Metric entropy estimation using o-minimality theory

A. Onshuus and A. J. Quiroz

Dpto. de Matemáticas, Universidad de Los Andes. Address: Dpto. de Matemáticas, Universidad de Los Andes, Carrera 1, Nro. 18A-10, edificio H, Bogotá, Colombia. Phone: (571)3394949, ext. 2710. Fax: (571)3324427.

E-mail address: `aonshuus@uniandes.edu.co`

E-mail address: `aj.quiroz1079@uniandes.edu.co`

Abstract. It is shown how tools from the area of Model Theory, specifically from the Theory of o-minimality, can be used to prove that a class of functions is VC-subgraph (in the sense of Dudley (1987)), and therefore satisfies a uniform polynomial metric entropy bound. We give examples where the use of these methods significantly improves the existing metric entropy bounds. The methods proposed here can be applied to finite dimensional parametric families of functions without the need for the parameters to live in a compact set, as is sometimes required in theorems that produce similar entropy bounds (for instance Theorem 19.7 of van der Vaart (1998)).

1. Introduction

VC-dimension and metric entropy are fundamental concepts in modern asymptotic statistics and the theory of statistical learning, due to their applicability in establishing uniform convergence results, such as Uniform Laws of Large Numbers and Uniform Central Limit Theorems (see Dudley (1987); Pollard (1984); van der Vaart (1998); van der Vaart and Wellner (1996)). The metric entropy of a class of functions is a measure of the size of the class, and, for classes of functions, it plays a roll, with regards to asymptotics, very similar to the one played by VC-density (or VC-dimension) for collections of sets. Finding a tight upper bound for the VC-density (or VC-dimension) of a class of sets or for the metric entropy of a class of functions (when this is small enough) will be useful in establishing speed of convergence in the Uniform Law of Large Numbers, as shown in Section 2.6 of Pollard (1984) or in the Central Limit Theorem for certain functionals of the empirical process, as explained in Section 3.4 of van der Vaart and Wellner (1996).

In the present article, we will show how working with the VC-density of an associated class of sets and taking advantage of recent results from the theory of o-minimality, allows for significantly improving bounds on the metric entropy of certain classes of functions and for finding tight entropy bounds for certain classes for which other methods would not work.

Even though most of the concepts we have mentioned are by now classical, we briefly review, in the following subsection, their definitions and their use in theorems of asymptotic statistics. This will be useful for setting our notation and establishing a context for the calculations in the following section.

1.1. *VC theory in asymptotic statistics.*

Definition 1.1. Given a collection of measurable sets in d-dimensional Euclidean space, $\mathcal{A}$, and a finite set $F \subset \mathbb{R}^d$, let

$$F \cap \mathcal{A} = \{F \cap A : A \in \mathcal{A}\} \ \text{ and } \ \Delta_{\mathcal{A}}(n) = \sup_{\{F : |F| = n\}} |F \cap \mathcal{A}|,$$

where $|\cdot|$ denotes cardinality of a (usually finite) set.[1] When $\Delta_{\mathcal{A}}(n)$ is bounded by a polynomial in n, the class $\mathcal{A}$ is said to be a Vapnik-Cervonenkis class or, shortly, a VC-class. The VC-density of $\mathcal{A}$, $\mathrm{dens}_{\mathrm{VC}}(\mathcal{A})$, is the infimum of the set of positive reals, r, such that a constant $C > 0$ exists (possibly depending on r) such that $\Delta_{\mathcal{A}}(n) \leq C\,n^r$ for all $n \in \mathbb{N}$. The VC-dimension of $\mathcal{A}$, $\dim_{\mathrm{VC}}(\mathcal{A})$, is the largest positive integer m such that $\Delta_{\mathcal{A}}(m) = 2^m$. If no such m exists, $\dim_{\mathrm{VC}}(\mathcal{A}) = \infty$.

Under measurability conditions, VC-classes satisfy a non-parametric Uniform Law of Large Numbers, in the sense that if $\mathcal{A}$ is a VC-class, for i.i.d. data $X_1, \ldots, X_n$, sampled from a probability distribution P on $\mathbb{R}^d$ and with

$$P_n(A) = \frac{|\{i : 1 \leq i \leq n,\ X_i \in A\}|}{n}, \tag{1.1}$$

for a set A, then

$$\sup_{A \in \mathcal{A}} |P_n(A) - P(A)| \to 0, \ \text{ a.s., as } n \to \infty.$$

Examining the proof of this Uniform Strong Law reveals that the value of the VC-density is what is actually involved in the arguments leading to this result. Still, most often, asymptotic statisticians have resorted to bounds on the VC-dimension to establish this type of strong laws. This is probably due, at least in part, to the following facts:
(i) Finiteness of the VC-dimension is equivalent to finiteness of the VC-density and
(ii) During the 1980's, several methods were developed for bounding the VC-dimension of a class of sets.

A notion related to VC-dimension and density, that will be used below, is the dual dimension of Assouad, Assouad (1983), defined as follows.

Definition 1.2. For a class of sets in d-dimensional Euclidean space, $\mathcal{A}$, and a finite sub-collection, $\mathcal{H} \subset \mathcal{A}$, let $\mathrm{At}(\mathcal{H})$ denote the set of atoms of the finite algebra

---

[1] In other places $|\cdot|$ will denote absolute value, as usual, but (we hope) this will cause no confusion since the meaning should be clear from the context in each case.

generated by $\mathcal{H}$. For $n \in \mathbb{N}^+$, let $\Delta_{\mathcal{A}}^*(n) = \sup\{|\text{At}(\mathcal{H})| : \mathcal{H} \subset \mathcal{A}, |\mathcal{H}| = n\}$. Assouad's dual dimension is

$$\dim^*(\mathcal{A}) = \sup\{m \in \mathbb{N}^+ : \text{ there exists } \mathcal{H} \subset \mathcal{A}, \text{ with } |\mathcal{H}| = m, |\text{At}(\mathcal{H})| = 2^m\}. \tag{1.2}$$

The dual density of $\mathcal{A}$, $\text{dens}^*(\mathcal{A})$ is the infimum of the positive reals, r, such that for some constant $C > 0$, $\Delta_{\mathcal{A}}^*(m) \leq C\, m^r$ for every $m \in \mathbb{N}^+$.

When it comes to asymptotic results over classes of functions, the concept of metric entropy plays a role similar to that of VC-density for classes of sets.

Definition 1.3. Let $\mathcal{F} \subset L^p(Q)$ for $p = 1$ or 2, and a probability measure Q on $\mathbb{R}^d$. For $\epsilon > 0$, the ϵ-covering number of $\mathcal{F}$ with respect to Q, $N_p(\epsilon, \mathcal{F}, Q)$, is the minimum natural m such that there exist functions $g_1, g_2, \ldots, g_m \in L^p(Q)$ satisfying that, for every $f \in \mathcal{F}$, there is a $j \in \{1, \ldots, m\}$ such that $\|f - g_j\|_{p,Q} < \epsilon$ where $\|\cdot\|_{p,Q}$ is the norm of $L^p(Q)$. $H_p(\epsilon, \mathcal{F}, Q) = \log N_p(\epsilon, \mathcal{F}, Q)$ is called the metric entropy of $\mathcal{F}$.

