A lower bound for families of Natarajan dimension d

PAUL FISCHER†

Lehrstuhl Informatik II
Universität Dortmund
D-44221 Dortmund
Germany

JIŘÍ MATOUŠEK*†

Department of Applied Mathematics
Charles University
Malostranské nám. 25, 118 00 Praha 1
Czech Republic

Abstract

A system $\mathcal{F}$ of functions $\{1, 2, \ldots, n\} \to \{1, 2, \ldots, k\}$ has *Natarajan dimension* at most d if no $(d+1)$-element subset $A \subset X$ is *2-shattered*. A is 2-shattered if for each $x \in A$ there is a 2-element set $V_x \subseteq \{1, 2, \ldots, k\}$ such that for any choice of elements $c_x \in V_x$, a function $f \in \mathcal{F}$ exists with $f(x) = c_x$ for all $x \in A$. We improve a lower bound of $c_d k^d n^d$ (due to Haussler and Long) for the maximum size of $\mathcal{F}$ of Natarajan dimension at most d by a factor somewhat smaller than k (e.g., by $\sqrt{k}$ for $d = 1$). The problem of obtaining a tight bound is related to interesting questions in extremal graph theory.

*Research supported by Czech Republic Grant GAČR 0194 and by Charles University grants No. 193,194.

†Part of this work was done at the workshop on VC-dimension in Edinburgh in September 1996.

1 Introduction

Let X be a set and let $\mathcal{F}$ be a set of functions $f : X \to [2]$ (here and in the sequel, the notation $[k]$ for a natural number k stands for the set $\{1, 2, \ldots, k\}$). The Vapnik-Chervonenkis dimension of $\mathcal{F}$ is the maximum size of a shattered subset $A \subseteq X$, where A being shattered means that every possible function $g : A \to [2]$ is a restriction to A of some function $f \in \mathcal{F}$.

Any function $f : X \to [2]$ can be identified with a subset of X, and in the literature, the Vapnik-Chervonenkis dimension is mostly considered for set systems. It is a very useful measure of complexity of a set system; from many points of view, the set systems with a finite Vapnik-Chervonenkis-dimension are those that are "easy to handle". In statistics, systems of finite VC-dimension admit an efficient random sampling (see e.g., [12], [9]). In algorithmic learning theory, the Vapnik-Chervonenkis dimension essentially determines the number of samples needed to learn a concept (set) in a given class with a given accuracy (see [2] or [1]). Other applications include computational geometry (e.g., [5]) and discrepancy theory ([7]).

It is natural to ask what replaces the Vapnik-Chervonenkis dimension in the case of multi-valued functions. Consider a family $\mathcal{F}$ of functions $f : X \to [k]$ for an integer $k \geq 3$. Perhaps the first generalization coming to mind would be the maximum size of a subset A such that any function $g : A \to [k]$ is a restriction to A of some function $f \in \mathcal{F}$. But it is easy to see that even if this dimension is 1, the family $\mathcal{F}$ need not be "simple" in an intuitive sense, and it may have exponentially many (in $|X|$) members. A better generalization is the maximum size of a 2-shattered subset (as defined in the abstract), sometimes also called the Natarajan dimension (see [8]).

One of the key results about the Vapnik-Chervonenkis dimension is so-called Sauer's lemma (independently proved in [12], [10], and [11]), stating that the maximum possible cardinality of a family of two-valued functions on an n-point set of Vapnik-Chervonenkis dimension d is

$\sum_{i=0}^{d}\binom{n}{i}$. It is also easy to see that this bound is tight, an example being provided by the family of all functions attaining at most d values 2. If d is considered constant this number is of the order n^d, i.e. polynomial in n.

Haussler and Long [4] investigated an analogue of Sauer's lemma for the Natarajan dimension (among others), i.e. the maximum possible cardinality of a system of k-valued functions on an n-point set of Natarajan dimension d. By generalizing an inductive proof of Sauer's lemma, they obtained an upper bound of $O(k^{2d}n^d)$, where the constant of proportionality depends on d. As a lower bound example, they suggest the system of all functions attaining a value different from 1 at most d times; the number of such functions is, up to a multiplicative constant, $k^d n^d$, so a significant gap remains between the upper and lower bounds.

