The Annals of Statistics
1989, Vol. 17, No. 4, 1441–1446

SOME NEW VAPNIK–CHERVONENKIS CLASSES

BY G. STENGLE AND J. E. YUKICH

Lehigh University

The theory of semialgebraic sets is used to generate new Vapnik–Chervonenkis (VC) classes of positivity sets. It is also shown that certain analytic families of positivity sets are VC.

1. Introduction. Given a set X, a collection $\mathscr{C}$ of subsets of X and a finite set $F \subset X$, let $\Delta^{\mathscr{C}}(F)$ be the number of different sets $C \cap F$ for $C \in \mathscr{C}$. For $n \in \mathbb{N}^+$ let $m^{\mathscr{C}}(n) := \max\{\Delta^{\mathscr{C}}(F)\colon F \text{ has } n \text{ elements}\}$. Let

$$V(\mathscr{C}) := \begin{cases} \inf\{n\colon m^{\mathscr{C}}(n) < 2^n\} \\ +\infty \quad \text{if } m^{\mathscr{C}}(n) = 2^n \text{ for all } n. \end{cases}$$

Vapnik and Chervonenkis (1971) introduced $\Delta^{\mathscr{C}}$, $m^{\mathscr{C}}(n)$ and $V(\mathscr{C})$. If $m^{\mathscr{C}}(n) < 2^n$ for some n, i.e., if $V(\mathscr{C}) < \infty$, then $\mathscr{C}$ is called a *Vapnik–Chervonenkis* class (or VC class) and $V(\mathscr{C})$ is called its index. See Dudley (1984, 1985), Pollard (1984) and Assouad (1983) for expositions and recent structure results on VC classes.

Let $\mathscr{A}$ be a σ-algebra containing $\mathscr{C}$ and P a probability measure on $\mathscr{A}$. Let X_i, $i \geq 1$, be independent identically distributed X-valued random variables with common law P. We shall consider the X_i, $i \geq 1$, to be the coordinates for a countable product $(X^\infty, \mathscr{A}^\infty, P^\infty)$ of $(X, \mathscr{A}, P)$. Let the nth empirical measure for P be defined by

$$P_n := n^{-1}\left(\delta_{X_1} + \cdots + \delta_{X_n}\right),$$

where δ_x is the unit mass at $x \in X$. If $\mathscr{C}$ is a VC class then, under some measurability conditions [Dudley (1984)], an unknown P can be approximated uniformly on $\mathscr{C}$ by the empirical measure P_n, i.e.,

$$P^\infty\left(\sup_{A \in \mathscr{C}} |(P_n - P)(A)| \to 0,\ n \to \infty\right) = 1.$$

This by now classical result belongs to Vapnik and Chervonenkis (1971).

Under stronger measurability conditions [Dudley (1978) and Dudley and Philipp (1983)], VC classes $\mathscr{C}$ satisfy the asymptotic equicontinuity condition

$$\forall\ \varepsilon > 0, \exists\ n_0 < \infty, \exists\ \delta > 0 \text{ such that } \forall\ n \geq n_0,$$

$$P^{\infty *}\left\{\sup\left\{|\nu_n(A) - \nu_n(B)|\colon A, B \in \mathscr{C},\ P(A \triangle B) < \delta\right\} > \varepsilon\right\} < \varepsilon,$$

where $\nu_n := n^{1/2}(P_n - P)$. This condition, together with total boundedness of $\mathscr{C}$ for $d_P(A, B) := P(A \triangle B)$, characterizes "functional Donsker classes" of sets, for which the central limit theorem [Dudley (1978)] holds for the normalized

Received November 1988.

AMS 1980 *subject classifications*. Primary 60B99; secondary 32C05.

Key words and phrases. Vapnik–Chervonenkis class, semialgebraic sets, semianalytic positivity sets.

empirical measures ν_n with respect to uniform convergence on $\mathscr{C}$. In this way VC classes play a natural role in multivariate nonparametric statistics.

Unfortunately, there are not many known examples of VC classes. The most important examples are usually related to Dudley's theorem [Dudley (1978), Theorem 7.2], which implies that the positivity sets of polynomials of bounded degree are VC classes. In this article we provide two nontrivial extensions of Dudley's theorem. First, we link semialgebraic sets to VC classes and show how the former may be used in a natural way to manufacture relatively "large" VC families of positivity sets. The main tool here is the quantifier elimination theorem of Tarski and Seidenberg. In the process we answer a conjecture of R. Olshen. Second, tools from the theory of analytic functions are used to show that certain special analytic families of positivity sets are VC.