In order to state a law of large numbers over $\mathcal{F}$, let again $X_1, \ldots, X_n$ denote an i.i.d. sample from a probability distribution P on $\mathbb{R}^d$, and, for each integrable function f, let $P_n f = (1/n) \sum_{i \leq n} f(X_i)$, be the empirical integral of f, while $Pf = \int f(x)\mathrm{d}P(x)$. The class $\mathcal{F}$ is said to have an envelope function $F \in L^p(Q)$ whenever $|f(x)| \leq F(x)$ for all $f \in \mathcal{F}$ and every $x \in \mathbb{R}^d$. A Uniform Law of Large Numbers holds over $\mathcal{F}$ with respect to P, when

$$\sup_{f \in \mathcal{F}} |P_n(f) - Pf| \to 0, \text{ a.s. as } n \to \infty. \tag{1.3}$$

Different results exist in the literature connecting bounds on the metric entropy of a class to Uniform Laws as (1.3).

Suppose that the L^p covering number, $p = 1, 2$, of the class $\mathcal{F}$ with envelope function F is small enough as to satisfy a polynomial bound such as

$$\sup_Q N_p(\epsilon\|F\|_{p,Q}, \mathcal{F}, Q) \leq A\left(\frac{1}{\epsilon}\right)^B \tag{1.4}$$

where A and B are positive constants and the bound is uniform over all choices of the probability measure Q. In this case we will say that *the class $\mathcal{F}$ has polynomial L^p covering number (with exponent B)*. When a class of functions has polynomial covering number, more things can be said regarding asymptotics. When the L^1 covering number is polynomial and the envelope function F is bounded, then the uniform strong law (1.3) can be improved with a uniform speed of convergence:

$$\sup_{f \in \mathcal{F}} |P_n(f) - Pf| \ll \frac{\log n}{\sqrt{n}}, \text{ a.s. as } n \to \infty, \tag{1.5}$$

where $a_n \ll b_n$ means that $a_n/b_n \to 0$, as can be deduced from Theorem 37 in Pollard (1984). Similarly, for classes with polynomial L^2 covering number, results on the speed of convergence in the Central Limit Theorem for the Empirical Process over a class of functions can be obtained, as explained in Section 3.4 of van der Vaart and Wellner (1996). Thus, establishing the polynomial L^p covering number property for a class of functions is quite relevant from the asymptotic viewpoint.

In the current literature, there exist two ways of proving that a class of functions $\mathcal{F}$ has polynomial L^p covering number. One is through the notion of VC-subgraph classes, to be discussed in a moment. The other is through the total boundedness

of the finite dimensional set of parameters that index the functions in $\mathcal{F}$. This second method appears as Example 19.7 in van der Vaart (1998) and we will refer to it in the sequel as the *bounded parameter space* method. The first method has the advantage of not needing the parameter space to be totally bounded and other technical (smoothness) conditions required in the bounded parameter space method. The purpose of the main result in the present article is to significantly simplify the verification that a parametric class of functions is VC-subgraph, thus obtaining that the class has polynomial L^p covering numbers, for $p = 1$ and 2. In examples we will show several classes that appear in concrete applications to be VC-subgraph.

First, let us recall the definition of VC-subgraph classes introduced in Pollard (1982), although this name comes from Dudley (1987).

Definition 1.4. For a class of functions on $\mathbb{R}^d$, $\mathcal{F}$, and $f \in \mathcal{F}$, the subgraph of f is the set

$$\text{subgraph}(f) = \{(x,t) \in \mathbb{R}^{d+1} : 0 \leq t \leq f(x) \text{ or } 0 > t > f(x)\}. \qquad (1.6)$$

The class of all subgraphs of functions in $\mathcal{F}$, $\text{subgraph}(\mathcal{F})$, is a collection of sets in $\mathbb{R}^{d+1}$. When $\text{subgraph}(\mathcal{F})$ is a VC-class, $\mathcal{F}$ is called a VC-subgraph class.

Careful reading of the proof of Lemma 25 in Chapter 2 of Pollard (1984) gives the following:

Theorem 1.5. *If $\mathcal{F}$ is a VC-subgraph class with envelope $F \in L^p(Q)$ and $r = dens_{VC}(subgraph(\mathcal{F}))$, then, for any $\eta > 0$*

$$N_p(\epsilon\|F\|_{p,Q}, \mathcal{F}, Q) \leq A \left(\frac{1}{\epsilon}\right)^{r+\eta} \qquad (1.7)$$

where the constant A depends only on r and η (not on Q). That is, $\mathcal{F}$ has polynomial L^p covering number with exponent $r + \eta$, for every positive η.

The next subsection includes some facts found recently in the context of Model Theory to be used later.

1.2. *Some definitions and results from o-minimality.* O-minimality and results we will be using are a subarea of model theory, in the sense of mathematical logic. We will need some definitions, although we will try to give enough examples so that the un-familiarized reader can get an idea of the concepts we will need.

1.2.1. *First order logic.* In order to work in model theory, one fixes a language $\mathcal{L}$ (say, the language of rings with unity $\mathcal{L} := \{+, \cdot, 0, 1\}$) and a structure which interprets each symbol that appears in $\mathcal{L}$ (for example the real field and the complex fields are both structures in the language of rings with unity) which will be called an $\mathcal{L}$-structure.

In this paper we will always work with a language $\mathcal{L}$ which includes the language of ordered rings $\mathcal{L}_{o.ring} := \{+, \cdot, 0, 1, <\}$, and with structures whose universe is the real numbers as an ordered field, fixing an interpretation of the symbols in $\mathcal{L}$, which will usually be the standard interpretation. For example, the structure with universe $\mathbb{R}$, associated with the language $\mathcal{L} := \{+, \cdot, 0, 1, <, e^x\}$, will be the real numbers, with the natural interpretation of, respectively, the addition, multiplication, additive identity, multiplicative identity, order, and exponential function.

Definition 1.6. Let $\mathcal{L}$ be a fixed language which includes $\mathcal{L}_{o.ring}$, and let $(\mathbb{R}, +, \cdot, 0, 1, <, \dots)$ be an $\mathcal{L}$-structure (here the dots stand for whichever relations or function symbols we want to add to the language). An $\mathcal{L}$-*definable subset of* $\mathbb{R}$ is the set of realizations in our structure $(\mathbb{R}, +, \cdot, 0, 1, <, \dots)$ of a formula which uses only symbols from $\mathcal{L}_{o.ring}$, besides the logic symbols $=, \vee, \wedge, \Rightarrow , \Leftrightarrow, \neg$ (the symbol for "it is not true that"), $\forall$ and $\exists$.

Given an $\mathcal{L}$-structure $\mathcal{M}$, an $\mathcal{L}$-formula $\phi(\bar{x}, \bar{y})$ and $\mathcal{M}$-tuples $\bar{a}$ and $\bar{b}$, we will say that
$$\mathcal{M} \models \phi(\bar{a}, \bar{b})$$
if the formula $\phi(\bar{a}, \bar{b})$ is true in $\mathcal{M}$.