In this note, we observe that the problem can be re-stated in terms of hypergraphs with forbidden subhypergraphs, and using extremal hypergraph theory constructions, we obtain an improved lower bound (see Proposition 2 below). The problem of finding tight bounds appears challenging. A very interesting special case is for $n = 3$ and $d = 1$, which translates to the following extremal graph theory problem:

Problem 1 *Let X_1, X_2, and X_3 be disjoint sets of cardinality k each, and for $1 \le i < j \le 3$, let H_{ij} be a bipartite graph with vertex classes X_i and X_j containing no $K_{2,2}$ as a subgraph. What is the maximum possible number of* triangles *in the graph $H = H_{12} \cup H_{23} \cup H_{31}$?*

Since each H_{ij} can have up to roughly $k^{3/2}$ edges [6], a lower bound for this problem is of the order $k^{3/2}$. The best upper bound we can prove at present is $O(k^{7/4})$. To see this, define an auxiliary bipartite graph G with X_1 as one vertex class and the edges $E(H_{23})$ of H_{23} as the other vertex class, with a vertex $v \in X_1$ connected to an edge $e \in E(H_{23})$ if they form a triangle in H. The classes of G have sizes k and $O(k^{3/2})$ and it is easy to check that G has no $K_{2,2}$ subgraph. A well-known result in extremal graph theory implies that G has $O(k^{7/4})$ edges (this simple proof has been observed by Tomáš Kaiser; we had a more complicated argument).

For a special case we shall present a tight bound in Section 3. We exhibit a system of 3-valued functions on $[n]$ with Natarajan dimension 1 with $3n$ members, and we show that this is the maximum possible size of such a system.

2 A General Bound

Let $\mathcal{F}$ be a system of functions $f : [n] \to [k]$. Such an $\mathcal{F}$ can also be regarded as an n-regular n-partite hypergraph whose each class has k vertices. Namely, the vertex set is $[n] \times [k]$ and the edges are $\{(1, f(1)), (2, f(2)), \ldots, (n, f(n))\}$, $f \in \mathcal{F}$. In hypergraph terms, the condition that a d-element subset $A \subseteq [n]$ be 2-shattered means that the d-partite hypergraph induced in $\mathcal{F}$ by the set $A \times [k]$ contains the complete d-partite hypergraph $K_d(2, 2, \ldots, 2)$ with two vertices in each class.

Let $g_d(k)$ stand for the maximum possible number of edges of a $(d+1)$-regular $(d+1)$-partite hypergraph with classes of size k containing no $K_{d+1}(2, 2, \ldots, 2)$ as a subhypergraph. It is well-known that $g_1(k) = \Theta(k^{3/2})$; this is the case of a bipartite graph with a forbidden $K_{2,2}$. For larger d, Erdős and Simmonovits [3] proved an upper bound $g_d(k) = O(k^{d+1-1/2^d})$. A straightforward probabilistic counting gives a lower bound $g_d(k) = \Omega(k^{d+1-\delta(d)})$ with $\delta(d) = \frac{d+1}{2^{d+1}-1}$, and we haven't found a better lower bound mentioned anywhere.

Proposition 2 *For $k, n \geq 2$, there exists a family $\mathcal{F}$ of k-valued functions on $[n]$ of Natarajan dimension $\leq d$ with at least $cn^d g_d(k-1)$ members, with $c > 0$ depending on d only.*

Lemma 3 *For any constant $d \geq 2$, there exists a $(d+1)$-regular hypergraph $\mathcal{S}$ on $[n]$ with $|\mathcal{S}| = \Omega(n^d)$ whose each two edges have at most $d-1$ points in common.*

Proof. (This is probably known.) Let $\alpha > 0$ be a small constant. Choose a family of $(d+1)$-tuples at random, by picking αn^d random independent $(d+1)$-tuples. The probability that a given pair of random $(d+1)$-tuples intersect in d or $d+1$ points is at most C/n^d for a constant C. Hence the expected number of pairs with the forbidden intersection in the family is at most $\alpha^2 C n^d < \frac{\alpha}{2} n^d$. For each pair with a too large intersection, delete one of its sets. After this, at least $\frac{\alpha}{2} n^d$ $(d+1)$-tuples still remain.

A simple explicit construction (induction on d) also works. $\square$

Proof of Proposition 2. Fix a $(d+1)$-partite $(d+1)$-regular hypergraph $\mathcal{H}$ with classes of size $k-1$, containing no $K_{d+1}(2, 2, \ldots, 2)$ as a subhypergraph, and with $g_d(k-1)$ edges. Regard the edges of $\mathcal{H}$ as functions $[d+1] \to [k-1]$.

