VC Dimension and Sampling Complexity of Learning Sparse Polynomials and Rational Functions

Marek Karpinski
Thorsten Werther
TR-91-044
August, 1991

Abstract

This paper presents the recent results on the VC dimension and the sampling complexity of learning sparse polynomials and rational functions. Some of the direct applications of these results have also been presented.

VC Dimension and Sampling Complexity of Learning Sparse Polynomials and Rational Functions

Marek Karpinski * Thorsten Werther †

Abstract

This paper presents the recent results on the VC dimension and the sampling complexity of learning sparse polynomials and rational functions. Some of the direct applications of these results have been also presented.

1 Introduction

This paper presents the recent results (cf. also [KW 89], [We 91]) on the VC dimension and the learnability of sparse univariate polynomials over the real numbers. The framework is the model of probably approximately correct (pac) learning concepts from examples introduced by Valiant ([Val 84]). The results presented in this paper crucially depend on the connection between pac learnability and the Vapnik-Chervonenkis (VC) dimension.

We derive linear upper $(4t-1)$ and lower $(3t)$ bounds on the VC dimension of the class of t-sparse polynomials over the real numbers implying uniform pac learnability of this function class. We transfer these results to sparse rational functions and sparse, degree-bounded polynomials. The results generalize to uniform distribution-free learnability of sparse polynomials even in the extended metric model of Haussler ([Ha 89]). Applying this result, we solve Vapnik's open problem on uniform estimation of the polynomial regression function ([Vap 82]).

The interest in the computational complexity in learning sparse polynomials and rational functions has several motivations. The first motivation is the growing interest in sparse polynomials from the complexity theoretic point of view. The complexity analysis of algorithms manipulating polynomials usually measures the input length in terms of the degree of polynomials. For polynomials with a small number of terms, this model is not reasonable since sparse polynomials are usually represented by a list of non-zero coefficients and the corresponding exponents. Hence, the natural measure of the size of a polynomial is given in terms of its sparsity when applying the uniform cost model of computation (cf. [AHU 74]). Recent results indicate also that sparse polynomials play a key role in the harmonic analysis of Boolean circuits ([Br 90, BS 90]) and, surprisingly, in the area of learnability of Boolean functions as well ([KM 91]).

*Department of Computer Science, University of Bonn, and International Computer Science Institute, Berkeley, California. Supported in part by Leibniz Center for Research in Computer Science, by the DFG Grant KA 673/4-1, and by the SERC Grant GR-E 68297.

†Department of Computer Science, University of Bonn.

The second motivation are the issues of sparse polynomial interpolation. In the black box model, a learning algorithm for sparse polynomials has access to an oracle which gives the value of the polynomial for an arbitrary evaluation point ([Kar 89]). Grigoriev and Karpinski ([GK 87]), Ben-Or and Tiwari ([BeTi 88]) and Grigoriev et al. ([GKS 90]) show that in this oracle model there are efficient algorithms for exact learning (interpolation) of sparse determinants ([GK 87]), sparse polynomials over fields of characteristic zero ([BeTi 88]) and also over finite fields ([GKS 90]).

It is known (cf. [WD 81, Fl 89]) that elements of vector spaces of real valued functions are pac learnable. Therefore, the third motivation of this work is to explore the learnability of sparse polynomials as a function class which is not embedded in some vector space.

2 Previous Work

Throughout this paper, we employ the model of machine learning introduced by Valiant ([Val 84]), usually referred to as the PAC model. Valiant used this new model of distribution-free learning from examples to exhibit and analyze several learning algorithms for Boolean functions.

In this section, we review the definition of Valiant's distribution-free model of learning. This model was extended by Blumer et al. ([BEHW 86], [BEHW 89]) to classes of concepts defined by regions in Euclidean space E^n. Blumer et al. applied results of the pioneering work of Vapnik and Chervonenkis ([VC 71]) on the uniform convergence of empirical dependences to provide the necessary and sufficient conditions for feasible learnability.

The reader is referred for the basic notations of set theory, probability theory, and stochastic processes to [Coh 82, Po 84, Vap 82]. For a set X, 2^X will denote the set of all subsets of X. Whenever we talk of an unknown probability distribution P on X, P is assumed to be arbitrary but fixed. For a collection $C \subseteq 2^X$ of subsets of X, we assume each member of C to form a Borel set, such that unions, intersections, sequences, and limits of events result also in events, i.e. probabilities are assigned to them by any probability distribution P on X. This assumption is of crucial importance for the connection of learnability and convergence of stochastic processes.

2.1 The PAC Model and the Vapnik-Chervonenkis Dimension

There are many approaches proposed in the literature, especially from the area of artificial intelligence, to formalize the notation of *learning* and to give it a precise meaning. This paper explores the learnability of sparse polynomials in the framework of Valiant's model of *distribution-free learning from examples.*

In Valiant's probabilistic model of learning (concepts) from examples, each concept c from a non-empty class $\mathcal{C} \subseteq 2^X$ is a subset of a given instance space X (for example, X might be $\{0,1\}^n$ ([Val 84]) or n-dimensional Euclidean space E^n ([BEHW 86])). The unknown target concept t to be learned is assumed to be a member of the class $\mathcal{C}$.

In this model of learning, we assume a fixed but arbitrary (and unknown) probability distribution P defined on X. It is assumed that a learning algorithm has access to a finite set of *examples* of

the unknown target concept t. Each example (x, c) consists of an instance $x \in X$, which is drawn independently according to P, and its classification $c \in \{1, 0\}$ as either a positive instance ($x \in t$) or a negative instance ($x \notin t$). This set is called a *sample* of the target concept.

In the PAC (Probably Approximately Correct) model, a *learning function* for $\mathcal{C}$ is a function that, given a large enough randomly drawn sample, returns a *hypothesis* which is, with high probability, a good approximation (with respect to P) to the target concept, no matter which concept from $\mathcal{C}$ we are trying to learn. The error of the hypothesis is the probability that the hypothesis disagrees with the target on a (with respect to P) randomly drawn example.

We formalize this using the notation from [BEHW 89]:

Let $\mathcal{C} \subseteq 2^X$ be a non-empty class of concepts and $c \in \mathcal{C}$ is a Borel set. For $x = (x_1, \ldots, x_m) \in X^m$, $m \geq 1$, the m-sample of $c \in \mathcal{C}$ generated by x is given by

$$\text{sam}_c(x) = (< x_1, I_c(x_1) >, \ldots, < x_m, I_c(x_m) >),$$

where I_c is the $\{0, 1\}$-valued indicator function for c, that is $I_c(x_i) = 1$ iff $x_i \in c$. The sample space of $\mathcal{C}$, denoted $S_{\mathcal{C}}$, is the set of all m-samples over all $c \in \mathcal{C}$ and all $x \in X^m$, for all $m \geq 1$.

The learning algorithm takes an m-sample as an input and produces a hypothesis from some hypothesis space H. Usually, the hypothesis space is $\mathcal{C}$ itself, but in some cases it is preferable to approximate concepts from $\mathcal{C}$ in a different class H.

Let $H \subseteq 2^X$ now be a set of Borel sets, called the hypothesis space. Let $A_{\mathcal{C},H}$ denote the set of all functions that map the sample space $S_{\mathcal{C}}$ to the hypothesis space H.
$A \in A_{\mathcal{C},H}$ is called consistent if, for each sample $s = (< x_1, a_1 >, \ldots, < x_m, a_m >)$, the hypothesis produced by A, $h = A(s)$, agrees with s, that is $a_i = I_h(x_i)$ for all $1 \leq i \leq m$.

In the PAC model, the sample is generated according to an unknown, but fixed probability distribution P on X. The error rate of a hypothesis $h \in H$ (with respect to the target concept $t \in \mathcal{C}$ and P) is the probability that h and t classify a randomly drawn example differently, which is $P(h \oplus t)$, the probability of the symmetric difference of h and t.

Let $\mathcal{C} \subseteq 2^X$ be a non-empty class of concepts and let $H \subseteq 2^X$ be a hypothesis space. For $0 < \epsilon, \delta < 1$, let $m(\epsilon, \delta)$ be an integer-valued function of ϵ and δ. Let P be a probability distribution on X.

We say $\mathcal{C}$ is *uniformly learnable by H under the distribution P* if there is (a learning function) $A \in A_{\mathcal{C},H}$ such that for a randomly drawn sample of size $m(\epsilon, \delta)$ of any target concept in $\mathcal{C}$, A produces, with probability at least $1 - \delta$, a hypothesis in H with error rate no more than ϵ.