Example 1.7. Let $\phi(x, y)$ be the formula $\exists z \; z^2 = (y - x)$, and let $\mathbb{Q}$ and $\mathbb{R}$ be the rational and the real fields, respectively. Then
$$\mathbb{R} \models \phi(1, 3)$$
but
$$\mathbb{Q} \not\models \phi(1, 3).$$

Of course, if a formula ϕ has no free variables (so that all variables appearing in ϕ are quantified by either $\exists$ or $\forall$) we don't need to replace any variables to know the truth or falsehood of ϕ in any structure of the language. Such formulas are called *sentences*. So for example, $\phi := \forall x \exists y \; y \cdot y = x$ is a sentence true in the complex field, false in the real field, but true in the structure $(\mathbb{R}^{\geq 0}, +, \cdot, 0, 1)$.

Example 1.8.

- The unit disk in $\mathbb{R}^2$ is definable in the structure $(\mathbb{R}, +, \cdot, 0, 1, <)$ by the formula $x^2 + y^2 < 1$.
- The set of integers are a definable subset of $\mathbb{R}$ in the structure $(\mathbb{R}, +, \cdot, 0, 1, < , sin(x))$, since it is the set or realizations of the formula $\exists y, \; sin(y) = 0 \wedge x \cdot \pi = y$.
- By Fact 1.10 below, the integers are not definable in the structure $(\mathbb{R}, +, \cdot, 0, 1, <)$.
- The derivative of a function $f(x)$ is definable in the structure $(\mathbb{R}, +, \cdot, 0, 1, < , f(x))$ by replacing $f'(x) = y$ by the formula

  $$\phi(x, y) := \forall \epsilon \exists \delta \forall h \left(-\delta < h < \delta\right) \Rightarrow \left(|f(x + h) - f(x) - hy| < |h\epsilon|\right).$$

  (Here the absolute value can be defined in the standard way or, since by $|x| := \sqrt{x^2}$ we can define $|x| = y$ by the formula $\theta(x, y) := \left(y^2 = x^2\right) \wedge (y \geq 0)$.)

1.2.2. *o-minimality.* The main logic definition of this paper is the following.

Definition 1.9. Let $\mathcal{L}$ be a fixed language which includes $\mathcal{L}_{o.ring}$, and let $(\mathbb{R}, +, \cdot, 0, 1, <, \dots)$ be an $\mathcal{L}$-structure. We will say that $(\mathbb{R}, +, \cdot, 0, 1, <, \dots)$ is o-minimal if and only if every $\mathcal{L}$-definable subset of $\mathbb{R}$ is a finite union of open intervals and points.

The following is a well known theorem of Tarski (see Tarski (1951)) (not stated originally in this precise manner, since the concept of o-minimality came later).

Fact 1.10. $(\mathbb{R}, +, \cdot, 0, 1, <)$ is o-minimal.

Notice that the subset $\mathbb{Z}$ of $\mathbb{R}$ is not a finite union of intervals and points, so in particular o-minimality implies it is not definable in the real field (or in any o-minimal expansion of the real field).

The study of o-minimal theories started without any other examples of o-minimal expansions of the real field. But it really started becoming a main area of model theory with the following theorem due to Wilkie (1996).

Fact 1.11. $\mathbb{R}_{exp} := (\mathbb{R}, +, \cdot, 0, 1, <, e^x)$ is o-minimal.

This was later generalized by van den Dries et al. (1994) to the following statement:

Fact 1.12. Let $R_{exp,an}$ be the real field expanded by the exponential functions and a function symbol for every analytic function with domain $[-1, 1]^m$ for some m. Then, $R_{exp,an}$ is o-minimal.

For example, even though $(\mathbb{R}, +, \cdot, 0, 1, <, sin(x))$ is not o-minimal, the structure

$$\left(\mathbb{R}, +, \cdot, 0, 1, <, sin\left(x\right)|_{[-1,1]}\right),$$

where $sin(x)|_{[-1,1]}$ is the restriction of the sine function to the closed interval $[-1, 1]$, is o-minimal.

This, together with a result of Speissegger, will cover most of the examples we will consider in this paper. But in order to state Speissegger's result we need the following definition.

Definition 1.13. Let $(\mathbb{R}, +, \cdot, 0, 1, <, \dots)$ be any expansion of the real field. We will say that a differential equation is *Pfaffian over* $(\mathbb{R}, +, \cdot, 0, 1, <, \dots)$ if it is given by a system of equations of the form

$$\frac{\partial f_i}{\partial x_j} = P_{i,j}\left(\bar{x}, f_1\left(\bar{x}\right), \dots, f_i\left(\bar{x}\right)\right)$$

where $f_i\left(\bar{x}\right)$ and $P_{i,j}(\bar{y})$ are definable functions in $(\mathbb{R}, +, \cdot, 0, 1, <, \dots)$, j varies through the number of variables and $1 \leq i \leq N$ for some positive integer N.

The following is due to Speissegger (1999):

Fact 1.14. Let $(\mathbb{R}, +, \cdot, 0, 1, <, \dots)$ be any o-minimal expansion of the real field, and let $f(\bar{x})$ be the solution of a Pfaffian differential equation in $(\mathbb{R}, +, \cdot, 0, 1, <, \dots)$. Then the structure $(\mathbb{R}, +, \cdot, 0, 1, <, \dots, f)$ is o-minimal.

Notice that since e^x is a solution of $\frac{\partial f}{\partial x} = f$, Wilkie's result follows from the o-minimality of the real field $(\mathbb{R}, +, \cdot, 0, 1, <)$ and Speissegger's result.

It follows that the Pfaffian closure of $\mathbb{R}_{exp,an}$ (which we will denote $\mathbb{R}_{an,Pfaff}$) is o-minimal[2].

---

[2]Here we use closure in a manner analogous to "algebraic closure": a structure $\mathcal{M}$ is Pfaffian closed if given any function f, if f is Pfaffian over $\mathcal{M}$, then f is definable in $\mathcal{M}$. The Pfaffian closure of $\mathcal{R}$ is the smallest structure containing $\mathcal{R}$ which is Pfaffian closed

1.2.3. *Uniform definable families of sets.* We will begin with a definition. By a *Uniform definable family of definable sets in* $\mathcal{M}$ we mean a family of definable subsets, all of which are given by changing the parameters in a fixed formula in the language $\mathcal{L}$. Formally,

Definition 1.15. Let $\mathcal{L}$ be any language and let $\mathcal{M} := (M, \dots)$ be an $\mathcal{L}$-structure. We will say that $\mathcal{F}$ is a uniform definable family of definable subsets of M^n if there is an $\mathcal{L}$-formula $\phi(\bar{x}; \bar{y})$ such that

$$\mathcal{F} := \{X_{\bar{b}}\}_{\bar{b} \in M^d}$$

where

$$X_{\bar{b}} := \left\{ \bar{a} \in M^n : \ \phi(\bar{a}; \bar{b}) \text{ is true in } \mathcal{M} \right\}.$$

The tuples $\bar{b}$ vary in M^d (where d is the dimension of the variable $\bar{y}$ in the formula $\phi(\bar{x}; \bar{y})$) and will be called the "parameters" of the subset $X_{\bar{b}}$.