Also fix a $(d+1)$-uniform hypergraph $\mathcal{S}$ as in Lemma 3. For each edge $S \in \mathcal{S}$, choose a bijection $\varphi_S : S \to [d+1]$.

Now we can define the elements of the desired $\mathcal{F}$, i.e. functions $[n] \to [k]$. For $S \in \mathcal{S}$ and for each edge–function $H \in \mathcal{H}$, define a function $F_{S,H} \in \mathcal{F}$ by

$$F_{S,H}(i) = \begin{cases} H(\varphi_S(i)) & \text{for } i \in S \\ k & \text{for } i \notin S\,. \end{cases}$$

Since $F_{S,H}(i)$ equals k iff $i \notin S$, both S and H can be reconstructed uniquely from $F_{S,H}$, and hence $|\mathcal{F}| = |\mathcal{S}| \cdot |\mathcal{H}| = \Omega(n^d g_d(k-1))$. It remains to show that no $(d+1)$-point subset of $[n]$ is 2-shattered.

So suppose for contradiction that a $(d+1)$-point $A \subseteq [n]$ is 2-shattered. For $i \in A$, let $V_i \subseteq [k]$ be the two-element sets witnessing the 2-shattering.

We must have $A \in \mathcal{S}$, for otherwise any function of $\mathcal{F}$ attains at least one value k on A.

For $A \in \mathcal{S}$, at least one of the sets V_i with $i \in A$ contains k, because $\mathcal{H}$ has no copy of $K_{d+1}(2, 2, \ldots, 2)$. And if $k \in V_i$ then $\mathcal{F}$ has to contain a function attaining the value k precisely once on A. This is impossible, since functions of the form $F_{A,H}$ have no value k on A, and functions of the form $F_{S,H}$ for $A \neq S \in \mathcal{S}$ attain value k at least twice on A, as $|A \setminus S| \geq 2$. Proposition 2 is proved. $\Box$

3 A Tight Bound for a Special Case.

For the case $k = 3$ and $d = 1$ it is possible to find the precise size of a maximum system of functions.

Proposition 4 *For $n \geq 2$ there exists a family $\mathcal{F}$ of 3-valued functions on $[n]$ of Natarajan dimension 1 with $3n$ members. No such system can have more members.*

Proof. For the lower bound define a system by the following value matrix:

$$
\begin{array}{l|ccccccc}
 & 1 & 2 & 3 & \dots & & & n \\
\hline
f_1 & 1 & 1 & 1 & \dots & 1 & 1 & 1 \\
f_2 & 1 & 1 & 1 & \dots & 1 & 1 & 2 \\
f_3 & 1 & 1 & 1 & \dots & 1 & 2 & 2 \\
f_4 & 1 & 1 & 1 & \dots & 2 & 2 & 2 \\
\vdots & & \vdots & & & & \vdots & \\
f_n & 1 & 2 & 2 & \dots & 2 & 2 & 2 \\
f_{n+1} & 2 & 2 & 2 & \dots & 2 & 2 & 2 \\
f_{n+2} & 3 & 1 & 1 & \dots & 1 & 1 & 1 \\
f_{n+3} & 3 & 3 & 1 & \dots & 1 & 1 & 1 \\
f_{n+4} & 3 & 3 & 3 & \dots & 1 & 1 & 1 \\
\vdots & \vdots & & & & \vdots & & \\
f_{2n} & 3 & 3 & 3 & \dots & 3 & 3 & 1 \\
f_{2n+1} & 3 & 3 & 3 & \dots & 3 & 3 & 3 \\
f_{2n+2} & 2 & 3 & 3 & \dots & 3 & 3 & 3 \\
f_{2n+3} & 2 & 2 & 3 & \dots & 3 & 3 & 3 \\
\vdots & \vdots & & & & \vdots & & \\
f_{3n} & 2 & 2 & 2 & \dots & 2 & 2 & 3
\end{array}
$$

First, note that for any two columns i, j, $1 \leq i < j \leq n$, one can find rows u, v, w and values $a_i, b_i, a_j, b_j \in [3]$, $a_i \neq b_i$, $a_j \neq b_j$, such that the three pairs $(f_u(i), f_u(j))$, $(f_v(i), f_v(j))$, $(f_w(i), f_w(j))$ are all different and $f_u(i), f_v(i), f_w(i) \in \{a_i, b_i\}$ and $f_u(j), f_v(j), f_w(j) \in \{a_j, b_j\}$. It is not possible to find four rows with this property, however. Hence the function system is of Natarajan dimension 1. Let us remark that the bipartite graph induced by the considered system on each of the sets $\{i, j\} \times [3]$ is the 6-cycle, which is the only extremal $K_{2,2}$-free bipartite graph with classes of size 3.

