dimension and the Sauer/VC lemma that have been studied (e.g., [Ste78, Fra83, Alo83, Dud87, HL91, BDCBL92]).

2. PROOFS OF THE RESULTS

Throughout this section we assume that $V \subseteq \{0,1\}^n$ and the Vapnik–Chervonenkis dimension of V is d. We begin with the following simple lemma from [HLW90].

Let E be the set of all pairs $(\mathbf{u}, \mathbf{v})$ with $\mathbf{u}, \mathbf{v} \in V$ such that $\rho(\mathbf{u}, \mathbf{v}) = 1/n$. Thus E is the set of edges in the subgraph of the Boolean n-cube induced by V (see also [Bon72, AHW87]).

LEMMA 2. [HLW90]. $|E|/|V| \leq d$.

Although this result is already proved in [HLW90], for completeness we provide an alternate proof here. This proof was suggested to us by Nati Linial, and uses the simple technique of *shifting* [Fra87, Ste78, Alo83, Tal88] in place of the recursion in [HLW90].

Proof. For each index i, $1 \leq i \leq n$, and each $\mathbf{v} \in V$, if $v_i = 1$ and the vector $\mathbf{v}' = (v_1, \ldots, v_{i-1}, 0, v_{i+1}, \ldots, v_n)$ is not in V, then let $S_{i,V}(\mathbf{v}) = \mathbf{v}'$ (here we say that $\mathbf{v}$ is *shifted* to $\mathbf{v}'$), otherwise let $S_{i,V}(\mathbf{v}) = \mathbf{v}$. We define the *shift* of V on index i, denoted $S_i(V)$, by

$$S_i(V) = \{S_{i,V}(\mathbf{v}) : \mathbf{v} \in V\}.$$

Let $S_i(E)$ denote the set of edges in the subgraph of the n-cube induced by $S_i(V)$. We claim that

(1) $|S_i(V)| = |V|$,

(2) $|S_i(E)| \geq |E|$, and

(3) for any index set I, if I is shattered by $S_i(V)$ then I is shattered by V. Hence the Vapnik–Chervonenkis dimension of $S_i(V)$ is no more than that of V [Alo83].

The first claim is obvious. To verify the second claim, we map the edges of E in a 1–1 manner into the edges of $S_i(E)$. Assume $(\mathbf{u}, \mathbf{v}) \in E$. If neither $\mathbf{u}$ nor $\mathbf{v}$ are shifted then this edge is unaffected by the shift, so map it to itself. If both $\mathbf{u}$ and $\mathbf{v}$ are shifted then this edge is simply mapped to the edge $(S_{i,V}(\mathbf{u}), S_{i,V}(\mathbf{v}))$. Finally, let us assume that $\mathbf{v}$ is shifted, but $\mathbf{u}$ is not. In this case $\mathbf{u}$ and $\mathbf{v}$ must differ on some index $j \neq i$, and we must have $u_i = v_i = 1$. Since $\mathbf{u}$ is not shifted, $\mathbf{u}' = (u_1, \ldots, u_{i-1}, 0, u_{i+1}, \ldots, u_n) \in V$. It follows that $(\mathbf{u}', S_{i,V}(\mathbf{v})) \in S_i(E)$. Hence we can map $(\mathbf{u}, \mathbf{v})$ to $(\mathbf{u}', S_{i,V}(\mathbf{v}))$. It is easily verified that the resulting map is 1–1.

To verify the third claim, suppose that a sequence I of k indices is shattered by $S_i(V)$. If i is not in I, then clearly I is also shattered by V, since $V_{|I} = S_i(V)_{|I}$ in this case. So let us assume that i is in I. Without loss of generality, we may assume that $i = 1$ and $I = (1, \ldots, k)$. Since I is shattered by $S_i(V)$, for every $\mathbf{u} \in \{0, 1\}^k$ there is a $\mathbf{v} \in S_i(V)$ with $v_j = u_j$, $1 \leq j \leq k$. However, if $u_1 = 1$ then we must have $\mathbf{v}$ and $\mathbf{v}' = (0, v_2, \ldots, v_n)$ both in V, otherwise $\mathbf{v}$ would have been shifted, and hence not be in $S_i(V)$. This implies that I is shattered by V, establishing the last claim.

Now beginning with V, simply shift V repeatedly on any sequence of (not necessarily distinct) indices until no more non-trivial shifts are possible, i.e., until you obtain a set W such that $S_i(W) = W$ for all $1 \leq i \leq n$. This must happen eventually, since each non-trivial shift reduces the total number of ones in the vectors of V. Let F be the set of edges of the subgraph of the n-cube induced by W. By the above results, $|W| = |V|$, $|F| \geq |E|$, and the Vapnik–Chervonenkis dimension of W is at most d.