Example 1.16.

- Since the semi-spaces of $\mathbb{R}^n$ are all definable by a formula

$$b_1 \cdot x_1 + b_2 \cdot x_2 + \cdots + b_n \cdot x_n + b_{n+1} < 0,$$

  they are a uniform definable family in $(\mathbb{R}, +, \cdot, 0, 1, <)$.
- The family $\mathcal{A} := \{X_\lambda\}_{\lambda \in \mathbb{R}}$ where

$$X_\lambda := \left\{ x : x \geq 0, \ 0 \leq \frac{x^\lambda - 1}{\lambda} \right\}$$

  is uniformly definable in the structure $(\mathbb{R}, +, \cdot, 0, 1, <, e^x)$. (Technically, we would need to replace $x^\lambda - 1 = z$ with the formula $\exists y \ \left(e^{\lambda \cdot y} - 1 = z \right) \ \wedge \ (e^y = x)$, but this is all definable by a first order formula.)

1.2.4. *The main theorem.* The main theorem relating o-minimality to VC-density, was explicitly stated and proved in Aschenbrenner et al. (2011), although the result is already contained in the paper Karpinski and Macintyre (1997). Here we state the result, followed by an immediate implication in terms of the VC-density of the subgraphs of a class of functions.

Theorem 1.17. *Let $\mathcal{R} := (\mathbb{R}, +, \cdot, 0, 1, <, \dots)$ be an o-minimal expansion of the real field, and let $\mathcal{F} := \{X_{\bar{b}}\}_{\bar{b} \in \mathbb{R}^d}$ be a uniform definable family of sets defined by the formula $\phi(\bar{x}; \bar{y})$ with $\bar{x}$ an m-tuple of variables and $\bar{y}$ a d-tuple of variables.*

$$\mathcal{F} := \{X_{\bar{b}}\}_{\bar{b} \in \mathbb{R}^d} := \left\{ \left\{ \bar{a} \in \mathbb{R}^m : \mathbb{R} \models \phi\left(\bar{a}, \bar{b}\right) \right\} : \bar{b} \in \mathbb{R}^d \right\}.$$

Then the VC-density of $\mathcal{F}$ is at most d.

It follows for instance that, since the family $\mathcal{A} := \{X_\lambda\}_{\lambda \in \mathbb{R}}$ in Example 1.16 is a one parameter family uniformly defined in the o-minimal structure $\mathbb{R}_{exp}$, the VC-density of $\mathcal{A}$ is at most 1. Recall that by definition this means there is some real constant $C > 0$ such that $\Delta_{\mathcal{A}}(n) < C \cdot n$ for all n.

The proof of Theorem 1.17 in Aschenbrenner et al. (2011) is done by induction on the length of the parameter set, and the proof for the 1-case will actually give us a bound for C in this particular example (which we will work out in Subsection 2).

More generally, because the subgraph of the function $f(\bar{x})$ is the set

$$\{(\bar{x}, y) : 0 \leq y \leq f(\bar{x})\} \cup \{(\bar{x}, y) : 0 \geq y \geq f(\bar{x})\}$$

we can state the following general result about the VC-density of subgraphs of uniformly definable functions:

Corollary 1.18. *Let $\mathcal{R} := (\mathbb{R}, +, \cdot, 0, 1, <, \dots)$ be an o-minimal expansion of the real field, and let $\mathcal{F} := \{f_{\bar{b}}(\bar{x})\}_{\bar{b} \in \mathbb{R}^d}$ be a uniform definable family of functions defined by the formula $\phi(y, \bar{x}; \bar{z})$ with $\bar{x}$ an m-tuple of variables and $\bar{z}$ a d-tuple of variables. Explicitly, $\mathcal{F} := \{f_{\bar{b}}(\bar{x})\}_{\bar{b} \in \mathbb{R}^d}$ is defined so that for any $\bar{x}$ and y, we have $f_{\bar{b}}(\bar{x}) = y$ if and only if $\phi(y, \bar{x}; \bar{b})$ holds.*

    Then the VC-density of subgraph$(\mathcal{F})$ is at most d.

The following is a direct consequence of Theorems 1.5 and 1.17 and it is the tool proposed here for statistical applications.

Corollary 1.19. *Let*

$$\mathcal{F} := \{X_{\bar{b}}\}_{\bar{b} \in \mathbb{R}^d} := \left\{ \left\{ \bar{a} \in \mathbb{R}^m : \mathbb{R} \models \phi\left(\bar{a}, \bar{b}\right) \right\} : \bar{b} \in \mathbb{R}^d \right\}.$$

be a (parametric) family of functions on $\mathbb{R}^m$ (uniformly) definable in an o-minimal structure with d parameters, and assume also that $\mathcal{F}$ has bounded envelope function F. Then, $\mathcal{F}$ has polynomial L^p covering number with exponent $d + \eta$, for any $\eta > 0$ and $p = 1, 2$.

2. Bounding the metric entropy of certain classes of functions

Next, we consider certain classes of functions that have appeared in the statistical literature and show how to improve the bounds that have been reported on their metric entropy.

2.1. *Transformations to elliptical symmetry.* Our first example appeared in Quiroz et al. (1996) in connection to the estimation of transformations of multivariate data to elliptical symmetry. In order to establish the efficiency of the method proposed there, part of the problem reduces to the consideration of the class of functions on $\mathbb{R}^+$, $\mathcal{T}$, defined by

$$T_\lambda(x) = \frac{x^\lambda - 1}{\lambda}, \text{ for } x \in \mathbb{R}^+ \qquad \text{and} \qquad \mathcal{T} = \{T_\lambda : \lambda \in \Lambda\},$$

where Λ is a bounded interval. In order to study the class subgraph$(\mathcal{T})$, its dual class, subgraph$(\mathcal{T})^*$, formed by the sets

$$T^{dual}(x, t) = \{\lambda \in \Lambda : 0 \leq t \leq \frac{x^\lambda - 1}{\lambda} \text{ or } \frac{x^\lambda - 1}{\lambda} \leq t < 0\}$$

for $(x, t) \in (\mathbb{R}^+ \times \mathbb{R})$, was considered. Since (as we shall see in the proof of Lemma 1) each $T^{dual}(x, t)$ is the union of at most two intervals, it was argued in Quiroz et al. (1996) that the VC-dimension of subgraph$(\mathcal{T})^*$ is bounded by 4.[3] Then, by Proposition 2.13 in Assouad (1983), it follows that the VC-dimension of subgraph$(\mathcal{T})$ is bounded by $2^4 = 16$ and, therefore, dens$_{VC}$(subgraph$(\mathcal{T})$) will be bounded by 16.

