Using Existential Theory of the Reals to Bound VC-Dimension

Austin Watkins
University of Utah

UUCS-20-008

School of Computing
University of Utah
Salt Lake City, UT 84112 USA

22 April 2020

Abstract

We solve the open problem of bounding the VC-dimension of inflated polynomials. To achieve this bound, we use the decidability algorithms for existential theory of the reals. Further, our results are generalized to give an upper bound on the VC-dimension for all semialgebraic sets constructed from a finite set of bounded degree polynomials. The VC-dimension of a geometric object, represented by a range space, encodes its geometric complexity. The VC-dimension of range spaces has applications towards the learnability of corresponding function classes within computational learning theory. Finally, VC-dimension is important in probability theory, computational geometry, and model theory.

USING EXISTENTIAL THEORY OF THE REALS TO BOUND VC-DIMENSION

by

Austin Watkins
Advisor: Jeff Phillips

A senior thesis submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

Bachelor of Science

in

Computer Science

Department of Computer Science

The University of Utah

April 2020

USING EXISTENTIAL THEORY OF THE REALS TO BOUND VC-DIMENSION

by

Austin Watkins

A thesis submitted to the faculty of
The University of Utah
in partial fulfillment of the requirements for the degree of

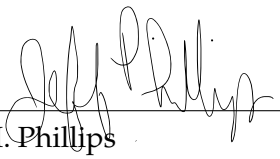
Bachelor of Science in Computer Science

School of Computing

The University of Utah

April 2020

Approved:



Jeff M. Phillips Supervisor	April 30, 2020 Date Approved
H. James de St. Germain Director of Undergraduate Studies School of Computing	Date Approved
Ross Whitaker Director School of Computing	Date Approved

ABSTRACT

We solve the open problem of bounding the VC-dimension of inflated polynomials. To achieve this bound, we use the decidability algorithms for existential theory of the reals. Further, our results are generalized to give an upper bound on the VC-dimension for all semialgebraic sets constructed from a finite set of bounded degree polynomials. The VC-dimension of a geometric object, represented by a range space, encodes its geometric complexity. The VC-dimension of range spaces has applications towards the learnability of corresponding function classes within computational learning theory. Finally, VC-dimension is important in probability theory, computational geometry, and model theory.

For my parents, Leisa and Craig.

CONTENTS

CHAPTER 1

INTRODUCTION

A bound of VC-dimension for different objects is of interest in the study of probability theory [14], computational learning theory [1], model theory [2], [3], and computational geometry [8]. The Vapnik-Chervonenkis-dimension (VC-dimension) viewed from a geometric perspective is a measurement of the underlying complexity of a set system [10], which we will define with an object called a range space. There is prior work on developing methods for bounding VC-dimension [1], [10], although these methods can be insufficient when bounding more complex range spaces. Our work uses these traditional tools along with algorithms in algebraic geometry to bound VC-dimension of complex range spaces. As mentioned, the study of VC-dimension is important to many fields. In bounding large and complex range spaces the opportunity for applications to these fields increases.

These traditional tools, circuit and composition arguments, are also not well-suited for evaluation of distance. Specifically, the square root operation needed for Euclidean distance is not included in the set of allowed operations for a circuit argument. Perhaps justified by this limitation bounding the VC-dimension of range spaces which involve Euclidean distance is not well developed. Therefore, in [6] an open problem was established to bound the VC-dimension of a set of inflated polynomials. Inflated polynomials are polynomials with a tubular neighborhood. For example, an inflated quadratic polynomial is a parabola shaped "band" through the plane. The study of inflated polynomials are important objects to study because polynomials are more complex than half-spaces and they encode Euclidean distance. We solve this problem by providing an upper and lower bound for the inflated polynomial range space. The methods we use are general, which is demonstrated by bounding the VC-dimension of semialgebraic sets.

In particular we consider semialgebraic sets composed of a finite set of bounded de-

gree polynomials. These sets are equivalent to finite union of polynomial equalities and inequalities. The VC-dimension of this space is finite and therefore learnable. As this function class is large the opportunities to reduce other problems to semialgebraic sets is present.

Through this work we have developed a connection between decidability algorithms in logic and VC-dimension using the circuit argument technique. The circuit argument bounds VC-dimension of a range space based on the number of simple operations required to perform a decision procedure to evaluate inclusion of a point. We can build these decision procedures using algorithms from algebraic geometry and, in the process, receive bounds on the simple operations involved. There are more potential applications to be found through studying the connections of logic, specifically decidability procedures, and VC-dimension.