If there exists $A \in A_{\mathcal{C},H}$ such that A is a learning function for C with sample size $m(\epsilon, \delta)$ for all probability distributions P on X, $\mathcal{C}$ *is uniformly learnable by* H. The smallest sample size $m(\epsilon, \delta)$ is called the *sample complexity* of A.

Note that this general definition of uniform learnability imposes no restrictions of feasibility or even computability of the learning function.

For finite concept classes $\mathcal{C} \subseteq 2^X$, Vapnik ([Vap 82]) gave an upper bound on the sample complexity for any uniform learning algorithm. For infinite classes, such as geometric concept classes

on E^n, there was no general characterization of uniform learnability known, until Blumer et al. ([BEHW 86]) employed ideas from Vapnik and Chervonenkis ([VC 71]) to show that the essential condition for distribution-free learnability is finiteness of a combinatorical parameter of the concept class $\mathcal{C}$, called the Vapnik-Chervonenkis dimension.

Let $\mathcal{C} \subseteq 2^X$ be a concept class on X. For any finite set $F \subseteq X$, let $\Pi_{\mathcal{C}}(F) = \{c \cap F | c \in \mathcal{C}\}$ denote the restriction of $\mathcal{C}$ to the set F. If $\Pi_{\mathcal{C}}(F) = 2^F$, then the set F is *shattered* by $\mathcal{C}$. In other words, each subset of F is of the form $c \cap F$ for some $c \in \mathcal{C}$. The *Vapnik-Chervonenkis dimension* (VC dimension) of the class $\mathcal{C}$ is the largest integer d such that some $S \subseteq X$ of size d is shattered by the class $\mathcal{C}$. If arbitrary large subsets of X are shattered by the class $\mathcal{C}$, then the VC dimension of $\mathcal{C}$ is infinite. The class consisting of one concept is of VC dimension 0, and, by convention, the empty class is of VC dimension -1. Let $\mathrm{VCdim}(\mathcal{C})$ denote the VC dimension of $\mathcal{C}$.

Vapnik and Chervonenkis ([VC 71]) give necessary and sufficient conditions for the uniform convergence of the empirical risk functional to the expected risk functional in terms of the VC dimension. Their work has been extended to handle much more general situations ([Po 84]). Blumer et al. ([BEHW 86]) were the first to draw the connection between distribution-free learning and the VC dimension.

To avoid measurability difficulties in Theorem 1, Blumer et al. assume that the concept class is *well-behaved* which is a relatively benign measure-theoretic condition. It is not likely to exclude any concept class considered in the context of machine learning applications. For common use, it is sufficient to show that a concept class is *universally separable* (cf. [Du 78]). The well-behavior of the concept class follows from this.

Theorem 1 [BEHW 86] Let $\mathcal{C}$ be a non-trivial, well-behaved concept class. $\mathcal{C}$ is uniformly learnable iff the VC dimension of $\mathcal{C}$ is finite.

Blumer et al. and Ehrenfeucht et al. give upper and lower bounds on the sample complexity $m(\epsilon, \delta)$ for distribution-free learning of a concept class $\mathcal{C}$ of finite VC dimension $d < \infty$.

Theorem 2 [BEHW 86] Let $\mathcal{C}$ be a non-trivial, well-behaved concept class of VC dimension $d < \infty$. Then, for $0 < \epsilon, \delta < 1$ and sample size at least

$$\max\left(\frac{4}{\epsilon}\log\frac{2}{\delta}, \frac{8d}{\epsilon}\log\frac{13}{\epsilon}\right),$$

any consistent function $A : S_{\mathcal{C}} \to \mathcal{C}$ is an (ϵ, δ)-learning function for $\mathcal{C}$.

Blumer et al. also give a lower bound on the sample size. This bound was improved by Ehrenfeucht et al.

Theorem 3 [EHKL 89] Let $\mathcal{C}$ be a concept class of VC dimension $d > 1$. Then, for $0 < \epsilon \leq \frac{1}{8}, 0 < \delta \leq \frac{1}{100}$, any (ϵ, δ)-learning function A for $\mathcal{C}$ must use sample size

$$m(\epsilon, \delta) \geq \max\left(\frac{1-\epsilon}{\epsilon}\log\frac{1}{\delta}, \frac{d-1}{32\epsilon}\right).$$

2.2 The Extended PAC Model

One of the shortcomings of the standard PAC model is that it is only defined for $\{0,1\}$-valued functions. Haussler ([Ha 89]) proposes a generalization of the PAC model for distribution-free learning of functions that take values in an arbitrary metric space. This is of particular interest when learning real-valued functions. Haussler generalizes the notation of VC dimension and shows that, similar to the standard model, small VC dimension implies fast uniform distribution-independent convergence. We recall some of his results briefly.

Let $\mathcal{F}$ (the hypothesis space) be a family of functions from a domain X to a set Y with metric d_Y. Let $\mathcal{D}$ be a family of probability distributions on $S = (X \times Y)$. A pair $(x,y) \in S$ is called an example, a sequence of examples is called a sample. A learning problem P is stated as follows: Given a sample chosen independently at random with respect to some unknown distribution $D \in \mathcal{D}$, find a hypothesis h from $\mathcal{F}$ that is close to the target t from $\mathcal{F}$, where the target is the element from $\mathcal{F}$ which is closest to the sample.

More formally: Let $\mathrm{er}_D(f)$ be the expectation (with respect to D) of $d_Y(f(x),y)$ when (x,y) is drawn at random from S (with respect to D). Let $\mathrm{opt}(D,\mathcal{F})$ be the infimum of $\mathrm{er}_D(f)$ over all $f \in \mathcal{F}$. Let $d_\nu(r,s) = \frac{|r-s|}{\nu+r+s}$, for $r,s,\nu \in \mathbf{R}^+$, be a metric on $\mathbf{R}^+$.

The set $\mathcal{F}$ is called *uniformly learnable* ([Ha 89]) iff there exists a function L from the set of all samples into $\mathcal{F}$ such that for all $\nu > 0, 0 < \alpha < 1, 0 < \delta < 1$ there exists a finite sample size $m = m(\nu,\alpha,\delta)$ such that for all $D \in \mathcal{D}$ and samples s of size m the (learning) function L produces with probability at least $1-\delta$ a hypothesis that is acceptably close to the optimal hypothesis in $\mathcal{F}$, that is

$$d_\nu(\mathrm{er}_D(L(s), \mathrm{opt}(D,\mathcal{F}))) \leq \alpha.$$

Similar to the results of Blumer et al. ([BEHW 86]), Haussler shows that the essential condition for distribution-free uniform learnability of $\mathcal{F}$ is the finiteness of the "VC dimension of the graphs of functions in $\mathcal{F}$" which is an extension of the standard VC dimension:

For each $f \in \mathcal{F}$, we denote by $I(f)$ the function from $X \times Y \times \mathbf{R}^+$ into $\{0,1\}$ defined by

$$I(f)(x,y,\epsilon) = \begin{cases} 1 & \text{if } d_Y(f(x),y) \leq \epsilon \\ 0 & \text{otherwise} \end{cases}.$$

Let $I(\mathcal{F}) = \{I(f) \mid f \in \mathcal{F}\}$. We define the *metric VC dimension* of $\mathcal{F}$ as the (standard) VC dimension of $I(\mathcal{F})$. Let m-VCdim$(\mathcal{F})$ denote the metric VC dimension of $\mathcal{F}$.

In Section 3, we investigate learnability of sparse real polynomials in this generalized model. In this case, $Y = \mathbf{R}$ and $d_Y(x,y) = |x-y|$, and the notation of metric VC dimension is similar to the notation of VC dimension of real-valued functions given in [Po 84] and [Vap 89].

Haussler gives bounds on the sample size required for uniform learnability in this generalized model depending on the metric VC dimension of $\mathcal{F}$ and on the metric dimension of the metric space (Y,d_Y). This bounds reduce to the bounds given in Theorem 2 for the standard PAC model.

3 Learnability of Sparse Polynomials

In this section, we prove uniform learnability of sparse univariate polynomials over the real numbers in the standard PAC model as well as in the generalized model defined by Haussler (cf. Section 2.2).

We prove upper and lower bounds on the VC dimension of sparse polynomials, and apply results of Blumer et al. to derive bounds on the sample size required for uniform learning.