Let us say that $\mathbf{u} \leq \mathbf{v}$ if $u_i \leq v_i$ for all i, $1 \leq i \leq n$. We claim that W is closed downward under the ordering $\leq$, in the sense that if $\mathbf{v} \in W$, then $\mathbf{u} \in W$ for all $\mathbf{u} \leq \mathbf{v}$. It is clear that if $\mathbf{u} \leq \mathbf{v} \in W$ and $\mathbf{u}$ differs from $\mathbf{v}$ on only one index i, then $\mathbf{u} \in W$: otherwise one more non-trivial shift of W would be possible. The claim follows by induction. It follows from this that if $\mathbf{v} \in W$, then the set of indices i for which $v_i = 1$ is shattered by W. Since the Vapnik–Chervonenkis dimension of W is at most d, this implies that no vector in W contains more than d ones. Therefore

$$|V| = |W| \leq \sum_{i=0}^{d} \binom{n}{i}$$

(which is the Sauer/VC lemma (Lemma 1)) and

$$|E|/|V| \leq |F|/|W| \leq d.$$

The last inequality can be verified by noting that a vector in $\{0, 1\}^n$ with at most d ones can have n-cube edges to at most d vectors with fewer ones. ∎

Lemma 2 is the key in proving the next result, which is main tool we use in the proof of Theorem 1. It is closely related to the results obtained in [HKS91]. Let P be a probability distribution on V. Hence V can now be viewed as a vector-valued random variable. For each i, $1 \leq i \leq n$, let V_i be the ith component of the random variable V. Thus $V_1, \ldots, V_n$ are correlated Bernoulli random variables, and the value of V_i is determined by choosing $\mathbf{v} \in V$ at random by the distribution P, and letting $V_i = v_i$. Recall that for any Bernoulli random variable B, the variance of B is $p(1 - p)$, where p is $P(B = 1)$, and for Bernoulli random variables $B_1, \ldots, B_m$, the *conditional variance* of B_m given $B_1, \ldots, B_{m-1}$ is defined by

$$\mathbf{Var}(B_m | B_1, \ldots, B_{m-1}) = \sum_{\mathbf{v} \in \{0,1\}^{m-1}} P(\mathbf{v}) P(B_m = 1 | \mathbf{v})(1 - P(B_m = 1 | \mathbf{v})),$$

where for $\mathbf{v} = (v_1, \ldots, v_{m-1})$, $P(\mathbf{v}) = P(B_1 = v_1, \ldots, B_{m-1} = v_{m-1})$, and $P(B_m = 1 | \mathbf{v}) = P(B_m = 1 | B_1 = v_1, \ldots, B_{m-1} = v_{m-1})$.

LEMMA 3. *For any probability distribution P on V,*

$$\sum_{i=1}^{n} \mathbf{Var}(V_i | V_1, \ldots, V_{i-1}, V_{i+1}, \ldots, V_n) \leq d.$$

Proof. Let E be the set of edges of the subgraph of the n-cube induced by V, as above. Consider any subgraph (V', E') of the graph (V, E) defined by selecting a subset V' of V and letting E' be all the edges in E between vectors in V' (i.e., induced edges). Since any subset V' of V is a set of vectors with Vapnik–Chervonenkis dimension at most d, it follows from Lemma 2 that for all subsets V', the *density* $|E'|/|V'|$ of the graph (V', E') is at most d. Using Hall's theorem, it is shown in [AT92, Lemma 3.1] that for any graph (V, E), the edges in E can be oriented such that for all $v \in V$, the outdegree of v (number of edges in E directed away from v) is at most the maximum density of any subgraph of (V, E) (see also [HLW90]). In our case this maximum density is at most d. Let us orient the edges of (V, E) such that for each vector $\mathbf{v} \in V$, the outdegree of $\mathbf{v}$, which we will denote outdeg($\mathbf{v}$), is at most d. For each edge $(\mathbf{u}, \mathbf{v}) \in E$, let tail($\mathbf{u}, \mathbf{v}$) denote the vector in the pair $\mathbf{u}, \mathbf{v}$ that the edge is directed away from. We will use the directions on the edges in E shortly.

Now let us consider $\mathbf{Var}(V_i | V_1, \ldots, V_{i-1}, V_{i+1}, \ldots, V_n)$. Partition E into $E_1, \ldots, E_n$ by letting E_i be the edges that cross the ith dimension of the n-cube, i.e., $E_i = \{(\mathbf{u}, \mathbf{v}) \in E : u_j = v_j, j \neq i\}$. It is readily verified that

$$\begin{aligned}
&\mathbf{Var}(V_i | V_1, \ldots, V_{i-1}, V_{i+1}, \ldots, V_n) \\
&\quad = \sum_{(\mathbf{u}, \mathbf{v}) \in E_i} (P(\mathbf{u}) + P(\mathbf{v})) \frac{P(\mathbf{u})}{(P(\mathbf{u}) + P(\mathbf{v}))} \frac{P(\mathbf{v})}{(P(\mathbf{u}) + P(\mathbf{v}))} \\
&\quad = \sum_{(\mathbf{u}, \mathbf{v}) \in E_i} \frac{P(\mathbf{u}) P(\mathbf{v})}{(P(\mathbf{u}) + P(\mathbf{v}))}.
\end{aligned}$$

