Sample Complexity for Learning Recurrent Perceptron Mappings[*] [†]

Bhaskar DasGupta[‡] **Eduardo D. Sontag**[§]

Abstract

Recurrent perceptron classifiers generalize the usual perceptron model. They correspond to linear transformations of input vectors obtained by means of "autoregressive moving-average schemes", or infinite impulse response filters, and allow taking into account those correlations and dependences among input coordinates which arise from linear digital filtering. This paper provides tight bounds on sample complexity associated to the fitting of such models to experimental data. The results are expressed in the context of the theory of probably approximately correct (PAC) learning.

Keywords

perceptrons, recurrent models, neural networks,
learning, Vapnik-Chervonenkis dimension

[*] This research was supported in part by US Air Force Grant AFOSR-94-0293.

[†] A preliminary version of some of the results reported here will also appear in David S. Touretzky, Michael C. Mozer and Michael E. Hasselmo (eds.), *Advances in Neural Information Processing Systems 8*, The MIT Press, 1996.

[‡] Department of Computer Science, University of Waterloo, Waterloo, Ontario N2L 3G1, CANADA. Email: `bdasgupt@daisy.uwaterloo.ca`

[§] Department of Mathematics, Rutgers University, New Brunswick, NJ 08903, USA. Email: `sontag@hilbert.rutgers.edu`

1 Introduction

One of the most popular approaches to binary pattern classification, underlying many statistical techniques, is based on *perceptrons* or *linear discriminants*; see for instance the classical reference [10]. In this context, one is interested in classifying k-dimensional input patterns $v = (v_1, \ldots, v_k)$ into two disjoint classes A^+ and A^-. A perceptron P which classifies vectors into A^+ and A^- is characterized by a vector (of "weights") $\vec{c} \in \mathbb{R}^k$, and operates as follows. One forms the inner product $\vec{c}.v = c_1 v_1 + \ldots + c_k v_k$. If this inner product is positive, v is classified into A^+, otherwise into A^-; see Figure 1. (A variation allows for an additional constant term c_0, corresponding geometrically to a partition of $\mathbb{R}^k$ by a hyperplane not passing through the origin, but this term, can be incorporated into the remaining weights if one input variable is always set to the value "1".)

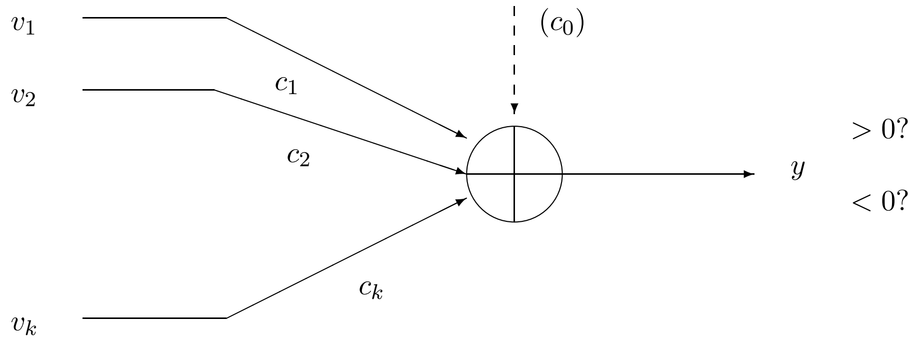

Figure 1: *Usual view of perceptron classifiers*

In practice, given a large number of labeled ("training") samples $(v^{(i)}, \varepsilon_i)$, where $\varepsilon_i \in \{+, -\}$, one attempts to find a vector $\vec{c}$ so that $\vec{c}.v^{(i)}$ is positive when $\varepsilon_i =$ "+" and negative (or zero) otherwise. Finding such a vector amounts to solving a linear programming problem, and recursive algorithms ("perceptron learning method") are popular for its solution. The resulting perceptron corresponding to one such vector $\vec{c}$ is then used to classify new, previously unseen, examples. There are two ways of justifying this procedure. The first is under the hypothesis that the sets A^+ and A^- are indeed linearly separable, that is, there is some hyperplane having them on opposite sides. In addition, it is assumed that the training samples are in either A^+ or A^-, and are labeled accordingly. Provided that the training set is large enough, a hyperplane separating the samples is a good approximation of a true separating hyperplane for A^+ and A^-. A second justification (called sometimes "agnostic learning" in computational learning theory) is based on the fact that, if a large proportion of samples can be linearly separated, then it is very likely that future samples will be correctly classified when using the same rule. Both of these justifications can be made precise on the basis of sample complexity bounds ("VC dimension" as discussed below), and can be found in classical references (see e.g. [27]) as well as [14]. These bounds give estimates of the number of random training samples needed so that a perceptron consistent with (a large proportion of) the seen samples will also, with high probability, perform well on unseen data; see in particular the exposition in [17]. The bounds are linear in the input dimensionality, k, for any fixed confidence levels.

Recurrent Perceptrons

In signal processing and control applications, the size k of the input vectors v is typically very large. As perceptron theory says that a number of training samples proportional to k is required for reliable prediction, this means that a very large number of samples is needed in such applications. However, perceptron theory does not take into account the fact that the signals of interest may exhibit context dependence and correlations, and this prior information can help in narrowing down the search for a classifier. It is often the case in such applications that the classes A^+ and A^- can be separated by means of a *linear dynamical system of fairly small dimensionality*. In that case, the inner product $\vec{c}.v$ represents a convolution by a separating vector $\vec{c}$ that is the impulse-response of a recursive digital filter of some order $n \ll k$. In this model, we think of the inputs as being presented *sequentially* instead of in parallel, to a linear filter, as shown in Figure 2. (In general, at each time t, v_t can be itself a vector, though for simplicity we will restrict our analysis to the case in which these are scalars.) This dynamic behavior can be represented in various ways, for instance by means of an “autoregressive moving average” update

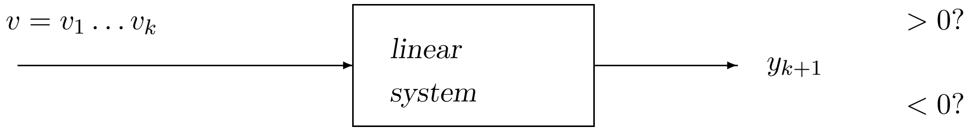

Figure 2: *Recurrent perceptron classifiers*

$$y_t \;=\; \alpha_1 y_{t-n} + \ldots + \alpha_n y_{t-1} + \beta_1 v_{t-n} + \ldots + \beta_n v_{t-1} \quad t = n+1, \ldots, k+1$$

for appropriate coefficients α_i's and β_i's (with the recursion initialized at $y_1 = \ldots = y_n = 0$, and where the sign of the last output y_{k+1} determines the classification), or equivalently, letting $\vec{c}$ denote the impulse response sequence, as a classical perceptron $y_{k+1} = \vec{c}.v$ in which the weight vector $\vec{c}$ has a special form, namely $\vec{c}$ is *n-recursive*, meaning that there exist real numbers $r_1, \ldots, r_n$ so that

$$c_j \;=\; \sum_{i=1}^{n} c_{j-i} r_i \,, \quad j = n+1, \ldots, k \,.$$

Seen in this context, the usual perceptrons are nothing more than the very special subclass of “finite impulse response” systems (all poles at zero); thus it is appropriate to call the more general class “recurrent” or “IIR (infinite impulse response)” perceptrons (as done in [1, 2]).