Since the subgraphs of T_λ are a uniformly definable family in the o-minimal structure $\mathbb{R}_{exp}$, by Theorem 1.17 we know that the VC-density of the subgraphs of the functions T_λ is bounded by the size of the parameter set, so it is bounded by a linear function. Now, a closer analysis of the methods in the one dimensional case of the proof of Theorem 1.17 in Aschenbrenner et al. (2011), will give a precise

---

[3]Taking points 1,2,3,4 and 5, one can not, with two intervals, pick the set {1,3,5}.

bound for this family which might be useful for getting precise bounds in any one dimensional set.

Lemma 2.1. *For any fixed $\lambda \in \mathbb{R}$, consider the subgraph*

$$S_\lambda = \left\{ (x,t) \in \mathbb{R}^{\geq 0} \times \mathbb{R} : 0 \leq t \leq \frac{x^\lambda - 1}{\lambda} \vee 0 \geq t \geq \frac{x^\lambda - 1}{\lambda} \right\}$$

and let $\mathcal{A} := \{S_\lambda\}_{\lambda \in \mathbb{R}}$. Then $\Delta_{\mathcal{A}}(n) \leq n + 1$.

Proof: We will again work with the dual subsets but in a different manner. Let

$$X_n := \{(x_1, t_1), (x_2, t_2), \dots, (x_n, t_n)\}$$

be any n points in $\mathbb{R}^2$, and we want to bound the number of sets in $X_n \cap \mathcal{A}$. Now, for each pair (x_i, t_i) let

$$T^{dual}(x_i, t_i) := \{\lambda : (x_i, t_i) \in S_\lambda\}.$$

The next observation is trivial, but it is the central piece of our argument.

Claim 2.2. If S_{λ_1} and S_{λ_2} define different subsets of X_n, then for some (x_i, t_i) we have that

$$\lambda_1 \in T^{dual}(x_i, t_i) \nLeftrightarrow \lambda_2 \in T^{dual}(x_i, t_i)$$

so that λ_1 is in the set $T^{dual}(x_i, t_i)$ and λ_2 isn't, or viceversa.

Notice that if

$$\lambda \in \bigcap_{i \in I} T^{dual}(x_i, t_i) \cap \bigcap_{j \notin I} \left(\mathbb{R} \setminus T^{dual}(x_j, t_j) \right),$$

then

$$S_\lambda \cap X_n = \{(x_i, t_i)\}_{i \in I}.$$

It follows from the claim and the above observation (by an easy and insightful argument left to the reader) that the number of sets $X_n \cap \mathcal{A}$ is equal to the number of non empty intersections of the sets $X_n \cap \mathcal{A}$ and their complements, so that

$$|X_n \cap \mathcal{A}| = \left| \left\{ I \subset \{1, \dots n\} : \bigcap_{i \in I} T^{dual}(x_i, t_i) \cap \bigcap_{j \notin I} \left(\mathbb{R} \setminus T^{dual}(x_j, t_j) \right) \neq \emptyset \right\} \right|.$$

For notation purposes, for any $I \subset \{1, \dots n\}$, let

$$B_I := \bigcap_{i \in I} T^{dual}(x_i, t_i) \cap \bigcap_{j \notin I} \left(\mathbb{R} \setminus T^{dual}(x_j, t_j) \right).$$

So we need to count the subsets I which give consistent (non-empty) boolean combinations.

For any (x_i, t_i) let $c_i := \inf(T^{dual}(x_i, t_i))$ if it exists ($-\infty$ otherwise) and let $d_i := \sup(T^{dual}(x_i, t_i))$ if it exists (∞ otherwise). The derivative of

$$h(\lambda) = \frac{x^\lambda - 1}{\lambda}$$

as a function of λ, is always positive, as one can easily verify[4]. Furthermore, for $x > 1$, $h(\lambda)$ is always positive with infimum 0, while for $x < 1$, $h(\lambda)$ is always

---

[4]In fact, the minimum of $h'(\lambda)$ is always 0, at $\lambda = 0$.

negative with supremum 0. It follows that

$$
\begin{aligned}
&T^{dual}(x_i, t_i) := [c_i, \infty) && \text{if } x_i > 1, t_i \geq \ln(x_i), \\
&T^{dual}(x_i, t_i) := [c_i, 0) \cup (0, \infty) && \text{if } x_i > 1, 0 < t_i < \ln(x_i), \\
&T^{dual}(x_i, t_i) := \mathbb{R} \setminus \{0\} && \text{if } x_i > 1, t_i = 0, \\
&T^{dual}(x_i, t_i) := \emptyset && \text{if } x_i > 1, t_i < 0, \\
&T^{dual}(x_i, t_i) := \emptyset && \text{if } x_i < 1, t_i > 0, \\
&T^{dual}(x_i, t_i) := \mathbb{R} \setminus \{0\} && \text{if } x_i < 1, t_i = 0, \\
&T^{dual}(x_i, t_i) := (-\infty, 0) \cup (0, d_i] && \text{if } x_i < 1, 0 > t_i > \ln(x_i), \\
&T^{dual}(x_i, t_i) := (-\infty, d_i] && \text{if } x_i < 1, 0 > \ln(x_i) \geq t_i,
\end{aligned}
$$

where all values of c_i and d_i are finite. Let $I \subset \{1, 2, \ldots, n\}$ be any subset for which $B_I \neq \emptyset$.

Working on the set $\mathbb{R} \setminus \{0\}$ (disregarding the zero element) and ignoring trivial values of the set $T^{dual}(x_i, t_i)$, we can assume that all our $T^{dual}(x_i, t_i)$ are of the form $[c_i, \infty)$ or $(-\infty, d_i]$. Assume, without loss of generality, that the c_i are listed in increasing order and so are the d_l: For $i < i'$, $c_i < c_{i'}$ and for $l < l'$, $d_l < d_{l'}$. Write T_i for $T^{dual}(x_i, t_i)$.

If B_I is non empty, $i \in I$ and T_i is of the form $[c_i, \infty)$, then for $i' < i$ and $T_{i'} = [c_{i'}, \infty)$, we must have $i' \in I$ (otherwise B_I would be empty). Similarly, if $l \in I$ and T_l is of the form $(-\infty, d_l]$, then for $l' > l$ and $T_{l'} = (-\infty, d_{l'}]$, we must have $l' \in I$.

In order to define a non empty B_I, we first choose i_1 as the largest i such that $T_i = [c_i, \infty)$ and $i \in I$. Let i_2 be the next i such that $T_i = [c_i, \infty)$. Then B_I must be contained in $[c_{i_1}, c_{i_2})$. If i_2 does not exist, B_I must be contained in $[c_{i_1}, \infty)$. Let $L(i_1)$ denote the set of indices l such that T_l is of the form $(-\infty, d_l]$ and $d_l \in [c_{i_1}, c_{i_2})$. Then, B_I is completely determined by choosing l_1, the smallest index l in $L(i_1)$ such that $l \in I$ (certainly, a possible choice is to include no element of $L(i_1)$ in I). For instance, if $L(i_1)$ is non-empty and l_1 is chosen as the smallest element of $L(i_1)$, then $B_I = [c_{i_1}, d_{l_1}]$, while if l_1 is not the smallest element of $L(i_1)$, B_I will be of the form $(d_{l_2}, d_{l_1}]$ for l_2 the largest element in $L(i_1)$ smaller than l_1.