Hence

$$\sum_{i=1}^{n} \mathbf{Var}(V_i | V_1, \ldots, V_{i-1}, V_{i+1}, \ldots, V_n) = \sum_{(\mathbf{u}, \mathbf{v}) \in E} \frac{P(\mathbf{u}) P(\mathbf{v})}{(P(\mathbf{u}) + P(\mathbf{v}))}.$$

Now note that for any $x, y \geq 0$, $xy \leq (x + y)\min(x, y)$. Hence

$$\begin{aligned}\sum_{i=1}^{n} \mathbf{Var}(V_i | V_1, \ldots, V_{i-1}, V_{i+1}, \ldots, V_n) &\leq \sum_{(\mathbf{u},\mathbf{v}) \in E} \min(P(\mathbf{u}), P(\mathbf{v})) \\ &\leq \sum_{(\mathbf{u},\mathbf{v}) \in E} P(\text{tail}(\mathbf{u}, \mathbf{v})) \\ &= \sum_{\mathbf{v} \in V} P(\mathbf{v}) \text{outdeg}(\mathbf{v}) \\ &\leq d \sum_{\mathbf{v} \in V} P(\mathbf{v}) \\ &= d. \quad \blacksquare\end{aligned}$$

The final lemma we will need in order to prove Theorem 1 is the following.

LEMMA 4. *Suppose that V is an ε-separated subset of $\{0,1\}^n$. Let P be the uniform distribution on V. For any integer m, $1 \leq m \leq n$, fix a sequence $I = (i_1, \ldots, i_{m-1})$ of $m - 1$ distinct indices between 1 and n and draw index i_m uniformly at random from the remaining $n - m + 1$ indices. Then*

$$\mathbf{E}\left[\mathbf{Var}\left(V_{i_m} | V_{i_1}, \ldots, V_{i_{m-1}}\right)\right] \geq \frac{\varepsilon n}{2(n - m + 1)}\left(1 - \frac{|V_{|I}|}{|V|}\right),$$

where $\mathbf{E}$ *denotes expectation over the random choice of* i_m.

Proof. Let us consider two vectors in V to be equivalent if they have the same value on all of the indices $i_1, \ldots, i_{m-1}$ in I. Suppose that this partitions V into $|V_{|I}| = M$ equivalence classes $C_1, \ldots, C_M$. Let $N_j = |C_j|$ and $N = |V|$. Now let us focus on a single equivalence class C_j. Suppose that an additional index i_m is selected at random from the remaining $n - m + 1$ indices, and two vectors $\mathbf{u}, \mathbf{v}$ are selected uniformly at random with replacement from C_j. Since C_j is ε-separated, if $\mathbf{u} \neq \mathbf{v}$ then they differ on at least εn of the remaining $n - m + 1$ indices. Hence the probability that $u_{i_m} \neq v_{i_m}$ is at least $\varepsilon n/(n - m + 1)$ times the probability that $\mathbf{u} \neq \mathbf{v}$, or $\varepsilon n(1 - 1/N_j)/(n - m + 1)$. The variance $p(1 - p)$ of a Bernoulli random variable is just half the probability that the value of this random variable differs on two independent trials. Hence

$$\mathbf{E}\left[\mathbf{Var}\left(V_{i_m} | \mathbf{v} \in C_j\right)\right] \geq \frac{\varepsilon n}{2(n - m + 1)}\left(1 - \frac{1}{N_j}\right),$$

where the expectation is over the random choice of i_m. From the above we have

$$\begin{aligned}\mathbf{E}\left[\mathbf{Var}\left(V_{i_m}|V_{i_1},\ldots,V_{i_{m-1}}\right)\right] &= \sum_{j=1}^{M} P(C_j)\mathbf{E}\left[\mathbf{Var}\left(V_{i_m}|\mathbf{v}\in C_j\right)\right] \\ &\geq \sum_{j=1}^{M}\left(\frac{N_j}{N}\right)\frac{\varepsilon n}{2(n-m+1)}\left(1-\frac{1}{N_j}\right) \\ &= \frac{\varepsilon n}{2(n-m+1)}\left(1-\frac{M}{N}\right). \quad \blacksquare\end{aligned}$$

We can now complete the proof of Theorem 1. Without loss of generality, let us assume that V itself is ε-separated, and obtain an upper bound on $|V|$. Let P be the uniform distribution on V. Recall that $k = \varepsilon n$. We can assume that $k \geq 3$, since it can be verified that the upper bound given in the statement of the theorem is greater than the trivial upper bound from Lemma 1 when $k = 1$ or $k = 2$. Let us choose

$$m = \left\lceil \frac{(2d+2)(n+1)}{k+2d+2} \right\rceil$$

indices $i_1,\ldots,i_m$ uniformly at random without replacement[1] from $\{1,\ldots,n\}$ and look at

$$\gamma = \mathbf{E}\left[\sum_{j=1}^{m}\mathbf{Var}\left(V_{i_j}|V_{i_1},\ldots,V_{i_{j-1}},V_{i_{j+1}},\ldots,V_{i_m}\right)\right].$$