Combining these bounds with the results for degree-bounded polynomials (cf. Section 3.2), we derive bounds on the VC dimension of the class of sparse and degree-bounded polynomials.

3.1 Notation

Let $C \subseteq 2^X$ be a concept class on X. For a sample $s = (< x_1, a_1 >, \ldots, < x_m, a_m >) \in S_C$, we call the vector $a = (a_1, \ldots, a_m) \in \{0,1\}^m$ the *labeling* of s. A concept $c \in C$ is said to satisfy the labeling a on $(x_1, \ldots, x_m)$ if c is consistent with the sample s.

Let $\mathcal{F}$ be a collection of real-valued functions on a set X. We investigate learnability of the concept class $\mathrm{pos}(f_0 - \mathcal{F})$ defined as the collection of all concepts

$$\mathrm{pos}(f_0 - f) = \{x \in X \mid f_0(x) - f(x) > 0\},$$

for $f \in \mathcal{F}$ and $f_0 \notin \mathcal{F}$ an arbitrary real function on X.

For each $t \in \mathbf{N}$, let $\mathcal{P}_t \subset \mathbf{R}[x]$ denote the set of t-sparse univariate polynomials over the real numbers, i.e, for each $p \in \mathcal{P}_t$, the number of non-zero coefficients in the expansion of p is bounded by t. Let $\mathcal{P}_t^+ \subset \mathbf{R}[x]$ denote the set of t-sparse polynomials where the domain is restricted to $\mathbf{R}^+$. We identify the VC dimension of $\mathcal{P}_t$ ($\mathcal{P}_t^+$) with the VC dimension of the concept class $\mathrm{pos}(y - \mathcal{P}_t) \subset \mathbf{R}^2$ ($\mathrm{pos}(y - \mathcal{P}_t^+) \subset \mathbf{R}^+ \times \mathbf{R}$).

3.2 Learnability of Degree-bounded Polynomials

In this section, we briefly survey results on the learnability of regions defined by elements of vector spaces of real-valued functions . Real polynomials of bounded degree fit in this setting as a special case.

Let $X = \mathbf{R}^2$, and consider the set $\mathcal{P}_n \subset \mathbf{R}[x]$ of univariate polynomials of degree less than n. Let $f_0(x, y) = y$. For $p \in \mathcal{P}_n$, the concept $\mathrm{pos}(f_0 - p) = \{(x, y) \in \mathbf{R}^2 \mid y > p(x)\}$ consists of all points in the plane that lie "above" the graph of p. It is simple to see that the VC dimension of $C = \mathrm{pos}(f_0 - \mathcal{P}_n)$ equals n. First, any subset $S \subset X$ of size n is shattered by C since a satisfying polynomial from $\mathcal{P}_n$ can be retrieved via interpolation for each labeling in $\{0,1\}^n$. Assume that a set $R \subset X$ of size $n+1$ is shattered by C_n. Then there are polynomials $p_1, p_2 \in \mathcal{P}_n$, $p_1 \not\equiv p_2$ satisfying the two alternating labelings $\sigma_1 = (1,0,1,0,\ldots)$ and $\sigma_2 = (0,1,0,1,\ldots)$ of size $n+1$. Hence, there are at least n points with $p_1(x) = p_2(x)$.This implies $p_1 \equiv p_2$. From Theorem 1, the class of polynomials of degree less than n is uniformly learnable for each fixed $n > 0$.

This result holds in the much more general case of vector spaces of real-valued functions. Wenocur and Dudley ([WD 81]) extended a result of Cover ([Cov 65]) and proved that the VC dimension equals the dimension of the vector space.

Theorem 4 [WD 81] Let $\mathcal{F}$ be an m-dimensional vector space of real functions on a set X. Let $f_0 \notin \mathcal{F}$ be a real function on X. Then the VC dimension of $\text{pos}(f_0 - \mathcal{F})$ equals m.

3.3 Bounds on the VC dimension of $\mathcal{P}_t$

We show that the VC dimension of $\mathcal{P}_t$ is linear in t. For $\mathcal{P}_t^+$, we determine its VC dimension exactly.

3.3.1 Lower Bounds

We start with a lower bound on the VC dimension of $\mathcal{P}_1$.

Lemma 5 The VC dimension of $\mathcal{P}_1$ is bounded from below by 3.

PROOF: We show that for each labeling $\sigma \in \{0,1\}^3$ there is a 1-sparse polynomial f_σ satisfying σ on the set $S = \{(-3,4),(1,2),(7,6)\}$ of size 3. Choose, for example, $f_{000} = 7$, $f_{001} = 5$, $f_{010} = x^2$, $f_{011} = -2x$, $f_{100} = 3x$, $f_{101} = 3$, $f_{110} = x$ and $f_{111} = 1$ (cf. Figure 1). Note that the VC dimension of $\mathcal{P}_t^+$ is at least 2. □

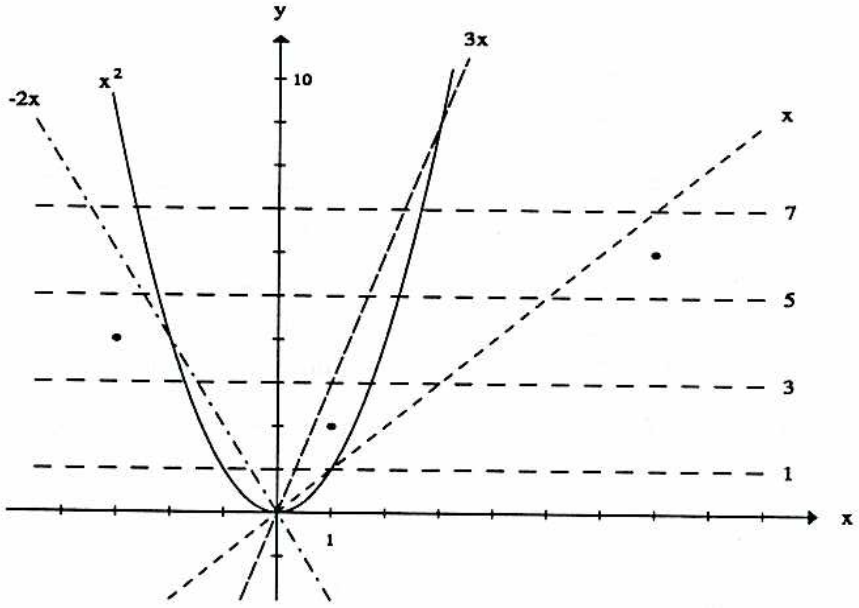

Figure 1: Monomials shattering the set S of size 3

In the following proofs, it will be convenient to assume that no element of a set S, which is shattered by some set of sparse polynomials, lies on the graph of these polynomials.

Remark 6 Let a set S of size d be shattered by the class of t-sparse polynomials. Then there are a set $Z = \{(x_i, y_i)\}_{i=1,\ldots,d}$ and constants $\epsilon_i > 0, i = 1, \ldots, d$ such that every set $S' = \{(\bar{x}_i, \bar{y}_i)\}_{i=1,\ldots,d}$ with $|(\bar{x}_i, \bar{y}_i) - (x_i, y_i)| \leq \epsilon_i$ is shattered by t-sparse polynomials.

PROOF: For each $\sigma \in \{0,1\}^d$, there is a t-sparse polynomial f_σ satisfying σ on S. For $i = 1, \ldots, d$, we define the regions

$$M_i = \{\, (x,y) \,|\, \forall \sigma \in \{0,1\}^d : \left\{ \begin{array}{ll} y > f_\sigma & \text{if } \sigma(i) = 1 \\ y < f_\sigma & \text{if } \sigma(i) = 0 \end{array} \right\} \}.$$

Since S is shattered by $\{f_\sigma\}_{\sigma \in \{0,1\}^d}$, there exists a point (x_i, y_i) and a constant $\epsilon_i > 0$ such that the ball

$$B_{\epsilon_i}(x_i, y_i) = \{\, (x,y) \,|\, |(x,y) - (x_i, y_i)| \leq \epsilon_i \,\}$$

is a proper subset of M_i. Hence, each set S' defined as above is shattered by the t-sparse polynomials $\{f_\sigma\}_{\sigma \in \{0,1\}^d}$. □

Lemma 7 states that the VC dimension of sparse polynomials is subadditive. We use this lemma to derive a lower bound.