The BPS (“backpropagation for sequences”) approach developed by Bengio and coauthors (see [6], Section 4.4) is an example of an application of these ideas in signal processing. The autoregressive equation is seen as determining the behavior of dynamical processing units (cf. [6], equation 4.17), and there is an output nonlinearity given by a “squashing” function, corresponding in our case to taking the sign of the output. (Sometimes cascades of these units are allowed, which makes the model capable of handling more highly nonlinear data as well.) The reference [6] describes experimental data regarding the use of the BPS architecture in several applications, including the speech recognition task of speaker-independent discrimination between the consonants "b" and "d" (in this case, at each t the input v_t is a vector whose coordinates consist of Fourier-like parameters associated to speech samples as well as some additional information on signal levels). There is also related work in control theory dealing with such classifying, or more generally quantized-output, linear systems; see [9, 16, 22]. Various dynamical system models for classification appear also when learning finite automata and languages —see e.g. [12]— and in signal processing as a channel equalization problem (at least in the simplest 2-level case) when modeling linear channels transmitting digital data from a quantized source —see [3] and also the related paper [19].

Thus we are motivated to look into the theoretical issue that arises from the fitting data to perceptrons in which the weight vector $\vec{c}$ is constrained to lie in the class of n-recursive (with fixed $n \ll k$) vectors. One may expect that the size of learning samples required in order to reliably classify future unlabeled inputs will be much smaller than k. Indeed, roughly speaking the main result is that the number of samples needed is proportional to the just *logarithm* of the length k (as opposed to k itself, as would be the case if one did not take advantage of the recurrent structure). This number is in general larger than the number of parameters $2n$, a perhaps surprising fact (see Remark 4.4). The precise formulation is in terms of computational leaning theory (or, in more classical statistical language, in terms of generalized Glivenko-Cantelli theorems for uniform convergence of empirical probabilities) and is reviewed below. We also make some remarks on the actual computational complexity of finding a vector $\vec{c}$ consistent with the training data, and we also discuss briefly the identification of linear dynamical systems, in which the complete output (as opposed to merely the sign) is of interest.

Sample Complexity and VC Dimension

We next very briefly review some (by now standard) notions regarding sample complexity, with the purpose of motivating the main results, which deal with the calculation of VC dimensions. For more details see the books [27, 28], the paper [7], or the survey [17].

In the general classification problem, an input space $\mathbb{X}$ as well as a collection $\mathcal{F}$ of maps $\mathbb{X} \to \{-1, 1\}$ are assumed to have been given. (The set $\mathbb{X}$ is assumed to be either countable or an Euclidean space, and the maps in $\mathcal{F}$ are assumed to be measurable. In addition, mild regularity assumptions are made which insure that all sets appearing below are measurable, but details are omitted since in our context these assumptions are always satisfied.) Let W be the set of all sequences

$$w = (u_1, \psi(u_1)), \ldots, (u_s, \psi(u_s))$$

over all $s \geq 1$, $(u_1, \ldots, u_s) \in \mathbb{X}^s$, and $\psi \in \mathcal{F}$. An *identifier* is a map $\varphi : W \to \mathcal{F}$. The value of φ on a sequence w as above will be denoted as φ_w. The *error* of φ with respect to a probability measure P on $\mathbb{X}$, a $\psi \in \mathcal{F}$, and a sequence $(u_1, \ldots, u_s) \in \mathbb{X}^s$, is

$$\mathrm{Err}_\varphi(P, \psi, u_1, \ldots, u_s) := \mathrm{Prob}\left[\varphi_w(u) \neq \psi(u)\right]$$

(where the probability is being understood with respect to P).

The class $\mathcal{F}$ is said to be (uniformly) *learnable* if there is some identifier φ with the following property: For each $\varepsilon, \delta > 0$ there is some s so that, for every probability P and every $\psi \in \mathcal{F}$,

$$\mathrm{Prob}\left[\mathrm{Err}_\varphi(P, \psi, u_1, \ldots, u_s) > \varepsilon\right] < \delta$$

(where the probability is being understood with respect to P^s on $\mathbb{X}^s$).

In the learnable case, the function $s(\varepsilon, \delta)$ which provides, for any given ε and δ, the smallest possible s as above, is called the *sample complexity* of the class $\mathcal{F}$. It can be proved that learnability is equivalent to finiteness of the *Vapnik-Chervonenkis (VC) dimension* ν of the class $\mathcal{F}$, a combinatorial concept whose definition we recall later. In fact, $s(\varepsilon, \delta)$ is bounded by a polynomial in $1/\varepsilon$ and $1/\delta$ and is proportional to ν in the following precise sense (cf. [7, 26]):

$$s(\varepsilon, \delta) \ \leq \ \max\left\{\frac{8\nu}{\varepsilon}\log\left(\frac{13}{\varepsilon}\right), \frac{4}{\varepsilon}\log\left(\frac{2}{\delta}\right)\right\}$$

Moreover, lower bounds on $s(\varepsilon, \delta)$ are also known, in the following sense (cf. [7]): for $0 < \varepsilon < \frac{1}{2}$, and assuming that the collection $\mathcal{F}$ is not trivial (i.e., $\mathcal{F}$ does not consist of just one mapping or a collection of two disjoint mappings, see [7] for details), we must have

$$s(\varepsilon, \delta) \ \geq \ \max\left\{\frac{1-\varepsilon}{\varepsilon}\ln\left(\frac{1}{\delta}\right), \nu(1 - 2(\varepsilon(1-\delta) + \delta))\right\}$$

The above bounds motivate the studies dealing with estimating VC dimension, as we pursue here.

When there is an algorithm that allows computing an identifier φ in time polynomial on the sample size, the class is said to be learnable in the PAC ("probably approximately correct") sense of Valiant (cf. [25]). In this paper, we first study the question of uniform learnability in the sample complexity sense, for recurrent perceptron concept classes, and we also prove a result, in Section 5 regarding PAC learnability for such classes.

There is a variation of the PAC learning results, in which the objective is not to obtain arbitrary small errors but merely to approximate the smallest possible error rate achievable with a given class of functions $\mathcal{F}$. This is much more realistic in applications, as there is no reason to assume that a given structure (such as recurrent perceptrons of a given order) will represent the data precisely. The VC dimension appears again in the sample complexity estimates associated to this "agnostic learning" problem (the term originates in the fact that we do not wish to assume a particular "target concept" that generates the observed samples). A typical result in this area is as follows (cf. [17], based on [18, 14], for more details). Let A be any distribution over $\mathbb{X} \times \{-1, 1\}$. Pick any $\varepsilon, \delta > 0$. Suppose that a sample $(u_1, y_1), \ldots, (u_s, y_s)$ of length $s = s(\varepsilon, \delta)$ is drawn according to A, where

$$s(\varepsilon, \delta) \geq \frac{576}{\varepsilon^2}(2\nu \ln \frac{48e}{\varepsilon} + \ln \frac{8}{\delta})\,.$$

Assume that we now approximately minimize the empirical risk, in the sense that we find a function $\psi \in \mathcal{F}$ so that the average number of missclassifications $\mu(\psi) := (1/s)\mathrm{card}\,\{i|\psi(u_i) \neq y_i\}$ when using ψ is within $\varepsilon/3$ of the minimal possible number $\inf_{\psi' \in \mathcal{F}} \mu(\psi')$. Then, with probability $\geq 1 - \delta$ (with respect to the random drawing of the sample), the expectation of the error made by ψ on samples drawn according to the same distribution A is within ε of the minimal possible expected error among all possible $\psi' \in \mathcal{F}$.