We first claim that Lemma 3 implies that $\gamma \leq d$. This can be verified by projecting V onto $I = (i_1,\ldots,i_m)$ and then defining the induced probability distribution $P_{|I}$ on $V_{|I}$ in the obvious way, i.e., $P_{|I}(u_1,\ldots,u_m) = P\{\mathbf{v}\in V: v_{i_j} = u_j,\ 1\leq j\leq m\}$. This projection does not change the conditional variances, hence the result follows.

Next we claim that

$$\begin{aligned}\gamma &= m\mathbf{E}\left[\mathbf{Var}\left(V_{i_m}|V_{i_1},\ldots,V_{i_{m-1}}\right)\right] \\ &\geq m\left(\frac{k}{2(n-m+1)}\left(1-\frac{\left|V_{|(i_1,\ldots,i_{m-1})}\right|}{|V|}\right)\right) \\ &\geq m\left(\frac{k}{2(n-m+1)}\left(1-\frac{(e(m-1)/d)^d}{|V|}\right)\right).\end{aligned}$$

[1]Since $k \geq 3$ and $n \geq d, k$, it is easy to see that $m \leq n$.

The first equality follows by symmetry (and linearity of expectation), the second from Lemma 4, and the third from the Sauer/VC lemma (Lemma 1). Now putting these two claims together, we obtain

$$d \geq m\left(\frac{k}{2(n-m+1)}\left(1-\frac{(e(m-1)/d)^d}{|V|}\right)\right)$$

or equivalently,

$$|V| \leq \frac{(e(m-1)/d)^d}{1-2d(n-m+1)/km},$$

so long as

$$\frac{2d(n-m+1)}{km} < 1.$$

Now it is clear that

$$m-1 \leq \frac{(2d+2)(n+1)}{k+2d+2},$$

so

$$\begin{aligned}(e(m-1)/d)^d &\leq \left(\left(\frac{e}{d}\right)\frac{(2d+2)(n+1)}{k+2d+2}\right)^d \\ &= \left((1+1/d)\left(\frac{2e(n+1)}{k+2d+2}\right)\right)^d \leq e\left(\frac{2e(n+1)}{k+2d+2}\right)^d.\end{aligned}$$

In addition, it is easily verified that

$$\begin{aligned}\frac{2d(n-m+1)}{km} &\leq \frac{2d(n+1-(2d+2)(n+1)/(k+2d+2))}{k(2d+2)(n+1)/(k+2d+2)} \\ &= \frac{d}{d+1}.\end{aligned}$$

Hence

$$\frac{1}{1-2d(n-m+1)/km} \leq d+1.$$

Putting these together, this gives the final bound

$$|V| \leq e(d+1)\left(\frac{2e(n+1)}{k+2d+2}\right)^d.$$

The second bound of the theorem follows easily from this one.

We close with the proof of Theorem 2.

Let $W = \{(000\ldots0), (100\ldots0), (110\ldots0), \ldots, (111\ldots1)\} \subset \{0,1\}^s$, and $V = W^d$, the set of all vectors obtained by concatenating d vectors from W. Since $n = sd$, $V \subset \{0,1\}^n$. It is easy to show that the Vapnik–Chervonenkis dimension of V is d: Say that indices $1 \leq i, j \leq n$ are equivalent if $\lceil i/s \rceil = \lceil j/s \rceil$. Then a sequence of indices is shattered by V if and only if it contains at most one index in each of the d equivalence classes. Thus no set of $d+1$ indices is shattered. Note also that the size of V is $(s+1)^d > s^d = (n/d)^d$.

For each $\mathbf{v} \in V$ and $1 \leq j \leq n$, let $N(\mathbf{v}, j)$ be the number of vectors $\mathbf{u} \in V$ with $\rho(\mathbf{u}, \mathbf{v}) = j/n$. Let $C(d, j)$ denote the number of ordered sequences of d non-negative integers that sum to j. We claim that for any $\mathbf{v} \in V$, $N(\mathbf{v}, j) \leq C(d, j)2^d$. This follows from the fact that there are at most $C(d, j)$ ways to choose the number of indices on which $\mathbf{u}$ differs from $\mathbf{v}$ in each of the d equivalence classes, and given any number of indices on which $\mathbf{u}$ and $\mathbf{v}$ must disagree in a given equivalence class, there are at most 2 choices for the values for $\mathbf{u}$ on the indices in that equivalence class. Hence, using well known identities (see, e.g., [GKP89])

$$\begin{aligned}
\sum_{j=0}^{k} N(\mathbf{v}, j) &\leq 2^d \sum_{j=0}^{k} C(d, j) \\
&= 2^d \sum_{j=0}^{k} \binom{j+d-1}{j} \\
&= 2^d \binom{k+d}{k} \\
&< 2^d (e(k+d)/d)^d \\
&= (2e(k+d)/d)^d.
\end{aligned}$$

Now choose any $\mathbf{v}_1$ in V, eliminate all vectors in V within ρ-distance k/n or less of $\mathbf{v}_1$, then choose $\mathbf{v}_2$ from the remaining vectors in V and eliminate all vectors within distance k/n of $\mathbf{v}_2$, etc., until V is exhausted.