Given a set shattered by t_1-sparse polynomials and a set shattered by t_2-sparse polynomials, we construct a set shattered by $(t_1 + t_2)$-sparse polynomials.

Lemma 7 For $t_1, t_2 \in \mathbf{N}$, let d_1, d_2 denote the VC dimension of $\mathcal{P}_{t_1}$, $\mathcal{P}_{t_2}$ respectively. Then the VC dimension of $\mathcal{P}_{t_1+t_2}$ is at least $d_1 + d_2$.

PROOF: Let S_1 and S_2 denote some sets of points of size d_1, d_2 respectively, shattered by t_1-sparse polynomials, t_2-sparse polynomials respectively.

Let $S_1 = \{(x_i^{(1)}, y_i^{(1)})\}_{i=1,\ldots,d_1}$ and $S_2 = \{(x_j^{(2)}, y_j^{(2)})\}_{j=1,\ldots,d_2}$. For a labeling $\sigma^{(1)} \in \{0,1\}^{d_1}$ let $f_{\sigma^{(1)}}$ satisfy $\sigma^{(1)}$ on S_1, and for a labeling $\sigma^{(2)} \in \{0,1\}^{d_2}$ let $g_{\sigma^{(2)}}$ satisfy $\sigma^{(2)}$ on S_2.

In order to show that VC dimension of $\mathcal{P}_{t_1+t_2} \geq d_1 + d_2$, we modify the sets S_1 and S_2 (and the corresponding polynomials shattering S_1 and S_2) such that the union of these modified sets is shattered by polynomials derived by adding some of the modified polynomials.

First, we pull the sets S_1 and S_2 apart such that the absolute values of the x-coordinates of points in S_1 are at most $\frac{1}{2}$ and the absolute values of the x-coordinates of points in S_2 are at least 2.

Let

$$c_1 > 2 \cdot \max_{(x_i,y_i) \in S_1} \{|x_i|\} \qquad \text{and} \qquad c_2 < \frac{1}{2} \cdot \min_{(x_j,y_j) \in S_2} \{|x_j|\}.$$

By Remark 6, we may assume that $c_2 > 0$.

Then, the set

$$\bar{S}_1 = \{(\bar{x}_i, \bar{y}_i)\}_{i=1,\ldots,d_1} \quad \text{with } (\bar{x}_i, \bar{y}_i) = (\frac{x_i}{c_1}, y_i),\ (x_i, y_i) \in S_1,$$

is of size d_1 and is shattered by the set of t_1-sparse polynomials $\{\bar{f}_{\sigma^{(1)}}\}_{\sigma^{(1)}\in\{0,1\}^{d_1}}$, where $\bar{f}_{\sigma^{(1)}}(x) = f_{\sigma^{(1)}}(c_1 x)$.

Similarly, the set

$$\bar{S}_2 = \{(\bar{x}_j, \bar{y}_j)\}_{j=1,\ldots,d_2} \quad \text{with } (\bar{x}_j, \bar{y}_j) = (\frac{x_j}{c_2}, y_j),\ (x_j, y_j) \in S_2,$$

is of size d_2 and is shattered by the set of t_2-sparse polynomials $\{\bar{g}_{\sigma^{(2)}}\}_{\sigma^{(2)}\in\{0,1\}^{d_2}}$, where $\bar{g}_{\sigma^{(2)}}(x) = g_{\sigma^{(2)}}(c_2 x)$.

$\bar{S}_1$ and $\bar{S}_2$ satisfy the conditions claimed above, i.e. $\forall (x,y) \in \bar{S}_1 : |x| < \frac{1}{2}$ and $\forall (x,y) \in \bar{S}_2 : |x| > 2$.

Let ϵ_i be the minimal distance of the point $(x_i, y_i) \in \bar{S}_1$ to some shattering polynomial in $\{\bar{f}_{\sigma^{(1)}}\}$, i.e.

$$\epsilon_i = \min_{f\in\{\bar{f}_{\sigma^{(1)}}\}} |f - y_i|.$$

Similarly, for each point $(x_j, y_j) \in \bar{S}_2$, define δ_j by

$$\delta_j = \min_{g\in\{\bar{g}_{\sigma^{(2)}}\}} |g - y_j|.$$

Again, by Remark 6, we assume that $\epsilon_i, \delta_j > 0$.

Our goal is to modify the polynomials from $\{\bar{f}_{\sigma^{(1)}}\}$ and $\{\bar{g}_{\sigma^{(2)}}\}$ such that the polynomials from $\{\bar{f}_{\sigma^{(1)}}\}$ do not interfere the shattering of the set $\bar{S}_2$ and vice versa.

We define a polynomial $F(x)$ to be an upper bound on the polynomials shattering $\bar{S}_1$ in the region according to $\bar{S}_2$, i.e.

$$F(x) > \max_{f(x)\in\{\bar{f}_{\sigma^{(1)}}\}} |f(x)| \quad \text{for all } |x| \geq 2.$$

$F(x)$ is an upper bound on the influence of the polynomials shattering $\bar{S}_1$ on the shattering of the set $\bar{S}_2$.

For some even integer N, we transform the set $\bar{S}_2$ into the set $\bar{S}_2^N$ by

$$(x_j, y_j) \in \bar{S}_2 \quad \Longrightarrow \quad (x_j, x_j^N \cdot y_j) \in \bar{S}_2^N.$$

Since N is even, x^N is positive, and the set $\bar{S}_2^N$ is shattered by the set of t_2-sparse polynomials $\{x^N \cdot \bar{g}_{\sigma^{(2)}}\}$. The minimal distance of the point $(x_j, y_j) \in \bar{S}_2^N$ to some shattering polynomial in $\{x^N \bar{g}_{\sigma^{(2)}}\}$ is $x_j^N \cdot \delta_j$.

We choose the parameter N to be large enough such that the following two conditions are fulfilled:

- The polynomials $\{x^N \cdot \bar{g}_{\sigma^{(2)}}\}$ may not interfere with the shattering of $\bar{S}_1$, i.e

$$x_i^N \cdot \bar{g}_{\sigma^{(2)}} < \epsilon_i \quad \text{for all } (x_i, y_i) \in \bar{S}_1 \text{ and for all } \sigma^{(2)} \in \{0,1\}^{d_2}.$$

Let G be the maximum of the absolute values of the polynomials $\{\bar{g}_{\sigma(2)}(x)\}$ for $|x| \leq 1/2$ (all $|x_i| < \frac{1}{2}$), and ϵ the maximum over all ϵ_i. Then we choose N according to

$$G \cdot (\frac{1}{2})^N < \epsilon, \qquad \text{i.e.} \quad N > \log_2(\frac{G}{\epsilon}).$$

- The polynomials $\{\bar{f}_{\sigma(1)}\}$ may not interfere with the shattering of $\bar{S}_2^N$, i.e

$$F(x_j) < x_j^N \cdot \delta_j \quad \text{for all } (x_j, y_j) \in \bar{S}_2.$$

There exists such an N since the absolute value of the x_j's is at least 2 and N is even.

Let $\bar{S}_1 = \{(x_i', y_i')\}_{i=1,\ldots,d_1}$ with $x_1' < \ldots < x_{d_1}'$, and let $\bar{S}_2^N = \{(x_j'', y_j'')\}_{j=1,\ldots,d_2}$ with $x_1'' < \ldots < x_{d_2}''$.

Let $S = \bar{S}_1 \cup \bar{S}_2^N$. $S = \{(x_k, y_k)\}_{k=1,\ldots,d_1+d_2}$ and $x_1 < \ldots < x_{d_1+d_2}$. For $\sigma \in \{0,1\}^{d_1+d_2}$, we define $\sigma_1 \in \{0,1\}^{d_1}$ and $\sigma_2 \in \{0,1\}^{d_2}$ by

$$\sigma_1(i) = \sigma(k) \text{ iff } x_k = x_i' \qquad \text{and} \qquad \sigma_2(j) = \sigma(k) \text{ iff } x_k = x_j''.$$

S is of size $d_1 + d_2$ and is shattered by the set of $(t_1 + t_2)$-sparse polynomials $\{h_\sigma\}_{\sigma \in \{0,1\}^{d_1+d_2}}$, where $h_\sigma = \bar{f}_{\sigma_1} + x^N \bar{g}_{\sigma_2}$. Hence, the VC dimension of the class of $(t_1 + t_2)$-sparse polynomials is at least $d_1 + d_2$. □

We are now able to state our lower bound on the VC dimension of t-sparse polynomials.