2. VC classes and the Tarski–Seidenberg theorem. By *semialgebraic set* we mean a subset of R^n defined by a finite system of polynomial inequalities. More precisely, such a set is any member of the Boolean algebra generated by all sets of the form $\{x|f(x) > 0\}$, where $f(x) = f(x_1, \ldots, x_n)$ is a polynomial. Another description, equally precise, is any set in R^n described by a first order unquantified statement in the language of ordered fields. Our main tool is a simple consequence of the principle of elimination of quantifiers of Tarski and Seidenberg [Bochnak, Coste and Roy (1987), Proposition 5.2.2]. This ensures that any elementary statement about the reals in the language of ordered fields which contains free variables together with variables bound by universal or existential quantifiers is equivalent to a quantifier-free statement in the free variables alone. Roughly speaking, an elementary statement is one that does not involve quantification of bound variables over general sets. However we note that quantification over a semialgebraic set can be regarded as shorthand for an elementary statement and is therefore an allowable constituent. A simple well-known example is the equivalence of the quantified statement $\forall\, x\ (x^2 + bx + c \geq 0)$ with the quantifier-free statement $b^2 - 4c \leq 0$. This principle is very powerful. It typically relates a compact, mathematically meaningful, quantified statement to an equivalent statement without quantifiers which is unproblematic from a theoretical standpoint but which, in actual detail, is diffuse and bulky to the point of unintelligibility. (One sign of this bulk would be very large estimates for the VC index if the hypotheses and results of this paper were further quantified.) The proof of the following, our main result, is a highly characteristic application.

THEOREM 1. *Let* $P(x_1, \ldots, x_m, y_1, \ldots, y_n, t_1, \ldots, t_p, u_1, \ldots, u_q) = P(x, y, t, u)$ *be a fixed real polynomial and* T *and* U *be fixed semialgebraic subsets of* R^p *and* R^q. *Then the family of subsets of* R^m *of the form*

$$W_a := \left\{x \middle| \sup_{t\in T} \inf_{u\in U} P(x, a, t, u) > 0\right\}, \qquad a \in R^n,$$

is a VC class.

PROOF. Let $W := \{(x, y) | \sup_{t \in T} \inf_{u \in U} P(x, y, t, u) > 0\}$. We can describe W by the elementary quantified formula

$$W = \{(x, y) | \exists\, t \in T\ \forall\, u \in U, P(x, y, t, u) > 0\}.$$

By elimination of quantifiers this is equivalent to a quantifier free formula in the unbound variables x and y alone. Such a formula describes a semialgebraic set. Hence there exists a finite set of polynomials $Q_1(x, y), \ldots, Q_s(x, y)$ such that W is generated by a finite number of Boolean operations from sets of the form $\{(x, y) | Q_j(x, y) > 0\}$. The same operations with y specialized to a determines W_a. Thus the set W_a can be generated by at most a fixed number of Boolean operations from the sets $\{x | Q_j(x, a) > 0\}$. But these sets are all contained in the family of positivity sets of all polynomials in x of degree not exceeding the largest degree of any Q_j in x. Since this larger family is known to be a VC class [Dudley (1978), Theorem 7.2] and a fixed number of lattice operations preserve the VC property [Dudley (1984), Theorem 9.2.3], the family $\{W_a\}$ is contained in a VC class and hence is also VC. □

REMARK. It is easy to see that Theorem 1 remains true of P is replaced by any semialgebraic function, that is, one whose graph is a semialgebraic set.

As an application, we prove the following corollary, thus answering a conjecture of R. Olshen (private communication). This result is used by Olshen, Biden, Wyatt and Sutherland (1989) in the statistical study of gait analysis.