Since we begin with more than $(n/d)^d$ vectors, and each step eliminates at most $(2e(k+d)/d)^d$ vectors, this process continues for at least

$$\frac{(n/d)^d}{(2e(k+d)/d)^d} = \left(\frac{n}{2e(k+d)}\right)^d$$

steps, and the resulting set $\mathbf{v}_1, \mathbf{v}_2, \ldots$ of vectors is clearly k/n-separated by construction. ∎

3. Remarks

3.1. *Bayes and Minimax Risk in Predicting Bits*

As argued in the proof of Theorem 1, the result given in Lemma 3 implies that if $i_1, \ldots, i_m$ are selected at random without replacement, then the expectation of $\mathbf{Var}(V_{i_m} | V_{i_1}, \ldots, V_{i_{m-1}})$ is at most d/m. When $m \gg d$, this means that the value of V_{i_m} is usually highly predictable given the values of $V_{i_1}, \ldots, V_{i_{m-1}}$. This is the basis for many of the applications of the Vapnik–Chervonenkis dimension in machine learning and statistics.

In particular, from a Bayesian perspective, we might imagine that a vector $\mathbf{v} \in V$ is selected at random according to a "prior" distribution P on V and hidden from us, indices $i_1, \ldots, i_m$ are selected uniformly at random without replacement, we are given $v_{i_1}, \ldots, v_{i_{m-1}}$, and we are asked to predict v_{i_m}. Suppose the loss is 1 if we predict incorrectly and 0 otherwise. Then Bayes optimal strategy is to compute the posterior probabilities $P(v_{i_m} = 1 | V_{i_1} = v_{i_1}, \ldots, V_{i_{m-1}} = v_{i_{m-1}})$ and $P(v_{i_m} = 0 | V_{i_1} = v_{i_1}, \ldots, V_{i_{m-1}} = v_{i_{m-1}})$, and predict according to which of these is larger. The probability that this prediction is wrong when $\mathbf{v}$ and $i_1, \ldots, i_m$ are chosen randomly as above, i.e., the Bayes risk, is the expectation of the minimum of the above two posterior probabilities. Looking into the proof of Lemma 3 and Theorem 1, it can be seen that this quantity is the same as

$$\frac{1}{m} \sum_{(\mathbf{u},\mathbf{v}) \in E} \min(P(\mathbf{u}), P(\mathbf{v})),$$

where E is the set of edges in the graph induced by projecting V onto $\{i_1, \ldots, i_m\}$. Hence, by the last series of inequalities in the proof of Lemma 3, the Bayes risk for this prediction problem is at most d/m for any prior P, as was shown in [HKS91]. As pointed out there and in [HLW90], in fact by using the orientations of the edges, we get the stronger result that there exists a (non-Bayesian) prediction strategy such that if $i_1, \ldots, i_m$ are chosen randomly without replacement, then given only the values

$v_{i_1}, \ldots, v_{i_{m-1}}$, the value v_{i_m} can be predicted such that for all $\mathbf{v} \in V$, the probability of a mistake is at most d/m, i.e., the minimax risk of this prediction problem is at most d/m.

3.2. $L_1(P)$ *Packing Numbers for Classes of* $\{0,1\}$*-Valued Functions*

For applications of the Vapnik–Chervonenkis dimension in empirical processes the following is a typical setup. Let X be a set (not necessarily finite) and $\mathscr{C}$ be a set of $\{0,1\}$-valued functions on X. For any sequence $\mathbf{x} = (x_1, \ldots, x_n)$ of points with $x_i \in X$, let

$$\mathscr{C}_{|\mathbf{x}} = \{(f(x_1), \ldots, f(x_n)) : f \in \mathscr{C}\}.$$

Hence $\mathscr{C}_{|\mathbf{x}} \subseteq \{0,1\}^n$. We can define the Vapnik–Chervonenkis dimension of $\mathscr{C}$ by

$$\dim_{VC}(\mathscr{C}) = \max_{\mathbf{x}} \dim_{VC}(\mathscr{C}_{|\mathbf{x}}),$$

where $\dim_{VC}(\mathscr{C}_{|\mathbf{x}})$ is the Vapnik–Chervonenkis dimension of $\mathscr{C}_{|\mathbf{x}}$ as defined in Section 1 above, and the maximum ranges over all sequences $\mathbf{x}$ of points of X of arbitrary finite length. When this maximum does not exist, we say that $\dim_{VC}(\mathscr{C}) = \infty$.