Generalizations to the learning of real-valued (as opposed to Boolean) functions, by evaluation of the "pseudo-dimension" of recurrent maps, are also possible; see the brief discussion in Section 6.

2 Definitions and Statements of Main Results

The concept of VC dimension is classically defined in terms of abstract concept classes. Assume that we are given a set $\mathbb{X}$, called the *set of inputs*, and a family of subsets $\mathcal{C}$ of $\mathbb{X}$, called the set of "concepts." A subset $X \subseteq \mathbb{X}$ is said to be shattered (by the class $\mathcal{C}$) if for each subset $B \subseteq X$ there is some $C \in \mathcal{C}$ such that $B = C \bigcap X$. The VC dimension is then the largest possible positive integer n (possibly $+\infty$) so that there is some $X \subseteq \mathbb{X}$ of cardinality n which can be shattered. An equivalent manner of stating these notions, somewhat more suitable for our purposes, proceeds by identifying the subsets of X with Boolean functions from X to $\{-1, 1\}$ (we pick $\{-1, 1\}$ instead of $\{0, 1\}$ for notational convenience): to each such Boolean function ϕ there is an associated subset, namely $\{x \in X \mid \phi(x) = 1\}$, and conversely, to each set $B \subseteq X$ one can associate its characteristic function ϕ_B defined on the set X. Similarly, we can think of the sets $C \in \mathcal{C}$ as Boolean functions on $\mathbb{X}$ and the intersections $C \bigcap X$ as the restrictions of such functions to X. Thus we restate the definitions now in terms of functions.

Given the set $\mathbb{X}$, and a subset X of $\mathbb{X}$, a *dichotomy* on X is a function

$$\delta \,:\, X \to \{-1, 1\}\,.$$

Assume given a class $\mathcal{F}$ of functions $\mathbb{X} \to \{-1, 1\}$, to be called the class of *classifier* functions. The subset $X \subseteq \mathbb{X}$ is *shattered* by $\mathcal{F}$ if each dichotomy on X is the restriction to X of some $\phi \in \mathcal{F}$. The *Vapnik-Chervonenkis dimension* $\textsc{vc}\,(\mathcal{F})$ is the supremum (possibly infinite) of the set of integers κ for which there is some subset $X \subseteq \mathbb{X}$ of cardinality κ which can be shattered by $\mathcal{F}$.

Pick any two integers $n{>}0$ and $q{\geq}0$. A sequence

$$\vec{c} = (c_1, \ldots, c_{n+q}) \in \mathbb{R}^{n+q}$$

is said to be *n-recursive* if there exist real numbers $r_1, \ldots, r_n$ so that

$$c_{n+j} \;=\; \sum_{i=1}^{n} c_{n+j-i} r_i \,,\quad j = 1, \ldots, q\,.$$

(In particular, every sequence of length n is n-recursive, but the interesting cases are those in which $q \neq 0$, and in fact $q \gg n$.) Given such an n-recursive sequence $\vec{c}$, we may consider its associated *perceptron* classifier. This is the map

$$\phi_{\vec{c}} \,:\, \mathbb{R}^{n+q} \to \{-1, 1\} \,:\quad (x_1, \ldots, x_{n+q}) \;\mapsto\; \text{sign}\left(\sum_{i=1}^{n+q} c_i x_i\right)$$

where the sign function is understood to be defined by sign$\,(z) = -1$ if $z \leq 0$ and sign$\,(z) = 1$ otherwise. (Changing the definition at zero to be $+1$ would not change the results to be presented in any way.) We now introduce, for each two fixed n, q as above, a class of functions:

$$\mathcal{F}_{n,q} \;:=\; \left\{ \phi_{\vec{c}} \,|\; \vec{c} \in \mathbb{R}^{n+q} \text{ is } n\text{-recursive} \right\}.$$

This is understood as a function class with respect to the input space $\mathbb{X} = \mathbb{R}^{n+q}$, and we are interested in estimating $\textsc{vc}\,(\mathcal{F}_{n,q})$.

Our main result will be as follows (in this paper, all logarithms are understood to be in base 2):

Theorem 1 $\boxed{\max\left\{n, n\lfloor\log(\lfloor 1+\frac{q-1}{n}\rfloor)\rfloor\right\} \;\leq\; \mathrm{VC}\,(\mathcal{F}_{n,q}) \;\leq\; \min\{n+q\,,\, 18n+4n\log(q+1)\}}$.

The upper bound is a simple consequence of an argument based on parameter counts, and is given in Section 4. Much more interesting is the almost matching lower bound, which will involve a result on dual VC dimensions which we prove in Section 3.

Some particular cases are worth discussing. When $q = O(n)$ then both the upper and the lower bounds are of the type cn for some (different) constants c. If $q = \Omega(n^{1+\epsilon})$ (for any constant $\epsilon > 0$), then both the upper and the lower bounds are of the form $cn\log(\frac{q}{n})$ for some constants c. In this latter case, assume that one is interested in the behavior of $\mathrm{VC}\,(\mathcal{F}_{n,q})$ as $n \to +\infty$ while q grows polynomially in n; then the upper and lower bounds are both of the type $cn\log n$, for some constants c. If instead q grows exponentially on n, both the upper and lower bounds are polynomial in n.

The organization of the rest of the paper is as follows. In Section 3 we prove an abstract result on VC-dimension, which is then used in Section 4 to prove Theorem 1. In Section 5, we show that the consistency problem for recurrent perceptrons can be solved in polynomial time, for any fixed n; some recent facts regarding representations of real numbers and decision problems for real-closed fields, needed in this Section, are reviewed in an Appendix. Finally, in Section 6 we make some final comments about bounds on the sample complexity needed for identification of linear dynamical systems, that is to say, the real-valued functions obtained when not taking "signs" when defining the maps $\phi_{\vec{c}}$.

3 An Abstract Result on VC Dimension

Assume that we are given two sets $\mathbb{X}$ and Λ, to be called in this context the set of *inputs* and the set of *parameter values* respectively. Suppose that we are also given a function

$$F \;:\; \Lambda \times \mathbb{X} \to \{-1, 1\}\,.$$

Associated to this data is the class of functions

$$\mathcal{F} \;:=\; \{F(\lambda,\cdot) : \mathbb{X} \to \{-1,1\} \,|\, \lambda \in \Lambda\}$$

obtained by considering F as a function of the inputs alone, one such function for each possible parameter value λ. We will prove lower bounds in Theorem 1 by studying the VC dimension of classes obtained in this parametric fashion.

Note that, given the same data one could, dually, study the class

$$\mathcal{F}^* \;:\; \{F(\cdot,\xi) : \Lambda \to \{-1,1\} \,|\, \xi \in \mathbb{X}\}$$

which is obtained by fixing the elements of $\mathbb{X}$ and thinking of the parameters as inputs. It is well-known (cf. [11], Theorem 9.3.2, and in any case, a consequence of the much more general result to be presented below) that

$$\mathrm{VC}\,(\mathcal{F}) \geq \lfloor\log(\mathrm{VC}\,(\mathcal{F}^*))\rfloor,$$

which provides a lower bound on $\mathrm{VC}(\mathcal{F})$ in terms of the "dual VC dimension." A sharper estimate is possible when Λ can be written as a product of n sets

$$\Lambda = \Lambda_1 \times \Lambda_2 \times \ldots \times \Lambda_n \tag{1}$$

and that is the topic which we develop next.