We also detail several applications of our results. VC-dimension establishes conditions for learnability and existence of ε-nets. Specifically, we provide the sample complexity of inflated polynomials, inflated splines, and semiagebraic sets. We also briefly detail applications to smoothed range spaces, which has applications to range spaces involving noise and regression.

The remaining chapters will continue as follows. First, we will consider the relevant definitions, theorems and algorithms already established in several fields. Next, we will give proofs for our results. Finally, we give a detailed account of the applications of our work.

CHAPTER 2

BACKGROUND, DEFINITIONS, AND PRIOR WORK

In our proofs we will need some established results from several fields. First, we will define terms from logic and computational geometry. Second, we include the theorem that is primarily used for bounding sample complexity with VC-dimension. Third, in our constructivist proof of a lower bound on the VC-dimension of inflated polynomial range space we will need multivariate Lagrange interpolation. Next, the important circuit and composition arguments are detailed with examples. Then, we provide a prior connection established between the square root operation and circuit arguments. Finally, we define algebraic and algebraic sets and give several algorithms that operate on these sets.

2.1 Definitions

Central to our study are polynomials with real coefficients.

Definition 1 (Real Polynomials)**.** *The set of all polynomials over the reals is* $\mathbb{R}[X_1, \ldots, X_d]$*. Where the degree of any polynomial is the maximum sum of the exponents of the variables,* $X_1, \ldots, X_d$ *in any monomial.*

There are several ways to view polynomials.

- The set of all d-variate polynomials with real coefficients
- Curves in $(d+1)$ space
- Functions $\mathbb{R}^d \mapsto \mathbb{R}$

A common operation between two sets is Minkowski addition.

Definition 2 (Minkowski Addition)**.** *Given sets* $A, B \in \mathbb{R}^d$*, the Minkowski sum is the set* $A \oplus B = \{a + b \mid a \in A, b \in B\}$

Extending the perspective of polynomials as curves in $\mathbb{R}^{d+1}$ we will define inflated polynomials as the Minkowski sum of a disk and a polynomial.

Definition 3 (Inflated Polynomials)**.** *Let $\mathcal{M}_d$ be the set of all possible Minkowski additions between a disk of variable radius and polynomials of bounded degree p. That is,*

$$\mathcal{M}_p = \{\mathbb{R}[X_1, \ldots, X_d] \oplus \bar{D}_r^{d+1} \mid r \in \mathbb{R}, p \in \mathbb{N}\}$$

where $\bar{D}_r^{d+1}$ is a $(d+1)$-dimensional disk with radius r.

A range space is a useful object in the study of sets of sets. That is, sets that contain other sets.

Definition 4 (Range space)**.** *A range space is a tuple $(X, \mathcal{R})$, where X is called the ground set and $\mathcal{R}$ is called the range set, where all sets in the range set are a subset of the ground set. Similar to a restriction over a set of functions to a subset of the functions domain, we will define $\mathcal{R}_{|Y} := \{R \cap Y \mid R \in \mathcal{R}\}$, for $Y \subseteq X$ [10].*

$\mathcal{R}$ is often defined in terms of geometric objects. $\mathcal{R}$ could be the set of disks on $\mathbb{R}^2$, intervals on $\mathbb{R}$, or more complex structures as polynomials in $\mathbb{R}^d$. Intuitively, polynomials are more complex than intervals, it is natural to want to define a measurement of this underlying complexity. The VC-dimension is a classic measurement of the underlying expressiveness of the range set [10]. First, we need to define shattering a set.