Let m denote the cardinality of the set of indices such that $T^{dual}(x_i, t_i)$ is of the form $[c_i, \infty)$ and let r be the cardinality of the set of indices such that $T^{dual}(x_i, t_i)$ is of the form $(-\infty, d_i]$. From the reasoning above, it follows that the number of choices for B_I is

$$
\sum_{\text{choices of } i_1} (|L(i_1)| + 1).
$$

Using that the sets $L(i_1)$ are disjoint and that the choices for i_1 are $m + 1$ (if we count the option that all i such that $T^{dual}(x_i, t_i)$ is of the form $[c_i, \infty)$ are in I^c), the sum above is bounded by $m + 1 + r$, which is bounded by $n + 1$, finishing the proof. $\square$

2.2. *Goodness of fit to multivariate normality.* In the context of testing for multivariate normality, Quiroz and Dudley (1991), in order to establish the asymptotic distribution of their proposed procedure, considered the following class of functions on $\mathbb{R}^d$: Let $\mathcal{H}_m$ denote the (finite collection of) polynomials in an orthogonal basis of spherical harmonics of degree up to m on the unit sphere in $\mathbb{R}^d$. For $h \in \mathcal{H}_m$,

$c \in \mathbb{R}^d$ and $A \in \mathrm{GL}(d, \mathbb{R})$ let

$$g_{A,c,h} = \begin{cases} h(A(x-c)/\|A(x-c)\|), & \text{for } x \neq c \\ -C, & \text{for } x = c, \end{cases}$$

with C a constant greater than $\sup_\eta |h(\eta)|$ (where the supremum runs over η in the unit sphere of $\mathbb{R}^d$). Let $\mathcal{G} = \{g_{A,c,h} : h \in \mathcal{H}_m, c \in \mathbb{R}^d, A \in \mathrm{GL}(d, \mathbb{R})\}$. In Quiroz and Dudley (1991) the metric entropy of the class $\mathcal{G}$ is estimated via an argument involving VC-hull classes (a concept introduced in Dudley (1987)). The uniform covering number bound obtained through this method, is the following:

$$\text{For} \quad s = 2\binom{m+d}{d} + \binom{2m+d}{d}, \text{ and any } v > 2s/(s+1),$$
$$\sup_Q N_2(\epsilon, \mathcal{G}, Q) \leq K_1 \exp\left(K_2/\epsilon^v\right). \tag{2.1}$$

In particular, it was not possible to show that $\mathcal{G}$ was a VC-subgraph class. Now, every polynomial is definable in the real field $(\mathbb{R}, +, \cdot, 0, 1)$, and so is the unitary sphere, so each of the finite polynomials in $\mathcal{H}_m$ will be definable in the real field. Since multiplication, subtraction, squaring and taking square roots are definable functions, the family of the subgraphs of $g_{A,c,h}$ will be a uniformly definable family in the real field, so by Theorem 1.17 it will have VC-density bounded by the number of free parameters in the family, and in particular this proves that $\mathcal{G}$ is a VC-subgraph class.

And we can do better in computing the VC-density. Adding constants to the language does not affect o-minimality, so we can take all the parameters involved in the polynomials in $\mathcal{H}_m$, add them as constants to the real field, and apply Theorem 1.17 to this new structure. The bound we get for the VC-density of the family of subgraphs of $g_{A,c,h}$ will be equal to the number of free parameters used in getting A and c, so $\mathrm{dens}_{\mathrm{VC}}(\mathrm{subgraph}(\mathcal{G})) \leq d^2 + d$, and by Corollary 1.19, $\mathcal{G}$ has polynomial L^p covering number with exponent $d^2 + d + \eta$, for any positive η.

The large variability of the functions in $\mathcal{G}$ when $\|A(x-c)\|$ approaches zero, makes it difficult to apply the method of bounded parameter space in this case. In Manzotti et al. (2002) the class $\mathcal{G}$ was modified, in order to avoid small values of $\|A(x-c)\|$, at the cost of sacrificing a fraction of the sample data, and only then a variation of the bounded parameter space method was applicable. The bound given here shows that the original $\mathcal{G}$ is a VC-subgraph class, without need for data truncation and may help in understanding the fast convergence reported in Quiroz and Dudley (1991) and Manzotti et al. (2002) of the finite sample distribution of the statistics proposed there to their limit distributions.

2.3. *Complexity penalties in model selection.* In Vapnik's paradigm of Structural Risk Minimization (see Vapnik (1998) and Devroye et al. (1996)) in order to choose between regression models, a complexity penalty is applied to each model depending on estimates of the metric entropy of the family of functions associated. On the other hand, van de Geer (2001), in a fairly general context, establishes the relationship between the metric entropy of classes of functions and the speed of convergence of penalized least squares estimators, in connection with model choice. Both paradigms highlight the need for sharp estimates of metric entropy for the classes of functions defining alternative models in regression.

For example, models of the form

$$Y_i = \eta(X_i^t \beta) + \epsilon_i, \qquad 1 \leq i \leq n, \tag{2.2}$$

appear in the context of generalized linear models McCullagh and Nelder (1989), where Y_i is the univariate response variable, X_i is a d-dimensional vector of covariates, β is a d-dimensional parameter and ϵ_i is the random error of the model. The function η, called the link function, is sometimes assumed to be a monotonically increasing function within a small finite set of candidates. But in a non-parametric setting (which we assume for now), η is only required to be a continuous non-decreasing function with values in [0,1]. Thus, in the non-parametric setting, the goal is to estimate a function in $\mathcal{H}$, the collection of real functions on $\mathbb{R}^d$ of the form $\eta(x^t\beta)$, for $\beta \in \mathbb{R}^d$ and η continuous and non-decreasing from $\mathbb{R}$ to $[0,1]$.

It is known that, if $\mathcal{M}$ denotes the collection of continuous non-decreasing functions from $\mathbb{R}$ to $[0,1]$, then

$$\frac{C_{1,p}}{\epsilon} \le \log \sup_Q N_p(\epsilon, \mathcal{M}, Q) \le \frac{C_{2,p}}{\epsilon}$$

for $p = 1, 2$ and positive constants $C_{1,p}$ and $C_{2,p}$ (see the discussion in Gao and Wellner (2007)). Since, clearly, covering numbers for $\mathcal{H}$ are larger than those for $\mathcal{M}$, we expect a relatively large metric entropy for $\mathcal{H}$, and in particular, this proves that the family of functions of the form $\eta(x^t\beta)$, for $\beta \in \mathbb{R}^d$ and η continuous and non-decreasing from $\mathbb{R}$ to $[0,1]$, is not a VC-subgraph class, so one cannot expect to have any such class definable in an o-minimal structure. In fact, if one composes the increasing function $x - \sin(x)$ with any of the standard maps from $\mathbb{R}$ into $[0,1]$, one can easily exhibit a function which is not definable in any o-minimal expansion of the real field.