We assume from now on that a decomposition of the form in Equation (1) is given, and will define a variation of the dual VC dimension by asking that only certain dichotomies on Λ be obtained from $\mathcal{F}^*$. We define these dichotomies only on "rectangular" subsets of Λ, that is, sets of the form

$$L = L_1 \times \ldots \times L_n \subseteq \Lambda$$

with each $L_i \subseteq \Lambda_i$ a nonempty subset. Given any index $1 \leq \kappa \leq n$, by a *κ-axis dichotomy* on such a subset L we mean any function $\delta : L \to \{-1, 1\}$ which depends only on the κth coordinate, that is, there is some function $\phi : L_\kappa \to \{-1, 1\}$ so that $\delta(\lambda_1, \ldots, \lambda_n) = \phi(\lambda_\kappa)$ for all $(\lambda_1, \ldots, \lambda_n) \in L$; an axis dichotomy is a map that is a κ-axis dichotomy for some κ. A rectangular set L will be said to be *axis-shattered* if every axis dichotomy is the restriction to L of some function of the form $F(\cdot, \xi) : \Lambda \to \{-1, 1\}$, for some $\xi \in \mathbb{X}$.

Theorem 2 *If $L = L_1 \times \ldots \times L_n \subseteq \Lambda$ can be axis-shattered and each set L_i has cardinality r_i, then $\mathrm{VC}(\mathcal{F}) \geq \lfloor \log(r_1) \rfloor + \ldots + \lfloor \log(r_n) \rfloor$.*

Note that in the special case $n = 1$ one recovers the result $\mathrm{VC}(\mathcal{F}) \geq \lfloor \log(\mathrm{VC}(\mathcal{F}^*)) \rfloor$. We will prove this theorem below, after a couple of small observations.

Remark 3.1 Assume that $L = L_1 \times \ldots \times L_n \subseteq \Lambda$ can be axis-shattered. Pick any indices (possibly equal) $\kappa_1, \kappa_2 \in \{1, \ldots, n\}$ and any functions $\phi_i : L_{\kappa_i} \to \{-1, 1\}$, $i = 1, 2$. By definition of axis-shattering, there exist elements $\xi_1, \xi_2 \in \mathbb{X}$, such that

$$F(\lambda_1, \ldots, \lambda_n, \xi_i) = \phi_i(\lambda_{\kappa_i}) \quad \forall (\lambda_1, \ldots, \lambda_n) \in L_1 \times \ldots \times L_n \,. \tag{2}$$

We then have:

(a) If $\kappa_1 = \kappa_2$ and $\xi_1 = \xi_2$ then $\phi_1 = \phi_2$.

(b) If $\kappa_1 \neq \kappa_2$ and $\xi_1 = \xi_2$ then both ϕ_1 and ϕ_2 are constant functions.

Property (a) is obvious. Property (b) is proved as follows. Without loss of generality, we may take $\kappa_1 = 1$ and $\kappa_2 = 2$. Now pick $\widehat{\lambda}_2, \ldots, \widehat{\lambda}_n$ arbitrarily. Then

$$\phi_1(\lambda) = F(\lambda, \widehat{\lambda}_2, \ldots, \widehat{\lambda}_n, \xi) = \phi_2(\widehat{\lambda}_2)$$

for all $\lambda \in L_1$, and a similar argument shows that ϕ_2 is constant as well. □

Remark 3.2 Let $\mathcal{S} = \{s_1, s_2, \ldots, s_r\}$ be a set of cardinality $r = 2^m$, where m is a positive integer. Let M be the $m \times r$ matrix whose columns are the 2^m possible vectors in $\{-1, 1\}^m$ and define the functions ϕ_i by the formula $\phi_i(s_j) = M_{ij}$ for all $1 \leq i \leq m$ and $1 \leq j \leq r$. Then, it is easy to see that the the set of m (distinct) dichotomies $\phi_1, \phi_2, \ldots, \phi_m$ on $\mathcal{S}$ have the following property: For each vector $(a_1, a_2, \ldots, a_m) \in \{-1, 1\}^m$, there exists a unique index $j \in \{1, \ldots r\}$ such that

$$\phi_i(s_j) = a_i \ , \quad i = 1, \ldots, m \, . \tag{3}$$

Moreover, none of the functions ϕ_i is a constant function. □

Proof of Theorem 2. We may assume without loss of generality that each $r_\kappa = 2^{m_\kappa}$ for some positive integers $m_1, \ldots, m_n$. This is because any possible indices so that $r_\kappa = 1$ can be dropped (and the result proved with smaller n), and, for each $r_\kappa > 1$, a subset L'_κ of L_κ, of cardinality $2^{\lfloor \log r_\kappa \rfloor}$, could be used instead of the original L_κ if r_κ is not a power of two.

To prove the Theorem, it will be enough to find n disjoint subsets $X_1, X_2, \ldots, X_n$ of $\mathbb{X}$, of cardinalities $m_1, \ldots, m_n$ respectively, so that the set $X = X_1 \bigcup X_2 \bigcup \ldots \bigcup X_n$ is shattered. Pick any $\kappa \in \{1, \ldots, n\}$. Consider the set $L_\kappa = \{l_{\kappa,1}, l_{\kappa,2}, \ldots, l_{\kappa,r_\kappa}\}$. By Remark 3.2 applied to this set, there exists a set of m_κ distinct and nonconstant dichotomies $\phi_{\kappa,1}, \phi_{\kappa,2}, \ldots, \phi_{\kappa,m_\kappa}$ on L_κ so that, for any vector $(a_1, a_2, \ldots, a_{m_\kappa}) \in \{-1, 1\}^{m_\kappa}$, there exists a unique index $1 \leq j_\kappa \leq r_\kappa$ so that

$$\phi_{\kappa,i}(l_{\kappa,j_\kappa}) = a_i \ , \quad i = 1, \ldots, m_\kappa \, . \tag{4}$$

Since L can be axis-shattered, each of the axis dichotomies $\phi_{\kappa,i}$ can be realized as a function $F(\cdot, \xi)$. That is, there exists a set inputs

$$X_\kappa \ = \ \{\xi_{\kappa,1}, \xi_{\kappa,2}, \ldots, \xi_{\kappa,m_\kappa}\}$$

so that, for each $i = 1, \ldots, m_\kappa$,

$$F(\lambda_1, \ldots, \lambda_n, \xi_{\kappa,i}) = \phi_{\kappa,i}(\lambda_\kappa) \ , \quad \forall (\lambda_1, \ldots, \lambda_n) \in L_1 \times \ldots \times L_n \, . \tag{5}$$

Note also that, by construction, $\xi_{\kappa,i} \neq \xi_{\kappa,i'}$ for $i \neq i'$, since the corresponding functions $\phi_{\kappa,i}$ are distinct (recall Remark 3.1, part (a)).