Lemma 8 The VC dimension of t-sparse polynomials is at least $3t$.

PROOF: Combine Lemma 7 and Lemma 5. □

Note that Lemma 7 is also valid for $\mathcal{P}_t^+$. Hence, $2t$ is a lower bound for the VC dimension of $\mathcal{P}_t^+$.

3.3.2 Upper Bounds

In Section 3.2, we have derived an upper bound on the VC dimension of degree-bounded univariate polynomials from the maximal number of roots.

The main tool in this section is *Descartes' Rule of Signs* used to derive an upper bound on the number of roots of t-sparse polynomials. This leads to a first upper bound on the VC dimension of t-sparse polynomials. Considering the structure of sparse polynomials with the maximal number of roots, we derive a (slight) improvement of the upper bound.

We begin with the well known Descartes' Rule (cf. [Coh 82]).

Let $f_t = \sum_{i=1}^t c_i x^{e_i} \in \mathbf{R}[x]$, $f \not\equiv 0$ be a t-sparse polynomial for $e_i < e_{i+1}, i = 1, \ldots, t-1$. The sequence $c = (c_1, c_2, \ldots, c_t)$ is said to have a sign alternation at position i if $c_i c_{i+1} < 0$ (zero coefficients are deleted from the sequence). Denote by $s(f_t)$ the number of sign alternations in c. Let $n^+(f_t)$ denote the number of positive real roots of f_t counted with multiplicity.

Theorem 9 [Descartes' Rule] Let $f \in \mathbf{R}[x], f \not\equiv 0$ be a t-sparse polynomial. Then $s(f) - n^+(f)$ is a non-negative even integer.

Hence, the number of positive real roots of a t-sparse real polynomial $f \not\equiv 0$ is strictly less than its sparsity t. The (total) number of real roots of f is bounded by $2t - 1$ (where the root at the origin is counted without multiplicity).

Let $f \in \mathbf{R}[x]$. f is said to be *even* iff $f_t(x) = f_t(-x)$ (i.e. $\forall i = 1, \ldots, t:$ e_i is even), and f is said to be *odd* iff $f_t(x) = -f_t(-x)$ (i.e. $\forall i = 1, \ldots, t:$ e_i is odd). We call f *symmetric* iff f is odd or even.

Lemma 10 Let $f_t \in \mathbf{R}[x]$, $f_t \not\equiv 0$ be a t-sparse polynomial. If f_t has the maximal number of $2t - 2$ non-zero real roots, then f is symmetric.

PROOF: Let $f_t = \sum_{i=1}^{t} c_i x^{e_i} \in \mathbf{R}[x]$, for $e_i < e_{i+1}, i = 1, \ldots, t-1$. Assume f_t has $2t - 2$ non-zero real roots. Then, f_t has $t - 1$ positive roots. Hence, the sequence of coefficients c of f has $t - 1$ sign alternations. Furthermore the $t - 1$ negative roots are positive roots for $f_t(-x)$. Let $c' = ((-1)^{e_1}c_1, \ldots, (-1)^{e_t}c_t)$ denote the sequence of coefficients of $f_t(-x)$. Suppose f_t is not symmetric. Then there is an index i such that e_i and e_{i+1} are not both even or odd. Therefore, $(-1)^{e_i}c_i \cdot (-1)^{e_{i+1}}c_{i+1} = (-1) \cdot c_i c_{i+1} > 0$, since $c_i c_{i+1} < 0$. Hence, c' has at most $t - 2$ sign alternations, contradicting the assumption that f_t has $t - 1$ negative real roots. □

Using Descartes' estimate on the number of positive real roots of a sparse polynomial, we deduce the (exact) VC dimension of $\mathcal{P}_t^+$.

Lemma 11 The VC dimension of the concept class $\text{pos}(y - \mathcal{P}_t^+)$ equals $2t$.

PROOF: Let d denote the VC dimension of $\mathcal{P}_t^+$. In Section 3.3.1, we proved $d \geq 2t$. Hence, we have to show $d \leq 2t$.

Let $S = \{(x_i, y_i)\}_{i=1,\ldots,d}$, where $0 < x_1 < x_2 < \ldots < x_d$ be a set of points shattered by t-sparse polynomials. Let f_1 and f_2 be t-sparse polynomials satisfying the two alternating labelings $\sigma_1 = (1, 0, 1, 0, \cdots, 1, 0)$ and $\sigma_2 = (0, 1, 0, 1, \cdots, 0, 1)$. Let $F = (f_1 - f_2)$. Note that F is $2t$-sparse and $s(F) \leq 2t - 1$. Furthermore, $F(x_i) \cdot F(x_{i+1}) < 0$ for $i = 1, \ldots, d-1$, forcing F to have at least $d - 1$ positive real roots. By Descartes' Rule $d - 1 < 2t$, proving the statement. □

By Lemma 11, the VC dimension of $\mathcal{P}_t$ is bounded by $4t$. With Lemma 8, the VC dimension of $\mathcal{P}_t$ is linear in t.

Lemma 12 gives an improvement of the upper bound on the VC dimension of $\mathcal{P}_t$. As a consequence, the VC dimension of 1-sparse polynomials is exactly 3.

Lemma 12 The VC dimension of $\mathcal{P}_t$ is at most $4t - 1$.

PROOF: Assume, for purpose of contradiction, that the set $S = \{(x_i, y_i)\}_{i=1,\ldots,4t}$, for $x_1 < x_2 < \ldots < x_{2t} < 0 < x_{2t+1} < \ldots < x_{4t}$ is shattered by t-sparse polynomials.

Consider the following 4 labelings on the set S:

$$\sigma_1 = (\underbrace{1,0,1,0,\ldots,1,0}_{2t},\underbrace{1,0,1,0,\ldots,1,0}_{2t}),$$

$$\sigma_2 = (\underbrace{0,1,0,1,\ldots,0,1}_{2t},\underbrace{0,1,0,1,\ldots,0,1}_{2t}),$$

and

$$\gamma_1 = (\underbrace{0,1,0,1,\ldots,0,1}_{2t},\underbrace{1,0,1,0,\ldots,1,0}_{2t}),$$

$$\gamma_2 = (\underbrace{1,0,1,0,\ldots,1,0}_{2t},\underbrace{0,1,0,1,\ldots,0,1}_{2t}).$$

Let the t-sparse polynomials f_1, f_2 and g_1, g_2 satisfy the labelings σ_1, σ_2, and γ_1, γ_2.

Define $F = f_1 - f_2$ and $G = g_1 - g_2$. Note that both F and G are $2t$-sparse. By the alternating structure of the labelings, both F and G have at least $4t - 2$ non-zero real roots. From Lemma 10, F and G are symmetric.

We show that F is odd and G is even. Assume, F is even and let $|x_{2t}| < x_{2t+1}$. Then, $F(-x_{2t}) = F(x_{2t}) > 0$ and $F(x_{2t+1}) < 0$, i.e. F has an 'extra' positive root in the interval $(-x_{2t}, x_{2t+1})$, contradicting the upper bound on the number of positive real roots. For $|x_{2t}| > x_{2t+1}$ F has an 'extra' negative root in the interval $(x_{2t}, -x_{2t+1})$. The proof that G is even is similar.

Note that F is odd implies that both f_1 and f_2 are odd (if some monomial occurs in f_1 and in f_2 as well, F would be at most $2t - 1$-sparse). Similarly, both g_1 and g_2 are even. Then, w.l.o.g., we may assume (for sake of simplicity of notation) that the x-values of the points from S are symmetric as well, i.e. $x_i = -x_{4t+1-i}$, $i = 1, \ldots, 2t$.

We define $2t - 1$ intervals J_i on the negative real line by $J_i = (x_i, x_{i+1})$, $i = 1, \ldots, 2t - 1$. We prove that for each $i = 1, \ldots, 2t - 1$ at least two polynomials from $\{f_1, f_2, g_1, g_2\}$ have a (negative) root in the interval J_i. We distinguish two cases:

1. Let y_i and y_{i+1} have different signs. Assume $y_i < 0$, $y_{i+1} > 0$ and i odd. Then, by definition of the labelings, $f_1(x_i), g_2(x_i) < y_i < 0$ and $f_1(x_{i+1}), g_2(x_{i+1}) > y_{i+1} > 0$. Hence, f_1 and g_2 have a root in J_i. If i is even, f_2 and g_1 have a root in J_i. The case $y_i > 0$, $y_{i+1} < 0$ is symmetric.