Still, in order to estimate η (and β) non-parametrically, one could consider a sequence of nested models, as follows: Let $\mathcal{H}^{(k)}$, $k \ge 2$, denote the collection of functions on $\mathbb{R}^d$ of the form $\eta_k(x^t\beta)$, where $\eta_k(\cdot)$ is continuous and non-decreasing from $\mathbb{R}$ onto $[0,1]$ and there exist numbers $a_1 < a_2 < \cdots < a_k$ and $0 < b_1 < b_2 < \cdots < b_k < 1$, such that, for every $i \le k$, $\eta_k(a_i) = b_i$, η_k is linear between (a_i, b_i) and (a_{i+1}, b_{i+1}), for $1 \le i < k$, while for $x \le a_1$ and $x \ge a_k$ we let

$$\eta_k(x) = B_1 e^{c_1(x-a_1)}, \text{ for } x \le a_1 \qquad \text{and} \qquad \eta_k(x) = 1 - B_k e^{-c_k(x-a_k)}, \text{ for } x \ge a_k,$$

for positive constants B_1, B_k, c_1 and c_k, chosen to make η_k and its derivative continuous on the set $(-\infty, a_1] \cup [a_k, \infty)$. It seems reasonable to believe that, for moderate values of k, the classes $\mathcal{H}^{(k)}$ will provide a good approximation to an unknown function in $\mathcal{H}$, specially when the unknown η is differentiable. And for the $\mathcal{H}^{(k)}$ the metric entropy is significantly smaller than for $\mathcal{H}$, as we see next. This implies (see van de Geer (2001)) a much faster speed of convergence to the best approximation within each $\mathcal{H}^{(k)}$.

Notice first that piecewise functions in an ordered set are very easy to define if each of the components is definable. For instance, given $\bar{a}$, $\bar{b}$, c_1, c_k, B_1, B_k as above, we can define the corresponding η_k by $\eta_k(x) = y$ if and only if

$$\left(y = B_1 e^{c_1(x-a_1)} \wedge x \le a_1\right) \vee \bigvee_{i=1}^{k-1} (y = \theta(x, b_i, b_{i+1}, a_i, a_{i+1}) \wedge x \in (a_i, a_{i+1}))$$

$$\vee \left(y = 1 - B_k e^{c_k(x-a_k)} \wedge x \ge a_k\right),$$

where $\theta(x, b_i, b_{i+1}, a_i, a_{i+1}) := \left(b_i + (b_{i+1} - b_i) \cdot \frac{x-a_i}{a_{i+1}-a_i}\right)$.

We are using $2k+2$ parameters in defining each function η_k, which means we will need $2k+d+2$ parameters to define the function $Y_i = \eta_k(X^t\beta)$ (with X and β in $\mathbb{R}^d$). Thus, the functions $\eta_k(X^t\beta)$ are uniformly definable in the o-minimal model $\mathbb{R}_{exp}$, and, therefore, the VC-density of the family

$$\left\{ \begin{array}{ll} \text{subgraph}\left(\eta_k\left(x^t\beta\right)\right): & a_1 < a_2 < \cdots < a_k,\, 0 < b_1 < b_2 < \cdots < b_k < 1, \\ & \text{with } c_1, c_k, B_1, B_k \text{ as defined above and } \beta \in \mathbb{R}^d \end{array} \right\}$$

is bounded by the number of parameters allowed in the definition, that is $2k+2+d$. This implies, by our Corollary 1, that each $\mathcal{H}^{(k)}$ has polynomial covering number, with exponent $2k+2+d+\delta$, for any $\delta > 0$.

Notice also that one can change the linear functions θ, in the argument just given, for slightly more complex functions in order to guarantee any level of differentiability at the intersections (a_i, b_i) without raising the VC-density (and therefore the complexity) too much. For example, using quadratic functions instead of linear, on each interval $[a_j, a_{j+1}]$, would raise the VC-density to $3k+d+2$, and would allow us to make all the functions in $\mathcal{H}^{(k)}$ differentiable.

In the example we have just described and the following one it would have been somewhat unnatural to impose the assumption of total boundedness on the set of parameters defining the functions in $\mathcal{H}^{(k)}$.

2.4. *Parametric estimation in Generalized Linear Models.* In the same context of generalized linear models of equation (2.2), let us move to a parametric setting by letting η vary over all the Gaussian cummulative distribution functions, with the mean and variance, μ and σ (as well as β), as free parameters to be estimated. We can use Corollary 1.19, to estimate the complexity of this model, as follows: If f is definable in an o-minimal expansion of $\mathbb{R}$, then its antiderivative (indefinite integral) belongs to the Pfaffian closure of such expansion and is therefore definable in an o-minimal structure (recall Fact 1.14). On the other hand, $\exp(-x^2)$ is a definable function in $\mathbb{R}_{exp}$. It follows that the Gaussian density and its cumulative distribution function (c.d.f.) are both definable in $\mathbb{R}_{an,pfaff}$ for any choice of the parameters μ and σ, and the family of functions $\mathcal{H} = \{\eta(x^t\beta) : \eta \text{ is a Gaussian c.d.f.}, \beta \in \mathbb{R}^d\}$ is uniformly definable in $\mathbb{R}_{an,pfaff}$. Since the number of parameters involved is $d+2$, using again Corollary 1.19, we have that the class $\mathcal{H}$ has polynomial L^p covering number with exponent $d+2+\eta$, for every $\eta > 0$, uniformly on all probability laws P over the pair (X, Y) in (2.2).

The analysis we have just outlined would hold in exactly the same manner if, in the definition of the link function of the generalized linear model, the family of Gaussian c.d.f. is replaced by a different parametric family of distributions whose densities are uniformly definable, such as the Gamma family of distributions and others.

In future work we intend to study in more detail, the use of o-minimality methods in the context of complexity penalties for model selection.

Appendix A. VC-dimension vs VC-density

In this appendix we show that, contrary to common belief (at least within the Asymptotic Statistics community), VC-density and VC-dimension can differ significantly over certain classes of sets.

If we restrict ourselves to finite families, it is quite easy to get any possible difference between VC-dimension and VC-density. For example, if we fix any k points in our universe and define $\mathcal{F}$ to be the family of all subsets of these fixed k points, then it is easy to verify that the VC-dimension is k, whereas the VC-density is 0 (the function $\Delta_{\mathcal{A}}(n)$ is bounded by 2^k for all n).