COROLLARY 2. *The family* $\{x | \sup_{0 \le \theta < 2\pi} \Sigma_{1 \le j \le N} P_j(x) Q_j(\theta) > 0\}$, *where the* P_j *range over any set of polynomials of bounded degree and the* Q_j *range over any set of trigonometric polynomials of bounded degree, is a VC class.*

PROOF. For each coefficient of a P or a Q, introduce a component of a new variable y. Let $t_1 = \cos\theta$ and $t_2 = \sin\theta$. Let T be the unit circle in the t-plane. Since any trigonometric polynomial $T(\theta)$ is a polynomial in $\cos\theta$ and $\sin\theta$, there is a single polynomial $P(x, y, t)$ of which any sum $\Sigma P_j(x) Q_j(\theta)$ is a specialization. All the sets in question then have the form $\{x | \sup_{t \in T} P(x, a, t) > 0\}$. By the theorem these belong to the VC class consisting of all sets of this form. □

It is clear that no estimate for the VC index can emerge from such purely existential reasoning without further quantification of our hypotheses. We remark that estimates for the complexity of semialgebraic sets (involving the number of inequalities which are needed, their degrees, etc.) which could yield bounds for the index are currently the subject of intensive investigation in semialgebraic geometry [Bröcker (1988)].

3. Analytic families of semianalytic positivity sets. It is natural to ask whether Dudley's result can be extended to certain positivity sets of real analytic functions. We shall consider analytic families of semianalytic positivity sets. Our reasoning will again be qualitative and the prospect of estimates for the index

will be even more remote. A key ingredient in our reasoning is a form of the Weierstrass preparation technique devised by Denef and van den Dries (1988) as a tool for studying subanalytic sets. The following theorem applies to certain analytic families of positivity sets on the torus, for example.

THEOREM 3. *Let X be a compact subset of a real analytic manifold M and let I denote $[-1,1]$. If $f(x,t)$ is a real analytic function on $X \times I$, then the family of subsets of X of the form*

$$W := \{\{x | f(x,t) > 0\}\}_{t \in I}$$

is a VC class.

PROOF. The underlying idea is simple. If we could apply the Weierstrass preparation theorem to f globally with respect to the variable t, then we could replace $f > 0$ by $P > 0$ where P is a distinguished polynomial

$$P(x,t) = t^N + a_1(x)t^{N-1} + \cdots + a_N(x).$$

But the positivity sets of P are positivity sets of the linear family obtained by replacing each power of t by a linear parameter which, again by Dudley's result, is VC. There are two difficulties here. The more serious problem is that we cannot assume that f is regular in t. Moreover the usual device of a slight linear change of coordinates to obtain regularity is not at our disposal since this requires, in general, transformation of (x,t). But here only transformations which respect the product structure are relevant and these are insufficient to achieve regularity. We will overcome this difficulty using methods of Denef and van den Dries to achieve regularity at the cost of increasing the dimension of M. A lesser difficulty is that such preparation is local in character. However, given our compactness hypothesis, simple properties of VC classes allow us to reduce our problem to a local one in the following way. First, if X is covered by a finite union of subsets, then a family is VC if its restriction to each subset is VC. Similarly, if I is covered by a finite number of subintervals, then it suffices to know that the subfamily parameterized by each subinterval is VC. Thus, as an example of such a reduction, since every point of $X \times I$ is contained in a product neighborhood in $M \times I$, we can cover $X \times I$ with a finite number of product neighborhoods $X_i \times I_j$ on each of which $f(x,t)$ can be represented in local coordinates by a convergent power series. By changing local coordinates, we can therefore suppose without loss of generality that $X = I^m$ and the power series for f converges on a slightly larger open cell containing $I^m \times I$. We will refer to such a reduction of the problem with respect to an open covering as "localizing and normalizing" and use it again with only brief further mention.

We first subject f to a preliminary preparation in t which does not require any regularity. By Lemma 4.12 of Denef and van den Dries (1988) there is a globally given finite set of analytic functions $a_i(x)$, $i = 1,2,\ldots,d$, such that each point (x_0, t_0) has a neighborhood (using local coordinates in which this is

the origin) on which f can be represented as a finite sum of the form

$$f(x, t) = \sum_{i \le d} a_i(x) t^i u_i(x, t),$$

where the u_i are given by convergent power series and $u_i(0, 0) \neq 0$. Localizing and normalizing it suffices to suppose that f has this form globally. We now partition X into subsets

$$X_j := \{x \mid |a_i(x)| \le |a_j(x)| \text{ for all } i\}, \qquad 1 \le j \le d.$$