2. Let y_i and y_{i+1} have equal signs. We show that f_1 or g_2 and f_2 or g_1 have a root in J_i.

   Assume $y_i, y_{i+1} > 0$ and i odd. Then $f_1(x_{i+1}), g_2(x_{i+1}) > y_{i+1} > 0$. Assume f_1 has no root in J_i (f_1 is strictly positive in J_i). Then f_1 is strictly negative in the interval (x_{4t-i}, x_{4t-i+1}) (f_1 is odd). Since g_2 is even, $g_2(x_{4t-i+1}) > 0$ and $g_2(x_{4t-i}) < f_1(x_{4t-i}) < 0$. Hence, g_2 has a root in the interval (x_{4t-i}, x_{4t-i+1}) and (g_2 is symmetric) g_2 has a root in J_i. Similarly, we can show that either f_2 or g_1 has a root in J_i. The remaining cases are symmetric.

Hence, the total number of negative roots of the polynomials from $\{f_1, f_2, g_1, g_2\}$ is at least $2 \cdot (2t - 1) = 4t - 2$ contradicting the assumption that each polynomial from $\{f_1, f_2, g_1, g_2\}$ is t-sparse (each polynomial has at most $t - 1$ negative roots summing up to at most $4t - 4$ negative roots). This proves the claimed upper bound of $4t - 1$ on the VC dimension of t-sparse polynomials. □

We state the main result of this section:

Theorem 13 For fixed $t \in \mathbb{N}$, the class of t-sparse polynomials is uniformly and distribution-free learnable. The sample size required for (ϵ, δ)-learning is at most

$$\frac{4}{\epsilon} \cdot \max\left(\log\frac{2}{\delta}, (8t-2)\log\frac{13}{\epsilon}\right).$$

PROOF: Apply the results of Blumer et al. (Theorem 1, Theorem 2). Note that the concept class $\mathrm{pos}(y - \mathcal{P}_t)$ is universally separable since any real polynomial can be written as the pointwise limit of some polynomial over the rational numbers. □

Note that the bounds derived in this subsection remain valid when restricted to t-sparse polynomials over the rational numbers and t-sparse polynomials over the integers.

Let $\mathcal{R}_t$ denote the set of real rational functions with t-sparse numerator and t-sparse denominator. Following the proof of Lemma 11, we derive the upper bound of $4t^2$ on the VC dimension of $\mathrm{pos}(y - \mathcal{R}_t)$ proving uniform learnability of t-sparse rational functions for any fixed t.

Theorem 14 The VC dimension of $\mathcal{R}_t$ is at most $4t^2$.

PROOF: Let d denote the VC dimension of $\mathrm{pos}(y - \mathcal{R}_t)$. Consider the two rational functions $f_1 = \frac{g_1}{h_1}$, $f_2 = \frac{g_2}{h_2}$ from $\mathcal{R}_t$ satisfying the alternating labelings. Then $f_1(x) = f_2(x)$ for at least $d - 1$ points, that is, the $2t^2$-sparse polynomial $g_1h_2 - g_2h_1$ has to have at least $d - 1$ real roots. From Theorem 9, we have $d - 1 \leq 4t^2 - 1$. □

Remark 15 It is interesting to compare our results on the learnability of polynomials with the learnability of trigonometric polynomials. Since degree-bounded trigonometric polynomials form a vector space of finite dimension, this concept class is of finite VC dimension (cf. Theorem 4) and, hence, uniformly learnable (cf. Theorem 1). On the other hand, sparse trigonometric polynomials may oscillate arbitrarily often. It is easily verified that the VC dimension of the class of sparse trigonometric polynomials is infinite and, hence, not learnable. □

3.4 Related Results

3.4.1 VC Dimension of Degree-bounded Sparse Polynomials

We give now sharp bounds on the VC dimension of sparse and degree-bounded univariate real polynomials. For practical applications, this is the most important case.

For each $t, n \in \mathbf{N}, t \leq n$, let $\mathcal{P}_{t,n} \subset \mathbf{R}[x]$ denote the set of t-sparse univariate polynomials of degree less than n, i.e. $\mathcal{P}_{t,n} = \{p \in \mathcal{P}_t \mid \deg(p) < n\}$. Let $\mathcal{P}^+_{t,n} = \{p \in \mathcal{P}^+_t \mid \deg(p) < n\}$.

From Section 3.2 and Section 3.3.2, we derive an upper bound of $\min\{n, 4t-1\}$ on the VC dimension of $\mathcal{P}_{t,n}$ and $\min\{n, 2t\}$ on the VC dimension of $\mathcal{P}^+_{t,n}$. In this section, we investigate the corresponding lower bounds.

Lemma 16

1. $\mathrm{VCdim}(\mathcal{P}_{t,n}) \geq \mathrm{VCdim}(\mathcal{P}_{t-1,n-3}) + 3$
2. $\mathrm{VCdim}(\mathcal{P}^+_{t,n}) \geq \mathrm{VCdim}(\mathcal{P}^+_{t-1,n-2}) + 2.$

PROOF: Let $d = \mathrm{VCdim}(\mathcal{P}_{t-1,n-3})$. Then there exists a set $S = \{(x_i, y_i)\}_{i=1,\ldots,d}$ that is shattered by the set $\{f_\sigma\}_{\sigma\in\{0,1\}^d}$ of $(t-1)$-sparse polynomials of degree less than $n-3$.

As shown in the proof of Lemma 7, we may assume $\max_i |x_i| < 1$. Let

$$0 < \epsilon < \min_{i,\sigma} |f_\sigma(x_i) - y_i|$$

be the minimal distance of some point from S to some shattering polynomial in $\{f_\sigma\}_{\sigma\in\{0,1\}^d}$. Since $\deg(f_\sigma) < n-3$, there exists $a > \max\{\frac{\epsilon}{27}, \frac{\epsilon^2}{243}\}$ such that

$$\forall \sigma \in \{0,1\}^d \; \forall x \geq 1 : \; |f_\sigma(x)| < a \cdot |x|^{n-4} =: M(x).$$

We show that three additional points are shattered by adding monomials of degree at most $n-1$ to the polynomials in $\{f_\sigma\}_{\sigma\in\{0,1\}^d}$, that is, by increasing the sparsity by one and the degree by three.

Consider, for instance, the points $(x_0, y_0) = (-81\frac{a}{\epsilon}, (-1)^n 30 \cdot M(-81\frac{a}{\epsilon}))$, $(x_{d+1}, y_{d+1}) = (9\frac{a}{\epsilon}, 2 \cdot M(9\frac{a}{\epsilon}))$ and $(x_{d+2}, y_{d+2}) = (81\frac{a}{\epsilon}, 60 \cdot M(-81\frac{a}{\epsilon}))$.

In Figure 2, we give monomials $\{g_j\}_{j=1,\ldots,8}$ of degree less than n shattering these three points.

Note that the minimal distance of the g_i's to (x_0, y_0), (x_{d+1}, y_{d+1}), and (x_{d+2}, y_{d+2}) is at least $M(x_0)$, $M(x_{d+1})$, respectively $M(x_{d+2})$, hence greater than the absolute values of each $f \in \{f_\sigma\}_{\sigma\in\{0,1\}^d}$ at these points. On the other hand, $g_j(x) < \epsilon$ for $x < 1$, $j = 1, \ldots, 8$.

Hence, the set $S' = \{(x_i, y_i)\}_{i=0,\ldots,d+2}$ is shattered by the set $\{f_\sigma + g_j \mid \sigma \in \{0,1\}^d, j = 1, \ldots, 8\}$ of t-sparse polynomials of degree less than n. This proves the first statement.

Note that the monomials $\{g_j\}_{j=1,\ldots,4}$ of degree at most $n-2$ shatter the two points (x_{d+1}, y_{d+1}) and (x_{d+2}, y_{d+2}). Hence the second statement follows. □

Lemma 17

1. $\mathrm{VCdim}(\mathcal{P}_{1,n}) = \min\{n, 3\}$.
2. $\mathrm{VCdim}(\mathcal{P}^+_{1,n}) = \min\{n, 2\}$.