Definition 5 (Shattering)**.** *Consider, a range space $(X, \mathcal{R})$, with the projection $\mathcal{R}_{|Y}$. If $\mathcal{R}_{|Y}$ contains all subsets of Y then it is said that $\mathcal{R}$ shatters Y [10].*

If $m = |Y|$ then the behavior of $\mathcal{R}_{|Y}$ may change as m becomes large. That is $\mathcal{R}$ may cease to shatter Y for large enough m. Intuitively, more expressive range sets will continue to shatter large m while their more simple counterparts are unable to do so. This gives us the VC-dimension:

Definition 6 (VC-dimension)**.** *VC-dimension of $(X, \mathcal{R})$ is the maximum cardinality of a shattered subset of X [10].*

By way of example [10]:

- The set of disks over $\mathbb{R}^2$ has VC-dimension 3
- The set of intervals over $\mathbb{R}$ has VC-dimension 2
- The set of all polynomials in $\mathbb{R}^d$ has VC-dimension ∞
- The set of all half-spaces in $\mathbb{R}^d$ shatter $d + 1$

That is, disks are unable to shatter a set of size 4, intervals are unable to shatter sets of size 3, and half-spaces in $\mathbb{R}^3$ are unable to shatter a set of size 5.

We will define the operations allowed by the algorithms here. These operations importantly do not include a square root.

Definition 7 (Simple Operations)**.** *A simple operation is one of the following [1]:*

- *the arithmetic operations* $+, -, \times,$ *and* $/$ *on real numbers,*
- *jumps conditioned on* $>, \geq, <, \leq, =$ *and* $\neq$ *comparisons of real numbers, and*
- *output* 0 *or* 1

Next we will consider definitions specific to logic for which the definitions that the decidability algorithms are defined. All of this notation is standard in [4].

Definition 8 ($\mathcal{P}$-atom)**.** *For our purposes (specifying the field to be* $\mathbb{R}$*), a* $\mathcal{P}$*-atom is a polynomial equality or inequality [4]. They are, if* $P \in \mathbb{R}[X_1, \ldots, X_k]$*:*

- $P = 0$
- $P \neq 0$
- $P > 0$
- $P < 0$

Definition 9 ($\mathcal{P}$-formula)**.** *A* $\mathcal{P}$*-formula is a combination of* $\wedge, \vee, \neg, \forall, \exists$ *with* $\mathcal{P}$*-atoms to form a logical statement [4].*

For example a $\mathcal{P}$-formula could be $\forall x \exists y (x^2 y + 2 > 0 \wedge y \leq 0)$.

2.2 Sample Complexity

Binary classification, the task assigning points in a space a label in $\{0,1\}$, is used in research and industry. The algorithm performing the classification is "learning" from the data. Below is a formal definition of learning from [1] that is used to study this phenomenon. First, we will need the definition of error with respect to the function the learning algorithm uses for classification.

Definition 10 (Error of h)**.** *With function $h : X \to \{0,1\}$ and P a probability distribution on $Z = X \times \{0,1\}$. The error of h with respect to P is defined as*

$$er_P(h) = P\{(x,y) \in Z : h(x) \neq y\}$$

Intuitively we want the learning algorithm to return a function with low error.

Definition 11 (Learning)**.** *Suppose that H is a class of functions that map from a set X to $\{0,1\}$. A learning algorithm L for H is a function*

$$L : \bigcup_{m=1}^{\infty} Z^m \to H$$

from the set of all training samples to H, with the following property:

- $\varepsilon \in (0,1)$
- $\delta \in (0,1)$

there is an integer $m_0(\varepsilon,\delta)$ such that if $m \geq m_0(\varepsilon,\delta)$ then,

- *for any probability distribution P on $Z = X \times \{0,1\}$*

if z is a training sample of length m, drawn randomly according to the product probability distribution P^m, then, with probability at least $1-\delta$, the hypothesis $L(z)$ output by L is such that

$$er_P(L(z)) < \inf_{g \in H} er_P(g) + \varepsilon$$

.

This defines what is known as (ε,δ)-learning H by L, with $m_0(\varepsilon,\delta)$ called the sufficient sample size. We want to know how much data a learning algorithm needs to achieve the desired results. Sample complexity is bounding $m_0(\varepsilon,\delta)$ for variable ε and δ.

Interestingly, the VC-dimension bounds the sample complexity of learning a function class. Like the definitions above, the theorem below is detailed in [1]. It shows that we can bound the sufficient sample complexity if we know the VC-dimension of H.

Theorem 1 (VC-Dimension Bounds Sample Complexity)**.** *There is a positive constant c such that the following holds. If H is a set of functions from a set X to* $\{0,1\}$ *and that H has VC-dimension* $d \geq 1$*, and L is a learning algorithm which minimizes sample error for H, then L is a learning algorithm for H and its sample complexity satisfies*

$$m_L(\varepsilon, \delta) \leq \frac{c}{\varepsilon^2}(d + \ln \frac{1}{\delta})$$

2.3 Interpolation

Given a set of points in space, the problem of interpolation is to produce a function which goes through the provided points. Interpolation is reproduction of a function given a set of points from domain to range. We will use the following result in the construction of our lower bound.