Summarizing, for each vector $(a_1, a_2, \ldots, a_{m_\kappa}) \in \{-1, 1\}^{m_\kappa}$ and for each $\kappa \in \{1, \ldots, n\}$ there is some $1 \leq j_\kappa \leq r_\kappa$ so that

$$F(\lambda_1, \ldots, \lambda_{\kappa-1}, l_{\kappa,j_\kappa}, \lambda_{\kappa+1}, \ldots, \lambda_n, \xi_{\kappa,i}) \ = \ \phi_{\kappa,i}(l_{\kappa,j_\kappa}) \ = \ a_i \ , \quad i = 1, \ldots, m_\kappa \tag{6}$$

for all $\lambda_q \in L_q$ $(q \neq \kappa)$. We do this construction for each κ and define $X := X_1 \bigcup X_2 \bigcup \ldots \bigcup X_n$. Note that the sets X_κ are disjoint, since $\xi_{\kappa,i} \neq \xi_{\kappa',i'}$ whenever $\kappa \neq \kappa'$ (by part (b) of Remark 3.1 and the fact that the functions $\phi_{\kappa,i}$ are all nonconstant). The set X can be shattered. Indeed, assume given any dichotomy $\delta : X \to \{-1, 1\}$. Using Equation (6), with the vector $a = (\delta(\xi_{\kappa,1}), \ldots, \delta(\xi_{\kappa,m_\kappa}))$ for each κ, it follows that for each $\kappa \in \{1, \ldots, n\}$ there is some $1 \leq j_\kappa \leq r_\kappa$ so that

$$F(l_{1,j_1}, \ldots, l_{n,j_n}, \xi_{\kappa,i}) \ = \ \delta(\xi_{\kappa,i}) \ , \quad i = 1, \ldots, m_\kappa.$$

That is, the function $F(\lambda, \cdot)$ coincides with δ on X, when one picks $\lambda = (l_{1,j_1}, \ldots, l_{n,j_n})$. ∎

Note that the lower bound in the above result is almost tight, because by Lemma 4.2 there is a set of the form $L = L_1 \times \ldots \times L_n \subseteq \Lambda$ which can be axis-shattered and for which $\text{VC}\,(\mathcal{F}) = O(n \log(rn))$, with cardinality of each L_i greater or equal to r for each i.

4 Proof of Main Result

We recall the following result; it was proved, using Milnor-Warren bounds on the number of connected components of semi-algebraic sets, by Goldberg and Jerrum:

Fact 4.1 ([13]) Assume given a function $F : \Lambda \times \mathbb{X} \to \{-1,1\}$ and the associated class of functions $\mathcal{F} := \{F(\lambda,\cdot) : \mathbb{X} \to \{-1,1\} \,|\, \lambda \in \Lambda\}$. Suppose that $\Lambda = \mathbb{R}^k$ and $\mathbb{X} = \mathbb{R}^n$, and that the function F can be defined in terms of a Boolean formula involving at most s polynomial inequalities in $k+n$ variables, each polynomial being of degree at most d. Then, $\mathrm{VC}(\mathcal{F}) \leq 2k\log(8eds)$. □

Lemma 4.2 $\mathrm{VC}(\mathcal{F}_{n,q}) \leq \min\{n+q\,,\, 18n + 4n\log(q+1)\}$

Proof. Since $\mathcal{F}_{n,q} \subseteq \mathcal{F}_{n+q,0}$,

$$\mathrm{VC}(\mathcal{F}_{n,q}) \leq \mathrm{VC}(\mathcal{F}_{n+q,0}) = n+q$$

where the last equality follows from the fact that $\mathrm{VC}(\mathrm{sign}(\mathcal{G})) = \dim(\mathcal{G})$ when $\mathcal{G}$ is a vector space of real-valued functions (the standard "perceptron" model). On the other hand, it is easy to see (by induction on j) that, for n-recursive sequences, c_{n+j} (for $1 \leq j \leq q$) is a polynomial in $c_1, c_2, \ldots, c_n, r_1, r_2, \ldots, r_n$ of degree exactly $j+1$. Thus one may see $\mathcal{F}_{n,q}$ as a class obtained parametrically, and applying Fact 4.1 (with $k=2n$, $s=1$, $d=q+1$) gives $\mathrm{VC}(\mathcal{F}_{n,q}) < 18n + 4n\log(q+1)$. ■

Lemma 4.3 $\mathrm{VC}(\mathcal{F}_{n,q}) \geq \max\{n, n\lfloor\log(\lfloor 1+\frac{q-1}{n}\rfloor)\rfloor\}$

Proof. As $\mathcal{F}_{n,q}$ contains the class of functions $\phi_{\vec{c}}$ with $\vec{c} = (c_1,\ldots,c_n,0,\ldots,0)$, which in turn being the set of signs of an n-dimensional linear space of functions, has VC dimension n, we know that $\mathrm{VC}(\mathcal{F}_{n,q}) \geq n$. Thus we are left to prove that if $q > n$ then $\mathrm{VC}(\mathcal{F}_{n,q}) \geq n\lfloor\log(\lfloor 1+\frac{q-1}{n}\rfloor)\rfloor$.

The set of n-recursive sequences of length $n+q$ includes the set of sequences of the following special form:

$$c_j = \sum_{i=1}^{n} \alpha_i l_i^{j-1}\,, \quad j = 1,\ldots,n+q \tag{7}$$

where $\alpha_i, l_i \in \mathbb{R}$ for each $i = 1,\ldots,n$. (More precisely, this is a characterization of those n-recursive sequences of length $n+q$ for which the characteristic roots, that is, the roots of the polynomial determined by the recursion coefficients, are all real and distinct; such facts are classical in the theory of recurrences.) In turn, this includes the sequences as in Equation (7) in which one uses only $\alpha_1 = \ldots = \alpha_n = 1$. Hence, to prove the lower bound, it is sufficient to study the class of functions induced by

$$F : \mathbb{R}^n \times \mathbb{R}^{n+q} \to \{-1,1\}\,, \quad (\lambda_1,\ldots,\lambda_n,x_1,\ldots,x_{n+q}) \mapsto \mathrm{sign}\left(\sum_{i=1}^{n}\sum_{j=1}^{n+q} \lambda_i^{j-1} x_j\right) \tag{8}$$

Let $r = \lfloor \frac{q+n-1}{n} \rfloor$ and let $L_1, \ldots, L_n$ be n disjoint sets of real numbers (if desired, integers), each of cardinality r. Let $L = \bigcup_{i=1}^{n} L_i$. In addition, if $rn < q+n-1$, then select an additional set B of $(q+n-rn-1)$ real numbers disjoint from L.