Also let $V_i(x) := a_i(x)/a_j(x)$ if $x \in X_j$ and $a_j(x) \neq 0$ and $V_i(x) := 0$ otherwise. Then on X_j we have

$$f(x, t) = a_j(x)\left\{t^j + \sum_{\substack{i \le d \\ i \neq j}} V_i(x) t^i u_i(x, t)\right\}.$$

We observe that, in general, the factors here are not analytic since the V's need not even be continuous. However we can repair this by introducing one component v_i of a new variable v for each V_i. Each V_i assumes values in the interval I. The preceding function is then the restriction to the graph of $v = V(x)$ over X_j of the analytic function on $I^{m+d} \times I$:

$$f(x, v, t) = a_j(x)\left\{t^j + \sum_{\substack{1 \le i \le d \\ i \neq j}} v_i t^i u_i(x, t)\right\}.$$

Since the restriction of a VC class to a subset is obviously VC, it will suffice to show that the positivity sets of this extended function form a VC class. But at any point of the form $(0, c, t)$, $c \in I^d$, the second factor reduces to

$$t^j + \sum_{\substack{1 \le i \le d \\ i \neq j}} c_i t^i u_i(0, t)$$

which must be regular. To see this, let $c_j = 1$ and let i_0 be the least integer i for which $c_i \neq 0$. If $i_0 < j$, then this function has order i_0 at $t = 0$ and otherwise it has order j. Thus every point $(0, c, 0)$ (representing an arbitrary point of $X_j \times I^d \times I$ in original coordinates) has a neighborhood on which this second factor can be prepared, that is,

$$f(x, v, t) = a_j(x) u(x, v, t)\left\{t^k + \sum_{i < k} b_i(x, v) t^i\right\},$$

where u does not vanish and $k := \min(i_0, j)$. It follows that each point of $X_j \times I$ has a neighborhood on which the positivity set of $f(x, t)$ is given by

$$\left\{x \mid a_j(x)\left(t^k + \sum_{i < k} b_i(x, V(x)) t^i\right) > 0\right\}.$$

Localizing and normalizing we can again suppose that this form is global. But such a set belongs to the VC class of all zero sets of the linear space of functions generated by $a_j(x)$ and the $a_j(x) b_i(x, V(x))$. This concludes the proof. □

It would be interesting to know if Theorem 3 holds for the collection of positivity sets generated by a real analytic function $f(x, t)$ on $X \times I^d$ for $d > 1$. Our methods seem to yield no information here.

Acknowledgment. The authors thank Professor R. Olshen for bringing his conjecture to our attention and for suggesting the link to algebraic geometry.

REFERENCES

ASSOUAD, P. (1983). Densité et dimension. *Ann. Inst. Fourier (Grenoble)* **33** (3) 233–282.

BOCHNAK, J., COSTE, M. and ROY, M.-F. (1987). *Géométrie Algébrique Réelle*. Springer, Berlin.

BRÖCKER, L. (1988). On basic semi-algebraic sets. Report, Mathematisches Institut, Münster.

DENEF, J. and VAN DEN DRIES, L. (1988). p-adic and real subanalytic sets. *Ann. of Math.* (2) **128** 79–138.

DUDLEY, R. M. (1978). Central limit theorems for empirical measures. *Ann. Probab.* **6** 899–929; correction **7** (1979) 909–911.

DUDLEY, R. M. (1984). A course on empirical processes. *École d'Été de Probabilités de Saint-Flour, XII—1982. Lecture Notes in Math.* **1097** 1–142. Springer, Berlin.

DUDLEY, R. M. (1985). The structure of some Vapnik–Chervonenkis classes. In *Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer* (L. M. Le Cam and R. A. Olshen, eds.) **2** 495–508. Wadsworth, Monterey, Calif.

DUDLEY, R. M. and PHILIPP, W. (1983). Invariance principles for sums of Banach space valued random elements and empirical processes. *Z. Wahrsch. verw. Gebeite* **62** 509–552.

OLSHEN, R. A., BIDEN, E. N., WYATT, M. P. and SUTHERLAND, D. H. (1989). Gait analysis and the bootstrap. *Ann. Statist.* **17** 1419–1440.

POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, Berlin.

VAPNIK, V. N. and CHERVONENKIS, A. YA. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory Probab. Appl.* **16** 264–280.

DEPARTMENT OF MATHEMATICS
CHRISTMAS-SAUCON HALL #14
LEHIGH UNIVERSITY
BETHLEHEM, PENNSYLVANIA 18015