Let P be a probability distribution on X such that the functions in $\mathscr{C}$ are *P-measurable*. Define the pseudo-metric σ_P on $\mathscr{C}$ by letting $\sigma_P(f, g) = P\{x \in X : f(x) \neq g(x)\}$ for any $f, g \in \mathscr{C}$. For any $\varepsilon > 0$, we say a subset T of $\mathscr{C}$ is ε-separated if for all distinct $f, g \in T$, $\sigma_P(f, g) > \varepsilon$. (Note that we use strict inequality here, unlike in our previous definition of an ε-separated set of binary vectors.) For any $\varepsilon > 0$, the ε *packing number* for $\mathscr{C}$ (under $L_1(P)$ metric), denoted $\mathscr{M}(\varepsilon, \mathscr{C}, \sigma_P)$, is defined as the cardinality of the largest ε-separated subset of $\mathscr{C}$ with respect to the pseudo metric σ_P, or ∞ if there are arbitrarily large finite ε-separated subsets of $\mathscr{C}$. Using a method due to Dudley [Dud78], as a corollary of Theorem 1, we have the following result.

COROLLARY 1. *For any set X, any probability distribution P on X, any set $\mathscr{C}$ of P-measurable $\{0,1\}$-valued functions on X with $\dim_{VC}(\mathscr{C}) = d < \infty$, and any $\varepsilon > 0$,*

$$\mathscr{M}(\varepsilon, \mathscr{C}, \sigma_P) \leq e(d+1)\left(\frac{2e}{\varepsilon}\right)^d.$$

Proof. Suppose to the contrary that $\mathscr{M}(\varepsilon, \mathscr{C}, \sigma_P) > B$, where $B = e(d+1)(2e/\varepsilon)^d$. Let T be an ε-separated subset of $\mathscr{C}$ with $|T| > B$.

Thus there exists $\gamma > 0$ such that for every distinct

$$f, g \in T, P\{x \in X : f(x) \neq g(x)\} \geq \varepsilon + \gamma.$$

Draw $x_1, \ldots, x_n$ independently at random according to the distribution P on X, and for each distinct $f, g \in T$, let $A_{f,g}$ be the event that

$$\frac{1}{n} \sum_{i=1}^{n} |f(x_i) - g(x_i)| < \varepsilon.$$

By choosing n large enough, we can make the probability of the event $A_{f,g}$ less than $1/|T|^2$ for each f, g. Then, since there are less than $|T|^2$ pairs of distinct $f, g \in T$, it follows that the probability that any of the events $A_{f,g}$ occur is less than one. Hence with positive probability $(1/n)\Sigma_{i=1}^{n}|f(x_i) - g(x_i)| \geq \varepsilon$ for each pair of distinct functions $f, g \in T$. It follows that there exist $x_1, \ldots, x_n$ in X such that $(1/n)\Sigma_{i=1}^{n}|f(x_i) - g(x_i)| \geq \varepsilon$ for each pair of distinct functions $f, g \in T$. This implies that the set $V = \{(f(x_1), \ldots, f(x_n)) : f \in T\}$ is an ε-separated set of binary vectors of cardinality $|T|$, which is greater than B. However, since $\dim_{VC}(\mathscr{C}) = d$, it follows that $\dim_{VC}(V) \leq d$, and hence this contradicts Theorem 1. ▮

This result has some applications in the theory of empirical processes (see, e.g., [Tal94]).

3.3. $L_1(P)$ *Packing Numbers for Classes of Real-Valued Functions*

There is also a natural extension of Corollary 1 to classes of real-valued functions. This result, given below, is due to Phil Long. Let $\mathscr{F}$ be a class of real-valued functions on a set X and let $\Re$ denote the real numbers. For each $f \in \mathscr{F}$, the function $U_f : X \times \Re \to \{0, 1\}$ is defined by letting

$$U_f(x, r) = \begin{cases} 1 & \text{if } f(x) \geq r \\ 0 & \text{otherwise,} \end{cases}$$

for $x \in X$ and $r \in \Re$. Let $U_{\mathscr{F}} = \{U_f : f \in \mathscr{F}\}$. The *pseudodimension* of $\mathscr{F}$, denoted $\dim_P(\mathscr{F})$, can be defined by $\dim_P(\mathscr{F}) = \dim_{VC}(U_{\mathscr{F}})$ [Pol90, Hau92].