We will apply Theorem 2, showing that the rectangular subset $L_1 \times \ldots \times L_n$ can be axis-shattered. Pick any $\kappa \in \{1, \ldots, n\}$ and any $\phi : L_\kappa \to \{-1, 1\}$. Consider the (unique) interpolating polynomial

$$p(\lambda) \;=\; \sum_{j=1}^{n+q} x_j \lambda^{j-1}$$

in λ of degree $q+n-1$ such that

$$p(\lambda) = \begin{cases} \phi(\lambda) & \text{if } \lambda \in L_\kappa \\ 0 & \text{if } \lambda \in (L \cup B) - L_\kappa. \end{cases}$$

One construction of such a polynomial is via the Lagrange formula

$$\sum_{l \in L_\kappa} \phi(l) \, \frac{\Pi_{l_j \in L \cup B \,;\, l_j \neq l} (\lambda - l_j)}{\Pi_{l_j \in L \cup B \,;\, l_j \neq l} (l - l_j)} \,.$$

Now pick $\xi = (x_1, \ldots, x_{n+q-1})$. Observe that

$$F(l_1, l_2, \ldots, l_n, x_1, \ldots, x_{n+q}) \;=\; \operatorname{sign}\left(\sum_{i=1}^{n} p(l_i) \right) \;=\; \phi(l_\kappa)$$

for all $(l_1, \ldots, l_n) \in L_1 \times \ldots \times L_n$, since $p(l) = 0$ for $l \notin L_\kappa$ and $p(l) = \phi(l)$ otherwise. It follows from Theorem 2 that $\text{VC}\,(\mathcal{F}_{n,q}) \geq n \lfloor \log(r) \rfloor$, as desired. ∎

Remark 4.4 The dependence of $\text{VC}\,(\mathcal{F}_{n,q})$ on q in Lemma 4.3 is perhaps a somewhat surprising combinatorial fact, since there are only $2n$ free parameters $c_1, \ldots, c_n, r_1, \ldots, r_n$. Intuitively, the explanation for this dependence is that, although the number of free parameters is independent of q, the degree of the polynomial computed does depend on q, and this degree influences the number of distinct sign assignments that the polynomial can achieve. In general, the VC dimension of a concept class may be far larger than the number of free parameters, even infinite (cf. [21]), and is roughly equal to the square of the number of parameters for general classes of "neural network" classifiers (cf. [15]). As a related remark, observe that, as follows from a simple continuity argument, once that parameters have been found to achieve the shattering of a set of samples, any other set of samples near this set can also be shattered (using the same sets of parameters). In other words, one can always shatter an open set of samples (when viewing such sequences of samples as elements of an appropriate product Euclidean space) of cardinality equal to the VC dimension. One may ask about the shattering of more arbitrary sequences, for instance, the shattering of all sequences in "general position". In [23], a result is given which implies, in particular, that when there are $2n$ parameters it is impossible to shatter all general position sets of more than $4n + 2$ points. So the "dimension" obtained when one asks for shattering of *all sets in general position* (a concept studied also in [21], and related to Cover's capacity measures) is linearly proportional to the number of parameters. □

5 The Consistency Problem

We next briefly discuss polynomial time learnability of recurrent perceptron mappings. As discussed in e.g. [24], in order to formalize this problem we need to first choose a *data structure* to represent the hypotheses in $\mathcal{F}_{n,q}$. In addition, since we are dealing with complexity of computation involving real numbers, we must also clarify the meaning of "finding" a hypothesis, in terms of a suitable notion of polynomial-time computation. Once this is done, the problem becomes that of solving the *consistency problem*:

> Given a set of $s \geq s(\varepsilon, \delta)$ inputs $\xi_1, \xi_2, \ldots, \xi_s \in \mathbb{R}^{n+q}$, and an arbitrary dichotomy $\Delta : \{\xi_1, \xi_2, \ldots, \xi_s\} \to \{-1, 1\}$ find a representation of a hypothesis $\phi_{\vec{c}} \in \mathcal{F}_{n,q}$ such that the restriction of $\phi_{\vec{c}}$ to the set $\{\xi_1, \xi_2, \ldots, \xi_s\}$ is identical to the dichotomy Δ (or report that no such hypothesis exists).

The representation to be used should provide an *efficient encoding* of the values of the parameters $r_1, \ldots, r_n, c_1, \ldots, c_n$: given a set of inputs $(x_1, \ldots, x_{n+q}) \in \mathbb{R}^{n+q}$, one should be able to efficiently check concept membership (that is, compute sign $(\sum_{i=1}^{n+q} c_i x_i)$). Regarding the precise meaning of polynomial-time computation, there are at least two models of complexity possible. The first, the *unit cost model* of computation, is intended to capture the algebraic complexity of the problem; in that model, each arithmetic and comparison operation on two real numbers is assumed to take unit time, and finding a representation in polynomial time means doing so in time polynomial on $s + n + q$. An alternative, the *logarithmic cost model*, is closer to the notion of computation in the usual Turing machine sense; in this case one assumes that the inputs $(x_1, \ldots, x_{n+q})$ are rational numbers, with numerators and denominators of size at most L bits, and the time involved in finding a representation of $r_1, \ldots, r_n, c_1, \ldots, c_n$ is required to be polynomial on L as well.

We study the complexity of the learning problem for constant n (but varying q). The key step is treating consistency, since if the decision version of a consistency problem is NP-hard, then the corresponding class is not properly polynomially learnable under the complexity theoretic assumption RP≠NP, cf. [7]. For a suitable choice of representation, we will prove the following result:

Theorem 3 *For each fixed $n > 0$, the consistency problem for $\mathcal{F}_{n,q}$ can be solved in time polynomial in q and s in the unit cost model, and time polynomial in q, s, and L in the logarithmic cost model.*

Since $\text{VC}\,(\mathcal{F}_{n,q}) = O(n + n\log(q+1))$, it follows from here that the class $\mathcal{F}_{n,q}$ is learnable in time polynomial in q (and L in the log model). Our proof will consist of a simple application of several recent results and concepts, given in [4, 5, 20], which deal with the computational complexity aspects of the first-order theory of real-closed fields. Note that we do not study scaling with respect to n: for $q = 0$, this reduces to the still-open question of polynomial time solution of linear programming problems, in the unit cost model.

Proof of Theorem 3. For asymptotic results we may assume, without loss of generality, that $s > 2n$ from the bound of Theorem 1. We will use the representation discussed in the Appendix for the coefficients $c_1, \ldots, c_n, r_1, \ldots, r_n$, seen as vectors in $\mathbb{R}^k$, $k = 2n$. We first write the consistency problem as a problem of the following type:

($\star$) find some $c_1, \ldots, c_n, r_1, \ldots, r_n \in \mathbb{R}$ such that $\wedge_{i=1}^{s} \; (\mathcal{Q}_i \;\, \Delta_i \;\, 0)$ (or report that no such parameter values exist)

where each $\mathcal{Q}_i$ is a certain real polynomial in the variables $r_1, \ldots, r_n, c_1, \ldots, c_n$ of degree at most $q+1$, and Δ_i is the relation $>$ (resp. $\leq$) if $\delta(\xi_i) = 1$ (resp. $\delta(\xi_i) = -1$). Next, we determine all non-empty sign conditions of the set $\mathcal{Q} = \{\mathcal{Q}_1 \ldots \mathcal{Q}_s\}$. See Fact A.2 in the Appendix for an algorithm achieving this. For constant n, and this can be done in polynomial time in either the unit cost or the logarithmic cost model. Now, we check each non-empty sign condition to see if it corresponds to the given dichotomy Δ, i.e. if all the $(\mathcal{Q}_i \; \Delta_i \; 0)$ hold. If there is no match, we report a failure. Otherwise, we output the representation of the coefficients $c_1, \ldots, c_n, r_1, \ldots, r_n$.