PROOF: The statement is clear for $n = 1, 2$. Note that $\mathrm{VCdim}(\mathcal{P}^+_1) = 2$ and $\mathrm{VCdim}(\mathcal{P}_1) = 3$. For the first statement, we proved in Lemma 5 that for each labeling $\sigma \in \{0,1\}^3$ there is a monomial f_σ of degree less than 3 satisfying σ on the set $S = \{(-3, 4), (1, 2), (7, 6)\}$ of size 3. □

	$\frac{g_i(x_0)}{M(x_0)}$	$\frac{g_i(x_{d+1})}{M(x_{d+1})}$	$\frac{g_i(x_{d+2})}{M(x_{d+2})}$	n even	n odd
$g_1 = 0$	0	0	0	011	111
$g_2 = \frac{\epsilon}{3} \cdot x^{n-3}$	$-27\,(-1)^n$	3	27	001	101
$g_3 = \frac{\epsilon^2}{81a} \cdot x^{n-2}$	$81\,(-1)^n$	1	81	010	110
$g_4 = \epsilon \cdot x^{n-3}$	$-81\,(-1)^n$	9	81	100	000
$g_5 = \frac{\epsilon^2}{27a} \cdot x^{n-2}$	$243\,(-1)^n$	3	243	000	100
$g_6 = \frac{2\epsilon}{3} \cdot x^{n-3}$	$-54\,(-1)^n$	6	54	101	001
$g_7 = -\frac{\epsilon^2}{27a} \cdot x^{n-2}$	$-243\,(-1)^n$	-3	-243	111	011
$g_8 = \frac{\epsilon^3}{243a^2} \cdot x^{n-1}$	$-2187\,(-1)^n$	1	2187	110	010

Figure 2: Monomials $\{g_i\}$ shattering the three additional points

Corollary 18

1. $\min\{n, 4t-1\} \geq \mathrm{VCdim}(\mathcal{P}_{t,n}) \geq \min\{n, 3t\}$.
2. $\mathrm{VCdim}(\mathcal{P}^+_{t,n}) = \min\{n, 2t\}$.

PROOF: Note that $\mathrm{VCdim}(\mathcal{P}_{t,n}) \leq \min\{n, 4t-1\}$, since $\mathrm{VCdim}(\mathcal{P}_n) = n$, $\mathrm{VCdim}(\mathcal{P}_t) \leq 4t-1$ and $\mathrm{VCdim}(\mathcal{P}^+_{t,n}) \leq \min\{n, 2t\}$, since $\mathrm{VCdim}(\mathcal{P}^+_n) = n$, $\mathrm{VCdim}(\mathcal{P}^+_t) = 2t$.

The statement follows by induction with Lemma 16 and Proposition 17. Then

$$\begin{aligned} \min\{n, 4t-1\} &\geq \mathrm{VCdim}(\mathcal{P}_{t,n}) \\ &\geq \mathrm{VCdim}(\mathcal{P}_{t-1,n-3}) + 3 \\ &= \min\{n-3, 3(t-1)\} + 3 \;=\; \min\{n, 3t\} \end{aligned}$$

and

$$\begin{aligned} \min\{n, 2t\} &\geq \mathrm{VCdim}(\mathcal{P}^+_{t,n}) \\ &\geq \mathrm{VCdim}(\mathcal{P}^+_{t-1,n-2}) + 2 \\ &= \min\{n, 2t\}. \end{aligned}$$

□

3.4.2 Learnability of Sparse Polynomials in the Metric PAC Model

In this subsection, we investigate the learnability of sparse polynomials in the generalized PAC model of Haussler (cf. Section 2.2). The essential condition for distribution-free uniform learnability (in this model) of the class of sparse polynomials is the finiteness of the metric VC dimension of $\mathcal{P}_t$. We prove linear bounds for m-$\mathrm{VCdim}(\mathcal{P}_t)$.

First, we construct a lower bound on the metric VC dimension of the class $\mathcal{P}_t$.

Lemma 19 m-VCdim$(\mathcal{P}_t) \geq$ VCdim$(\mathcal{P}_t)$.

PROOF: Let $d =$ VCdim$(\mathcal{P}_t)$, and let $S = \{(x_i, y_i)\}_{i=1,\ldots,d}$ be a set of points shattered (in the standard sense) by the set of t-sparse polynomials $\{f_\sigma\}_{\sigma \in \{0,1\}^d} \subset \mathcal{P}_t$, i.e.

$$\forall i = 1, \ldots, d \ \forall \sigma \in \{0,1\}^d \ : \ f_\sigma(x_i) - y_i \begin{cases} \leq 0 & \text{if } \sigma(i) = 1 \\ > 0 & \text{if } \sigma(i) = 0 \end{cases} .$$

Let ϵ be defined by

$$\epsilon = \max_{i=1,\ldots,d} \max_{\sigma, \sigma(i)=1} y_i - f_\sigma(x_i).$$

Then

$$\forall i = 1, \ldots, d \ \forall \sigma \in \{0,1\}^d \ : \ |f_\sigma(x_i) - (y_i - \epsilon)| \begin{cases} \leq \epsilon & \text{if } \sigma(i) = 1 \\ > \epsilon & \text{if } \sigma(i) = 0 \end{cases} ,$$

i.e. the set $S_\epsilon = \{(x, y - \epsilon, \epsilon) \mid (x, y) \in S\}$ of size d is shattered (in the metric sense) by the set of t-sparse polynomials $\{f_\sigma\}_{\sigma \in \{0,1\}^d} \subset \mathcal{P}_t$. Hence, m-VCdim$(\mathcal{P}_t) \geq$ VCdim$(\mathcal{P}_t)$. □

We introduce the following lemma to derive an upper bound on m-VCdim$(\mathcal{P}_t)$.

Lemma 20 Let $S = \{(x_i, y_i, \epsilon_i)\}_{i=1,\ldots,4}$ where $x_1 < x_2 < x_3 < x_4$. Let $\sigma_1 = (1,0,0,1)$, $\sigma_2 = (0,1,1,0)$, $\sigma_3 = (1,0,1,0)$, $\sigma_4 = (0,1,0,1)$ be labelings on S. Let $\{f_i\}_{i=1,\ldots,4}$ be continuous functions satisfying σ_i on S (in the metric sense). Then at least one of the following pairs of functions (f_1, f_2), (f_1, f_3), (f_1, f_4), (f_3, f_4) have an intersection point in the interval (x_1, x_4).

PROOF: Consider the 2^8 cases for $f_i(x_j) > y_j + \epsilon_j$ or $f_i(x_j) < y_j - \epsilon_j$ if $\sigma_i(j) = 0$. □

Theorem 21 The metric VC dimension of the class of t-sparse polynomials is at most $48t - 9$.

PROOF: Let $d =$ m-VCdim$(\mathcal{P}_t)$ and $S = \{(x_i, y_i, \epsilon_i)\}_{i=1,\ldots,d}$, where $x_1 < x_2 < \ldots < x_d$. Assume S is shattered by t-sparse polynomials. Consider the labelings $\sigma_1 = (1,0,0,1,0,0,1,0,0,\ldots)$, $\sigma_2 = (0,1,1,0,1,1,0,1,1,\ldots)$, $\sigma_3 = (1,0,1,0,1,0,\ldots)$, and $\sigma_4 = (0,1,0,1,0,1,\ldots)$. Let $f_1, \ldots, f_4$ be t-sparse polynomials satisfying $\sigma_1, \ldots, \sigma_4$. Then, by Lemma 20, there are two polynomials with at least $d/12$ intersections, and, with Lemma 10, we conclude that $d \leq 12(4t-1) + 2 = 48t - 10$ (there may exist two additional points, where we cannot apply Lemma 20). □

Corollary 22 For any fixed $t \in \mathbf{N}$, the class of t-sparse polynomials is uniformly and distribution-free learnable in the metric PAC model.

3.4.3 Approximating the Polynomial Regression

As an application of the results derived in the previous section, we consider the open problem stated by Vapnik ([Vap 82]) on computational approximation of the general regression functions used in the theory of empirical data dependences.

One of the central problems in computational regression theory is the problem of determining the number of terms in an arranged system of functions. The most important case of this problem is the approximation of polynomial regression (cf. [Vap 82], pp. 254–258).

The classical scheme of approximating polynomial regression, which involves the determination of the true degree n of regression and the expansion in a system of n orthogonal polynomials of degree $1, 2, \ldots, n$, can be successfully implemented only when large samples are used. The reason for this is the (possibly) large degree of regression and therefore the large metric VC dimension (capacity) of the class of polynomials of degree n. The problem for small samples remained open.