A.1. *Finite unions of families of subsets.* The finite case is of course a very artificial way to force a difference between VC-dimension and VC-density. A more common occurrence happens when $\mathcal{A}$ is the union of two families. Even at the level of the family $\mathcal{F}$ of semi-planes in $\mathbb{R}^2$, it is easy to verify that the VC-dimension of the upper semiplanes $\mathcal{F}^+$ is 2, as is the VC-dimension of the lower semiplanes $\mathcal{F}^-$, whereas the VC-dimension of the union $\mathcal{F}$ is 3. This implies by the Sauer-Shelah Lemma[5] that

$$\Delta_{\mathcal{F}^+}(n) \leq (1/2)(n^2 - n) + n + 1$$

and

$$\Delta_{\mathcal{F}^-}(n) \leq (1/2)(n^2 - n) + n + 1,$$

so by definition of Δ,

$$\Delta_{\mathcal{F}}(n) \leq \Delta_{\mathcal{F}^+}(n) + \Delta_{\mathcal{F}^-}(n) \leq n^2 + n + 2.$$

When taking finite unions of families, the VC-density is the maximum of the VC-densities of the individuals in the union whereas the VC-dimension might be increased.

This example shows a behavior that, although it happens often in the literature, it usually never brings the difference between VC-dimension and VC-density too far apart:

Let $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_k$ each of VC-density N and suppose that $\mathcal{F} := \bigcup \mathcal{F}_i$ has VC-density $N + l$, so that for some set X of size $N + l$ we have $|X \cap \mathcal{F}| = 2^{N+l}$. Now, trivially,

$$|X \cap \mathcal{F}| = |X \cap \bigcup \mathcal{F}_i| \leq \sum_i |X \cap \mathcal{F}_i| \leq k \sum_{j=0}^{N} \binom{N+1}{j},$$

so that k would need to be of the order of 2^l. This means that if we work with unions of k families of sets, the VC-dimension might be increased by not more than a factor of $\log(k)$.

A.2. *A bigger difference.* The final example in this section, is inspired in the finite case, but we provide a one parameter uniformly definable family of subsets of $\mathbb{R}$, with VC-dimension N, for any N.

Fix a set of N points $A := \{a_1, \ldots, a_N\}$ in the interval $(0,1)$, and for each $X_i \subset A$, let I_i be a union of subintervals of $(0,1)$ such that $I_i \cap A = X_i$ (so in particular, we have such I_i for $1 \leq i \leq 2^N$). Now, let $J_i := i + I_i$ be the shift of the set I_i by a number of units equal to its index (so that $J_i \cap J_j = \emptyset$ for $i \neq j$), let

$$\mathbf{J} := \bigcup_{j=1}^{2^N} J_i$$

---

[5]A reference to the Sauer-Shelah Lemma can be found in Sauer (1972). An interesting discussion about the name is available in Bottou (2011)

and finally let

$$\mathcal{A} := \{x + \mathbf{J}\}_{x \in \mathbb{R}}.$$

Then $\mathcal{A}$ has VC-density one (by Theorem 1.17, since it is a one parameter uniformly definable family in the real field) but since for any subset $X_k \subset A$ by definition $X_k = (-k) + \mathbf{J}$, we have that $|\mathcal{A} \cap A| = 2^N$, so that the VC-dimension of $\mathcal{A}$ is at least N, witnessed by A.

References

M. Aschenbrenner, A. Dolich, D. Haskell, D. Macpherson and S. Starchenko. Vapnik-chervonenkis density in some theories without the independence property, i. *ArXiv Mathematics e-prints* (2011). arXiv: 1109.5438.

P. Assouad. Densité et dimension. *Ann. Inst. Fourier (Grenoble)* **33** (3), 233–282 (1983). MR723955.

L. Bottou. About the origins of the vc lemma (2011). Available in: `http://leon.bottou.org/_media/papers/vapnik-symposium-2011.pdf`.

L. Devroye, L. Györfi and G. Lugosi. *A probabilistic theory of pattern recognition*, volume 31 of *Applications of Mathematics (New York)*. Springer-Verlag, New York (1996). ISBN 0-387-94618-7. MR1383093.

L. van den Dries, A. Macintyre and D. Marker. The elementary theory of restricted analytic fields with exponentiation. *Ann. of Math. (2)* **140** (1), 183–205 (1994). MR1289495.

R.M. Dudley. Universal Donsker classes and metric entropy. *Ann. Probab.* **15** (4), 1306–1326 (1987). MR905333.

F. Gao and J.A. Wellner. Entropy estimate for high-dimensional monotonic functions. *J. Multivariate Anal.* **98** (9), 1751–1764 (2007). MR2392431.

S. van de Geer. Least squares estimation with complexity penalties. *Math. Methods Statist.* **10** (3), 355–374 (2001). Meeting on Mathematical Statistics (Marseille, 2000). MR1867165.

M. Karpinski and A. Macintyre. Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks. *J. Comput. System Sci.* **54** (1, part 2), 169–176 (1997). 1st Annual Dagstuhl Seminar on Neural Computing (1994). MR1463257.

A. Manzotti, F.J. Pérez and A.J. Quiroz. A statistic for testing the null hypothesis of elliptical symmetry. *J. Multivariate Anal.* **81** (2), 274–285 (2002). MR1906381.

P. McCullagh and J.A. Nelder. *Generalized linear models*. Monographs on Statistics and Applied Probability. Chapman & Hall, London (1989). ISBN 0-412-31760-5. Second edition [of MR0727836]. MR3223057.

D. Pollard. A central limit theorem for empirical processes. *J. Austral. Math. Soc. Ser. A* **33** (2), 235–248 (1982). MR668445.

D. Pollard. *Convergence of stochastic processes*. Springer Series in Statistics. Springer-Verlag, New York (1984). ISBN 0-387-90990-7. MR762984.

A.J. Quiroz and R.M. Dudley. Some new tests for multivariate normality. *Probab. Theory Related Fields* **87** (4), 521–546 (1991). MR1085180.

A.J. Quiroz, M. Nakamura and F.J. Pérez. Estimation of a multivariate Box-Cox transformation to elliptical symmetry via the empirical characteristic function. *Ann. Inst. Statist. Math.* **48** (4), 687–709 (1996). MR1440265.

N. Sauer. On the density of families of sets. *J. Combinatorial Theory Ser. A* **13**, 145–147 (1972). MR0307902.

P. Speissegger. The Pfaffian closure of an o-minimal structure. *J. Reine Angew. Math.* **508**, 189–211 (1999). MR1676876.

A. Tarski. *A decision method for elementary algebra and geometry.* University of California Press, Berkeley and Los Angeles, Calif. (1951). 2nd ed. MR0044472.

A.W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics.* Cambridge University Press, Cambridge (1998). ISBN 0-521-49603-9; 0-521-78450-6. MR1652247.

A.W. van der Vaart and J.A. Wellner. *Weak convergence and empirical processes.* Springer Series in Statistics. Springer-Verlag, New York (1996). ISBN 0-387-94640-3. With applications to statistics. MR1385671.

V.N. Vapnik. *Statistical learning theory.* Adaptive and Learning Systems for Signal Processing, Communications, and Control. John Wiley & Sons, Inc., New York (1998). ISBN 0-471-03003-1. A Wiley-Interscience Publication. MR1641250.

A.J. Wilkie. Model completeness results for expansions of the ordered field of real numbers by restricted Pfaffian functions and the exponential function. *J. Amer. Math. Soc.* **9** (4), 1051–1094 (1996). MR1398816.