∎

6 A Comment on Real-Valued Function Learning

As a final comment, we wish to simply remark that it is possible to obtain results on the learnability of linear systems dynamics, that is, the class of functions obtained if one does *not* take the sign when defining recurrent perceptrons. The connection between VC dimension and sample complexity is only meaningful for classes of Boolean functions; in order to obtain learnability results applicable to real-valued functions one needs metric entropy estimates for certain spaces of functions. These can be in turn bounded through the estimation of Pollard's pseudo-dimension. The reader is referred to [14] for the appropriate definitions and the results linking pseudo-dimension PD and learnability. One example result possible in our context is as follows. For any two nonnegative integers n, q, consider the class

$$\mathcal{F}'_{n,q} \;:=\; \left\{ \widehat{\phi}_{\vec{c}} \,\middle|\, \vec{c} \in \mathbb{R}^{n+q} \text{ is } n\text{-recursive} \right\}$$

where

$$\widehat{\phi}_{\vec{c}} \;:\; \mathbb{R}^{n+q} \to \mathbb{R} \;:\quad (x_1, \ldots, x_{n+q}) \;\mapsto\; \sum_{i=1}^{n+q} c_i x_i \;.$$

Assume that we wish to learn with respect to the loss function $\ell(y_1, y_2) = \max\{|y_1 - y_2|^2, 1\}$ and that $n + q \geq 4$. Then we have that

$$\text{PD}\left[\mathcal{F}'_{n,q}\right] \leq 20n \log(n+q) \,.$$

The proof follows easily from the Milnor-type bounds and the appropriate definitions.

A Appendix: Representations of Real Numbers and Decision Problems

We collect here some facts regarding Thom encodings of real numbers and their use in decision problems for real-closed fields.

Let $f(x)$ be a real univariate polynomial of degree d, and let α be a real root of f. The *Thom encoding* of α relative to $f(x)$, denoted $\mathrm{Th}\,(\alpha, f)$, or just $\mathrm{Th}\,(\alpha)$ if f is clear from the context, is the sign vector

$$\Big(\mathrm{sg}[f(\alpha)], \mathrm{sg}[f'(\alpha)], \ldots, \mathrm{sg}[f^{(d)}(\alpha)]\Big) \;\in\; \{-1, 0, 1\}^{d+1}$$

where $\mathrm{sg}[x] = x/|x|$ if $x \neq 0$ and $\mathrm{sg}[0] = 0$. It is known (cf. [8]) that $\mathrm{Th}\,(\alpha, f)$ uniquely characterizes α among the roots of f.

In this paper, by a *representation* of a vector $(y_1, y_2, \ldots, y_k) \in \mathbb{R}^k$ we mean a vector

$$(f(t), g_0(t), \ldots, g_k(t), \rho)$$

consisting of:

(a) a univariate polynomial $f(t)$,

(b) $k+1$ univariate polynomials $g_0(t), \ldots, g_k(t)$, and

(c) a vector $\rho \in \{-1, 0, 1\}^{\deg(f)+1}$,

so that ρ is the Thom encoding $\mathrm{Th}\,(\alpha)$ of some root α of f, and $y_i = \frac{g_i(\alpha)}{g_0(\alpha)}$ for each $1 \leq i \leq k$. The polynomials are represented by vectors providing their degrees and listing all coefficients. When dealing with the logarithmic cost model, we assume in addition that the coefficients of the polynomials f and g_i are all rational numbers. In the unit cost model, the *size* of such a representation is defined to be the total number of reals needed so as to specify the coefficients, that is, the sum of the degrees of all the polynomials plus $k + 3 + \deg(f)$. In the logarithmic cost model, the size is the above plus the total number of bits needed in order to represent the coefficients of the polynomials, each written in binary as the quotient of two integers.

In the paper, we use these representations for the parameters defining concepts, while inputs are given directly as real numbers (rationals in the log model); thus we need to know that signs of polynomial expressions involving vectors represented in the above manner as well as reals can be evaluated efficiently. We next state a result that assures this. By the *complexity* of a multi-variable polynomial $H(z_1, \ldots, z_q)$ we mean the sum of the number of nonzero monomials plus the sum of the total degrees of all these monomials (for instance, $2z_1^2z_2^3 - z_1^7$ has complexity $2 + 5 + 7 = 14$); in the log cost model, we assume that the coefficients of H are rational and we add the number of bits needed to represent the coefficients.

Lemma A.1 In the unit cost model, there is an algorithm $\mathcal{A}$ which, given a polynomial H of complexity h on variables $x_1, \ldots, x_l, y_1, \ldots, y_k$, and given real numbers $x_1, \ldots, x_l$ and a representation $(f(t), g_0(t), \ldots, g_k(t), \rho)$ of a vector $y_1, \ldots, y_k$, can compute $\mathrm{sg}[H(x_1, \ldots, x_l, y_1, \ldots, y_k)]$ in time polynomial on l, h, and the size of this representation. The same result holds in the logarithmic cost model, assuming that the inputs x_i are all rational, with time now polynomial on the size of these inputs as well. □

Proof. Note that, in general, if $p_1(t)$ and $p_2(t)$ are two rational functions with numerator and denominators of degree bounded by d, then both $p_1(t)p_2(t)$ and $p_1(t)+p_2(t)$ are rational functions with numerator and denominator of degree at most $2d$. Moreover, these algebraic operations can be computed in time polynomial on d as well as, in the log model, on the size of coefficients. Working iteratively on all monomials of H, we conclude that it is possible to construct from the g_i's and x_j's, in polynomial time, two polynomials $R_1(t)$ and $R_2(t)$ with real (rational, in the log model) coefficients so that $H(x_1, \ldots, x_l, y_1, \ldots, y_k) = R_1(\alpha)/R_2(\alpha)$, where α is the root encoded by ρ. Note that

$$\mathrm{sign}\left(\frac{R_1(\alpha)}{R_2(\alpha)}\right) = \begin{cases} 1 & \text{if } \mathrm{sign}\,(R_1(\alpha)) = \mathrm{sign}\,(R_2(\alpha)) \text{ and } R_1(\alpha) \neq 0 \\ -1 & \text{otherwise} \end{cases}$$

Thus it is only necessary to evaluate $\mathrm{sign}\,(R_i(\alpha))$, $i = 1, 2$. The evaluation can be done efficiently because of the following fact from [20]:

> There is an algorithm $\mathcal{B}$ with the following property. Given any univariate real polynomial $f(t)$, a real root α of f specified by means of its Thom encoding $\mathrm{Th}\,(\alpha)$, and another univariate polynomial $g(t)$, $\mathcal{B}$ outputs $\mathrm{sign}\,(g(\alpha))$, using a number of arithmetic operations polynomial on $\deg(f) + \deg(g)$; in the logarithmic cost model, if all input coefficients are rationals of size at most L, then $\mathcal{B}$ uses a number of bit operations polynomial on $\deg(f) + \deg(g) + L$.

This provides the desired $\mathrm{sg}[H(x_1, \ldots, x_l, y_1, \ldots, y_k)]$. ∎

The main reason that representations of the type $(f(t), g_0(t), \ldots, g_k(t), \rho)$ are of interest is that one can produce solutions of algebraic equations and inequalities represented in that form. We explain this next.

One says that a vector $\sigma = (\sigma_1, \sigma_2, \ldots, \sigma_s) \in \{-1, 0, +1\}^s$ is a *nonempty sign condition* for an ordered set of s real polynomials $\mathcal{P} = \{\mathcal{P}_1, \mathcal{P}_2, \ldots, \mathcal{P}_s\}$ in $k < s$ real variables if there exists some point $(y_1, \ldots, y_k) \in \mathbb{R}^k$ such that $\sigma_i = \mathrm{sg}[\mathcal{P}_i(y_1, y_2, \ldots, y_k)]$ for all i; the corresponding point $(y_1, y_2, \ldots, y_k) \in \mathbb{R}^k$ is said to be a *witness* of σ.