We prove linear bounds (Theorem 21) on the metric VC dimension of t-sparse polynomials (independent of the degree) implying Corollary 23.

Corollary 23 The polynomial regression can be estimated for small samples (depending only on the number of required terms).

4 Open Problems and Further Research

In Section 3, we have proved uniform and distribution-free learnability of sparse univariate polynomials, but several related problems remain open. In this section, we list some of the open problems in the area of learnability of sparse polynomials.

Learnability of Multivariate Polynomials

From Theorem 4, degree-bounded multivariate polynomials are of finite VC dimension for any fixed number of variables. There is no corresponding result for sparse multivariate polynomials. As described in Section 3, the main tool for proving the finiteness of the VC dimension in the sparse univariate case is the upper bound on the number of roots of sparse polynomials derived from Descartes' Rule. A promising approach for the multivariate case might be the work of Khovanskii ([Kh 83]). Khovanskii generalizes the Descartes' estimate to the sparse multivariate case and proves that the number of non-degenerate roots of sparse polynomials as well as the number of connected components of a singular real algebraic variety can be estimated in terms of the sparsity and the number of variables. In spite of these results, it is not clear in the multidimensional case how to relate the VC dimension to the upper bounds on the roots of multivariate polynomials.

Efficient Learning Algorithms

It is an open problem if there exists a hypothesis finder for the class of t-sparse polynomials such that the time complexity of the algorithm is bounded only in terms of the sparsity and the sample size.

A related problem is the problem of whether or not the class of sparse polynomials is learnable with respect to target complexity. The results of Linial et al. ([LMR 88]) imply that this is equivalent to the question of whether or not the class of sparse polynomials is polynomially uniformly decomposable. This reduces to the problem of the existence of a polynomial-time algorithm for sparse linear programming. Note that the existence of such an algorithm would not imply polynomial learnability of the class of t-sparse polynomials for fixed t since the appropriate degree is unknown.

Exact VC Dimension of Sparse Polynomials

From Lemma 11, the VC dimension of sparse univariate polynomials on the right half space equals $2t$. Lemma 8 and 12 give a lower bound of $3t$ and an upper bound of $4t - 1$ on the VC dimension in the unrestricted case. The exact VC dimension remains unknown. For the metric VC dimension the tradeoff between lower bound ($3t$) and upper bound ($48t - 10$) is even larger.

Data Compression Schemes

Given a finite set of examples, labeled consistently with some concept from a concept class, a data compression scheme of size d saves at most d of those examples. From the d saved examples, the data compression scheme reconstructs a hypothesis that is consistent with the original sample. Data compression schemes are of crucial importance in the context of on-line and space-bounded learning algorithms. It is an open problem whether or not there exists a data compression scheme of small size for the class of sparse univariate polynomials. For the case of univariate real polynomials of degree less than n, Floyd ([Fl 89]) shows that there is a data compression scheme of size n and gives an on-line learning algorithm saving at most n examples at a time. The techniques used by Floyd depend mainly on the vector space structure, induced by degree-bounded polynomials, implying that any interpolation problem is solvable. For sparse polynomials, these techniques are not applicable.

Acknowledgements

We thank Manuel Blum, Allan Borodin, Sally Floyd, Dima Grigoriev, Les Valiant, Vladimir Vapnik, and Manfred Warmuth for the number of stimulating discussions.

References

[AHU 74] Aho, A., Hopcroft, J., Ullman, J., *The Design and Analysis of Computer Algorithms*, Addison-Wesley, London, 1974.

[BEHW 86] Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.K., *Classifying Learnable Geometric Concepts with the Vapnik-Chervonenkis Dimension*, Proc. 18^{th} ACM STOC, pp. 273–282, 1986.

[BEHW 89] Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.K., *Learnability and the Vapnik-Chervonenkis Dimension*, Journal ACM 36 (4), pp. 929–965, 1989.

[BeTi 88] Ben-Or, M., Tiwari, P.A., *A Deterministic Algorithm for Sparse Multivariate Polynomial Interpolation*, Proc. 20^{th} ACM STOC, pp. 301–309, 1988.

[BoTi 89] Borodin, A., Tiwari, P.A., *On the Decidability of Sparse Univariate Polynomial Interpolation*, IBM Research Report RC 14923 (#66763), 1989.

[Br 90] Bruck, J., *Harmonic Analysis of Polynomial Threshold Functions*, SIAM J. Discrete Math. 3 (2), pp. 282–287, 1990.

[BS 90] Bruck, J., Smolensky, R., *Polynomial Threshold Functions, AC^0 Functions, and Spectral Norms*, Proc. 31^{th} IEEE FOCS, pp. 632–641, 1990.

[Coh 82] Cohn, P.M., Algebra, Vol. 1, 2^{nd} ed., John Wiley & Sons Ltd., 1982.

[Cov 65] Cover, T.M., *Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition*, IEEE Trans. Electron. Comput. 14, pp. 326–334, 1965.

[Du 78] Dudley, R.M., *Central Limit Theorems for Empirical Measures*, The Annals of Probability 6 (6), 1978, pp. 899–929.

[EHKL 89] Ehrenfeucht, A., Haussler, D., Kearns, M., Valiant, L., *A General Lower Bound on the Number of Examples Needed for Learning*, Information and Computation 82 (3), pp. 247–261, 1989.

[Fl 89] Floyd, S., *On Space-bounded Learning and the Vapnik-Chervonenkis Dimension*, Technical Report TR-89-061, Ph.d. dissertation, International Computer Science Institute, Berkeley, 1989.

[GK 87] Grigoriev, D.Yu., Karpinski, M., *The Matching Problem for Bipartite Graphs with Polynomially Bounded Permanent is in NC*, Proc. 28^{th} IEEE FOCS, pp. 166–172, 1987.

[GKS 90] Grigoriev, D.Yu., Karpinski, M., Singer, M., *Fast Parallel Algorithms for Sparse Multivariate Polynomial Interpolation over Finite Fields*, SIAM J. Comp. 19 (6), pp. 1059–1063, 1990.

[Ha 89] Haussler, D., *Generalizing the PAC Model: Sample Size Bounds from Metric Dimension-based Uniform Convergence Results*, Proc. 30^{th} IEEE FOCS, pp. 40–45, 1989.

[Kar 89] Karpinski,M., *Boolean Circuit Complexity of Algebraic Interpolation Problems*, Proc. CSL'88, LNCS 385, Springer Verlag, pp. 138–147, 1989.

[KW 89] Karpinski, M., Werther, T., *VC Dimension and Learnability of Sparse Polynomials and Rational Functions*, Technical Report TR-89-060, International Computer Science Institute, Berkeley, 1989.

[Kh 83] Khovanskii, A.G. *Fewnomials and Pfaff Manifolds*, Proc. of the Intern. Congress of Math., Warsaw, 1983.

[KM 91] Kushilevitz, E., Mansour, Y., *Learning Decision Trees Using the Fourier Spectrum*, Proc. 23^{th} ACM STOC, pp. 455–464, 1991.

[LMR 88] Linial, N., Mansour, Y., Rivest, R.L., *Results on Learnability and the Vapnik-Chervonenkis Dimension*, Proc. 29^{th} IEEE FOCS, pp. 120–129, 1988.

[Po 84] Pollard, D., *Convergence of Stochastic Processes*, Springer Verlag, 1984.

[Val 84] Valiant, L.G., *A Theory on the Learnable*, Comm. ACM, 27(11), pp. 1134–1142, 1984.

[Vap 82] Vapnik, V.N., *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, 1982.

[Vap 89] Vapnik, V.N., *Inductive Principles of the Search for Empirical Dependences (Methods Based on Weak Convergence of Probability Measures)*, Proc. of the 2nd Workshop on Computational Learning Theory, pp. 3–21, 1989.

[VC 71] Vapnik, V.N., Chervonenkis, A.Y., *On the Uniform Convergence of Relative Frequencies of Events and their Probabilities*, Th. Prob. and its Appl., 16(2), pp. 264–280, 1971.

[WD 81] Wenocur, R.S., Dudley, R.M., *Some Special Vapnik-Chervonenkis Classes*, Discrete Mathematics 33, pp. 313–318, 1981.

[We 91] Werther, T., *VC Dimension and Learnability of Sparse Polynomials*, Ph. D. Thesis, University of Bonn, 1991.