Let P be a probability distribution on X such that each $f \in \mathscr{F}$ is P-measurable. Let Q be a cumulative probability distribution function on $\Re$. We can define a pseudo metric on $\mathscr{F}$ by letting

$$\sigma_{P,Q}(f, g) = \int_X |Q(f(x)) - Q(g(x))| \, dP(x)$$

for each f, $g \in \mathcal{F}$. Let $\mathcal{M}(\varepsilon, \mathcal{F}, \sigma_{P,Q})$ denote the ε packing number for this pseudo metric, in analogy with the packing numbers $\mathcal{M}(\varepsilon, \mathcal{C}, \sigma_P)$ defined above. Using Corollary 1, we have the following result.

COROLLARY 2 (P. Long). *For any set X, any probability distribution P on X, any distribution Q on $\Re$, any set $\mathcal{F}$ of P-measurable real-valued functions on X with $\dim_P(\mathcal{F}) = d < \infty$, and any $\varepsilon > 0$,*

$$\mathcal{M}(\varepsilon, \mathcal{F}, \sigma_{P,Q}) \leq e(d+1)\left(\frac{2e}{\varepsilon}\right)^d.$$

Proof. Follows directly from Corollary 1 with $X = X \times \Re$, $\mathcal{C} = U_{\mathcal{F}}$, and $P = P \times Q$ by noting that $\sigma_P(U_f, U_g) = \sigma_{P,Q}(f, g)$. ∎

As a special case of this corollary, let the $L_1(P)$ distance between two real-valued functions f and g be defined by

$$\|f - g\|_1 = \int_X |f(x) - g(x)|\, dP(x),$$

and let $\mathcal{M}(\varepsilon, \mathcal{F}, L_1(P))$ denote the ε packing numbers for this pseudo metric.

COROLLARY 3. *For any set X, any probability distribution P on X, any set $\mathcal{F}$ of P-measurable functions on X taking values in the interval $[0, 1]$ with $\dim_P(\mathcal{F}) = d < \infty$, and any $\varepsilon > 0$,*

$$\mathcal{M}(\varepsilon, \mathcal{F}, L_1(P)) \leq e(d+1)\left(\frac{2e}{\varepsilon}\right)^d.$$

Proof. Follows directly from Corollary 2 setting Q to be the uniform distribution on $[0, 1]$. ∎

This result is also useful in the theory of empirical processes. Note that our original result, Theorem 1, can be viewed as a special case of this result in which P is a uniform distribution on a finite set of n points and the functions in $\mathcal{F}$ are $\{0, 1\}$-valued.

ACKNOWLEDGMENTS

I thank Phil Long, Michael Kearns, Michel Talagrand, Richard Dudley, and Andrew Barron for helpful discussions of this material. I also thank the referee for suggesting a simplification of Lemma 3.

REFERENCES

[AHW87] N. ALON, D. HAUSSLER, AND E. WELZL, Partitioning and geometric embedding of range spaces of finite Vapnik–Chervonenkis dimension, *in* "Proceedings 3rd Symp. on Computational Geometry," pp. 331–340, Waterloo, June 1987.

[Alo83] N. ALON, On the density of sets of vectors, *Discrete Math.* **24** (1983), 177–184.

[Ass83] PATRICE ASSOUAD, Densité et dimension, *Ann. Inst. Fourier (Grenoble)* **33**, No. 3, (1983), 233–282.

[AT92] N. ALON AND M. TARSI, Colorings and orientations of graphs, *Combinatorica* **12**, (1992) 125–134.

[BDCBL92] S. BEN-DAVID, N. CESA-BIANCHI, AND P. M. LONG, Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions, *in* "Proceedings of the 1992 Workshop on Computational Learning Theory," ACM, New York, 1992.

[BEHW89] A. BLUMER, A. EHRENFEUCHT, D. HAUSSLER, AND M. K. WARMUTH, Learnability and the Vapnik–Chervonenkis dimension, *J. Assoc.* Comput. Mach. **36**, *No.* 4 (1989), 929–965.

[Bon72] J. A. BONDY, Induced subsets, *J. Combin. Theory Ser. B* **12** (1972), 201–202.

[CF88] B. CHAZELLE AND J. FRIEDMAN, A deterministic view of random sampling and its use in geometry, *in* "Proceedings of the 29th Annual Symposium on Foundations of Computer Science," pp. 539–549 IEEE, New York, 1988.

[CW89] B. CHAZELLE AND E. WELZL, Quasi-optimal range searching and VC-dimensions. *Discrete Comput. Geom.* **4** (1989), 467–490.

[Dud78] R. M. DUDLEY, Central limit theorems for empirical measures, *Ann. Probab.* **6**, No. 6 (1978), 899–929.

[Dud84] R. M. DUDLEY, A course on empirical processes, *in* "Lecture Notes in Mathematics," Vol. 1097, pp. 2–142, Springer-Verlag, New York, 1984.

[Dud87] R. M. DUDLEY, Universal Donsker classes and metric entropy, *Ann. Probab.* **15**, No. 4 (1987), 1306–1326.