Fact A.2 ([4, 5]) There is an algorithm $\mathcal{A}$ as follows. Given any set $\mathcal{P}$ of s real polynomials in $k < s$ variables, where each polynomial is of degree at most d, $\mathcal{A}$ computes, for each nonempty sign-condition of $\mathcal{P}$, the sign condition σ as well as a representation of a witness for σ. Moreover, $\mathcal{A}$ runs in $O((sd)^{O(k)})$ time in the unit cost model, and in the corresponding representation, $\deg(f) \leq (sd)^{O(k)}$. In the logarithmic cost model, assuming that coefficients of the given polynomials are rationals of size at most L, $\mathcal{A}$ runs in time $O(s^k d^{O(k)} L^{O(1)})$, and the degrees and coefficients of all the polynomials $f, g_0, \ldots, g_k$ (and, consequently the number of components in $\mathrm{Th}\,(\alpha)$) are rational numbers of size at most $O(d^{O(k)} L^{O(1)})$. □

References

[1] A.D. Back and A.C. Tsoi, "FIR and IIR synapses, a new neural network architecture for time-series modeling", in *Neural Computation*, vol. 3, pp. 375–385, 1991.

[2] A.D. Back and A.C. Tsoi, "A comparison of discrete-time operator models for nonlinear system identification", in *Advances in Neural Information Processing Systems (NIPS'94)*, Morgan Kaufmann Publishers, 1995, to appear.

[3] A.M. Baksho, S. Dasgupta, J.S. Garnett, and C.R. Johnson, “On the similarity of conditions for an open-eye channel and for signed filtered error adaptive filter stability”, in *Proc. IEEE Conf. Decision and Control*, Brighton, UK, Dec. 1991, IEEE Publications, 1991, pp. 1786–1787.

[4] S. Basu, R. Pollack, and M.-F. Roy, “A New Algorithm to Find a Point in Every Cell Defined by a Family of Polynomials”, in *Quantifier Elimination and Cylindrical Algebraic Decomposition*, B. Caviness and J. Johnson eds., Springer-Verlag, to appear.

[5] S. Basu, R. Pollack, and M.-F. Roy, “On the Combinatorial and Algebraic Complexity of Quantifier Elimination”, in *Proc. 35th IEEE Symp. on Foundations of Computer Science*, 1994.

[6] Y. Bengio, *Neural Networks for Speech and Sequence Recognition*, Thompson Computer Press, Boston, 1996.

[7] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth, “Learnability and the Vapnik-Chervonenkis dimension”, *J. of the ACM*, vol. 36, pp. 929-965, 1989.

[8] M. Coste and M.F. Roy, “Thom’s Lemma, the Coding of Real Algebraic Numbers and the Computation of the Topology of Semi-algebraic sets”, *J. Symbolic Computation*, vol. 5, pp. 121-129, 1988.

[9] D.F. Delchamps, “Extracting State Information from a Quantized Output Record”, *Systems and Control Letters*, vol. 13, pp. 365-372, 1989.

[10] R.O. Duda and P.E. Hart, *Pattern Classification and Scene Analysis*, Wiley, New York, 1973.

[11] R.M. Dudley, *A Course on Empirical Processes*, École d’été de probabilités de Saint-Flour, XII—1982, 1–142, Lecture Notes in Math., 1097, Springer, Berlin-New York, 1984.

[12] C.E. Giles, G.Z. Sun, H.H. Chen, Y.C. Lee, and D. Chen, “Higher order recurrent networks and grammatical inference”, in *Advances in Neural Information Processing Systems 2,* D.S. Touretzky (ed.), Morgan Kaufmann, San Mateo, CA, 1990.

[13] P. Goldberg and M. Jerrum, “Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers”, *Machine Learning*, vol. 18, pp. 131-148, 1995.

[14] D. Haussler, “Decision theoretic generalizations of the PAC model for neural nets and other learning applications”, *Information and Computation*, vol. 100, pp. 78-150, 1992.

[15] P. Koiran, P and E.D. Sontag, “Neural networks with quadratic VC dimension,” *J. Computer Systems Sciences*, to appear. (Summarized version in *Advances in Neural Information Processing Systems* (NIPS95), to appear.)

[16] R. Koplon and E.D. Sontag, “Linear systems with sign-observations”, *SIAM J. Control and Optimization*, vol. 31, pp. 1245-1266, 1993.

[17] W. Maass, “Perspectives of current research about the complexity of learning in neural nets”, in *Theoretical Advances in Neural Computation and Learning*, V.P. Roychowdhury, K.Y. Siu, and A. Orlitsky (eds.), Kluwer Acedemic Publishers, pp. 295-336, 1994.

[18] D. Pollard, *Empirical Processes: Theory and Applications*, NSF-CBMS Regional Conf. Series in Probability and Statistics, vol.2, 1990, American Statistical Association, Alexandria, VA, 1990.

[19] G.W. Pulford, R.A. Kennedy, and B.D.O. Anderson, “Neural network structure for emulating decision feedback equalizers”, in *Proc. Int. Conf. Acoustics, Speech, and Signal Processing*, Toronto, Canada, pp. 1517-1520, May 1991.

[20] M.-F. Roy and A. Szpirglas, “Complexity of Computation on Real Algebraic Numbers”, *J. Symbolic Computation*, vol. 10, pp. 39-51, 1990.

[21] E.D. Sontag, “Feedforward nets for interpolation and classification,” *J. Comp. Syst. Sci.* **45**(1992): 20-48.

[22] E.D. Sontag, “Neural networks for control”, in *Essays on Control: Perspectives in the Theory and its Applications* (H.L. Trentelman and J.C. Willems, eds.), Birkhauser, Boston, pp. 339-380, 1993.

[23] E.D. Sontag, “Shattering all sets of k points in general position requires (k-1)/2 parameters,” SYCON (Rutgers Center for Systems and Control) Report 96-01, February 1996[¶]. Submitted for publication.

[24] György Turán, “Computational Learning Theory and Neural Networks: A Survey of Selected Topics”, in *Theoretical Advances in Neural Computation and Learning*, V.P. Roychowdhury, K.Y. Siu, and A. Orlitsky (eds.), Kluwer Academic Publishers, pp. 243-293, 1994.

[25] L.G. Valiant “A theory of the learnable”, *Comm. of the ACM*, vol. 27, pp. 1134-1142, 1984.

[26] V.N. Vapnik and A.Ja. Chervonenkis, *Theory of Pattern Recognition (in Russian)*, Moscow, Nauka, 1974. (German translation: W.N. Wapnik and A.Ja. Chervonenkis, *Theorie der Zeichenerkennung*, Berlin, Akademia-Verlag, 1979).

[27] V.N. Vapnik, *Estimation of Dependencies Based on Empirical Data*, Springer, Berlin, 1982.

[28] M. Vidyasagar, *Learning and Generalization with Applications to Neural Networks*, Springer, London, 1996, to appear.

[¶] available via the WWW at **http://www.math.rutgers.edu/~sontag**