We now show that $3n$ is also an upper bound on the size of a system of functions $f : [n] \to [3]$ of Natarajan dimension 1; the argument is the same as in [4], but since the results in that paper are formulated in a much more general way and since the explicit formula given there is slightly worse, we give a proof for reader's convenience. We proceed by induction on n. The upper bound is clear for the case $n = 1$. For $n \geq 2$, let $\mathcal{F} = \{f_1, f_2, \ldots, f_m\}$ be a system of functions $[n] \to [3]$ of Natarajan dimension 1, and consider the value matrix M of this system. Certainly, all rows of this matrix are different. Now consider the effect of deleting the last column. Let $f'_1, f'_2, \ldots, f'_m$ denote the rows of the resulting matrix M'. Some rows in M' might be equal. Let M'' be the result of deleting all but one occurrence of multiple rows. Clearly, M'' defines a system of functions $f : [n-1] \to [3]$ of Natarajan dimension 1. We want to show that M has at most 3 rows more than M''. To this end, it suffices to establish the following claim: for any 2-element set $\{a, b\} \subset [3]$, there exists at most one pair $\{f_s, f_t\}$ of rows of M with $f'_s = f'_t$ and $f_s(n) = a$, $f_t(n) = b$.

Suppose for contradiction that there are four rows in M, say f_1, f_2, f_3, and f_4, such that $f'_1 = f'_2 \neq f'_3 = f'_4$ and $f_1(n) = f_3(n) = a \neq b = f_2(n) = f_4(n)$, $a, b \in [3]$. Suppose (w.l.o.g) that f'_1 and f'_3 differ in column $n-1$, i.e. $f'_1(n-1) = f'_2(n-1) = c \neq d = f'_3(n-1) = f'_4(n-1)$. Then the top of M looks as in the following scheme:

$$
\begin{array}{c|ccccccc}
 & 1 & 2 & 3 & \ldots & & n-1 & n \\
\hline
f_1 & * & * & * & \ldots & * & c & a \\
f_2 & * & * & * & \ldots & * & c & b \\
f_3 & * & * & * & \ldots & * & d & a \\
f_4 & * & * & * & \ldots & * & d & b \\
\vdots & & \vdots & & & & \vdots &
\end{array}
$$

Clearly, the functions f_1, f_2, f_3, f_4 2-shatter the set $\{n-1, n\}$, thus contradicting the fact that the system has dimension 1. This proves the claim, and since there are 3 possible choices for $\{a, b\}$, we see that the number of rows of M is at most 3 more than that of M''. $\square$

References

[1] M. Anthony and N. Biggs. *Computational Learning Theory*. Cambridge University Press, Cambridge, 1992.

[2] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36:929–965, 1989.

[3] P. Erdős and M. Simmonovits. Supersaturated graphs and hypergraphs. *Combinatorica*, 3:181–192, 1983.

[4] D. Haussler and P. M. Long. A generalization of Sauer's lemma. *J. Comb. Theory, Ser. A*, 71:219–240, 1995.

[5] D. Haussler and E. Welzl. Epsilon-nets and simplex range queries. *Discrete Comput. Geom.*, 2:127–151, 1987.

[6] T. Kővári, V. Sós, and P. Turán. On a problem of Zarankiewicz. *Coll. Math.*, 3:50–57, 1954.

[7] J. Matoušek, E. Welzl, and L. Wernisch. Discrepancy and ε-approximations for bounded VC-dimension. *Combinatorica*, 13:455–466, 1993.

[8] B. K. Natarajan. On learning sets and functions. *Machine Learning*, 4:67–97, 1989.

[9] D. Pollard. *Empirical processes: Theory and applications.* NSF-CBMS Regional Conference Series in Probability and Statistics vol. 2. Institute of Math. Stat. and Am. Stat. Assoc., 1990.

[10] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory Ser. A*, 13:145–147, 1972.

[11] S. Shelah. A combinatorial problem, stability and order for models and theories in infinitary languages. *Pacific J. Math.*, 41:247–261, 1972.

[12] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.*, 16:264–280, 1971.