[EGS88] H. EDELSBRUNNER, L. GUIBAS, AND M. SHARIR, The complexity of many faces in arrangements of lines and of segments, *In* "Proceedings 4th Ann. ACM Symp. on Computational Geometry," pp. 44–55, ACM, New York, 1988.

[FC90] M. FULK AND J. CASE, Eds. "Proceedings of the 1990 Workshop on Computational Learning Theory," Kaufmann, San Mateo, CA, 1990.

[Fra83] P. FRANKL, On the trace of finite sets, *J. Combin. Theory Ser. A* **34** (1983) 41–45.

[Fra87] P. FRANKL, The shifting technique in extremal set theory, *in* "Surveys in Combinatorics," (*C. Whitehead, Ed.*), *pp.* 81–110, *Cambridge Univ. Press, London*, 1987.

[GKP89] R. GRAHAM, D. E. KNUTH, AND O. PATASHNIK, "Concrete Mathematics," Addison–Wesley, Reading, MA, 1989.

[GZ84] E. GINÉ AND J. ZINN, Some limit theorems for empirical processes, *Ann. Probab.* **12** (1984), 929–989.

[Hau92] D. HAUSSLER, Decision theoretic generalizations of the PAC model for neural net and other learning applications, *Inform. and Comput.* **100**, *No.* 1, (*Sept.* 1992), 78–150.

[HKS91] D. HAUSSLER, M. KEARNS, AND R. SCHAPIRE, Bounds on the sample complexity of Bayesian learning using information theory and the VC dimension, *Machine Learning* **14** No. 1, (1994), 83–114.

[HL91] D. HAUSSLER AND P. LONG, A generalization of Sauer's lemma, *J. Combin. Theory Ser. A*, to appear.

[HLW90] D. HAUSSLER, N. LITTLESTONE, AND M. WARMUTH, "Predicting {0, 1}-Functions on Randomly Drawn Points," Technical Report UCSC-CRL-90-54, University of California Santa Cruz, Computer Research Laboratory, December 1990; *Inform. and Comput.* to appear.

[HP88] D. HAUSSLER AND L. PITT, Eds., "Proceedings of the 1988 Workshop on Computational Learning Theory," Kaufmann, San Mateo, CA, 1988.

[HW87] D. HAUSSLER AND E. WELZL, Epsilon nets and simplex range queries, *Discrete Compute Geom.* **2** (1987), 127–151.

[M94] J. MATOUSEK, Tight upper bounds for the discrepancy of halfspaces, manuscript, 1994.

[MSW90] J. MATOUSEK, R. SEIDEL, AND E. WELZL, How to net a lot with a little: Small epsilon-nets for disks and halfspaces, *in* Proceedings 6th Ann. ACM Symp. on Computational Geometry," 1990.

[Pol84] D. POLLARD, "Convergence of Stochastic Processes," Springer-Verlag, New York, 1984.

[Pol90] D. POLLARD, Empirical processes: Theory and Applications, *in* "NSF-CBMS Regional Conference Series in Probability and Statistics," Vol. 2, Institute of Math. Stat. and Am. Stat. Assoc., 1990.

[RHW89] R. RIVEST, D. HAUSSLER, AND M. WARMUTH, EDS., "Proceedings of the 1989 Workshop on Computational Learning Theory," M. Kaufmann, San Mateo, CA, 1989.

[Sau72] N. SAUER, On the density of families of sets, *J. Combin. Theory Ser. A* **13** (1972), 145–147.

[Ste78] J. M. STEELE, Existence of submatrices with all possible columns, *J. Comb. Theory Ser. A* 24 (1978), 84–88.

[Tal87a] M. TALAGRAND, Donsker classes and random geometry, *Ann. Probab.* **15** (1987), 1327–1338.

[Tal87b] M. TALAGRAND, The Glivenko–Cantelli problem, *Ann. Probab.* **15** (1987), 837–870.

[Tal88] M. TALAGRAND, Donsker classes of sets, *Probab. Theory Related Fields* **78** (1988), 169–191.

[Tal94] M. TALAGRAND, Sharper bounds for Gaussian and empirical processes, *Ann. Probab.* **22**, No. 1, (1994), 28–76.

[Vap82] V. N. VAPNIK, "Estimation of Dependences Based on Empirical Data," Springer-Verlag, New York, 1982.

[VC71] V. N. VAPNIK AND A. YA. CHERVONENKIS, On the uniform convergence of relative frequencies of events to their probabilities, *Theory Probab. Appl.* **16**, No. 2 (1971), 264–80.

[VW91] L. G. VALIANT AND M. WARMUTH, EDS. "Proceedings of the 1991 Workshop on Computational Learning Theory," Kaufmann, San Mateo, CA, 1991.

[W92] L. WERNISCH, Note on stabbing numbers and sphere packing numbers, manuscript, 1992.

[Wel88] E. WELZL, Partition trees for triangle counting and other range search problems, *in* "Proceedings 4th Ann. ACM Symp. on Computational Geometry," pp. 23–33, ACM, New York 1988.