Theorem 2 (Multivariate Polynomial Interpolation)**.** *We can uniquely interpolate* $\binom{d+p}{p}$ *points with a polynomial* $P \in \mathbb{R}[X_1, \ldots, X_d]$ *so long as the determinant of the sample matrix for the* $\binom{d+p}{p}$ *points and is non-zero [13].*

The definition of the sample matrix can be found in [13], which corresponds with terms like full rank and general position. This constraint is listed for rigor's sake. This is not a concern for our purposes as it is always possible to pick points where the determinant of the sample matrix is nonzero.

2.4 Methods of Bounding VC-Dimension

There are two powerful methods for bounding complex range spaces. The first is composition arguments, where we break the range spaces into more simple parts then bound via composition. The second is circuit arguments, where computing set inclusion within a computational framework is used to derive an upper bound for the range space. The next theorem details the former of these two tools.

2.4.1 Composition

Theorem 3 (k-fold composition)**.** *Let $S_1 = (X, \mathcal{R}^1), \ldots, S_n = (K, \mathcal{R}^n)$ be range spaces with VC-dimension $\delta_1, \ldots, \delta_n$, respectively. Next, let $f(r_1, \ldots, r_n)$ be a function that maps any n-tuple of sets [1] $r_1 \in \mathcal{R}^1, \ldots, r_n \in \mathcal{R}^n$ into a subset of X. Consider the range set*

$$R' = \{f(r_1, \ldots, r_n) \mid r_1 \in \mathcal{R}_1, \ldots, r_n \in \mathcal{R}_n\}$$

and the associtated range space $T = (X, \mathcal{R}')$. Then the VC dimension of T is bounded by $O(n\delta \log n)$, where $\delta = \max_i \delta_i$ [10].

Next I will provide an example of how to use Theorem 3 to bound a range space. As far as we are aware this proof is original.

Claim 1 (Bound on Simple Polygons)**.** *The range space $(\mathbb{R}^2, \mathcal{P}_m)$ where $\mathcal{P}_m$ is the set of all simple polygons [2] with m sides has VC-dimension $O(m \log m)$ for $m > 3$.*

Proof. Let $B \in \mathcal{P}_n$ where $m > 3$. We know that the VC-dimension of convex polygons with m sides is $2m - 1$ [10]. It has been shown that all simple polygons have a triangular decomposition, consisting of $m - 2$ triangles [5]. Decompose B into $m - 2$ triangles. Each triangle [3] has VC-dimension 7. The set of all triangles bounds the set of all triangles decomposed from a polygon. By Theorem 3.6 the bound on the set of simple polygons is $O(m \log(m - 2))$. □

Similarly, circuit arguments provide a powerful tool for bounding VC-dimension of range spaces. This theorem is from the perspective of computational learning theory which uses function classes. In this case a is in the domain of a polynomial in $\mathbb{R}^d$ which "thresholds" the points in $\mathbb{R}^{d+1}$.

2.4.2 Circuits

Theorem 4 (Circuit Argument)**.** *Suppose h is a function from $\mathbb{R}^d \times \mathbb{R}^k$ to $\{0, 1\}$ and let*

$$H = \{x \mapsto h(a, x) : a \in \mathbb{R}^d\}$$

---

[1] Recall that $\mathcal{R}^n$ is a set of sets for $n = 1, \ldots, k$.

[2] Recall that a simple polygon is a polygon that does not intersect itself nor has holes.

[3] This is because triangles are convex with $m = 3$.

be the class determined by h. Suppose that h can be computed by an algorithm that takes as input the pair $(a, x) \in \mathbb{R}^d \times \mathbb{R}^k$ *and returns* $h(a, x)$ *after no more than t simple operations. Then* $VCdim(H) \leq 4d(t+2)$ *[1].*

There arises a difficulty in applying Theorem 4 to applications in geometry, that is the lack of a square root operation, which is needed to encode distance into our algorithms. Thus, when attempting to approach problems like Problem 1, distance between a point and curve becomes a barrier for applying Theorem 4.

2.5 Results for Combining Distance and Circuit Arguments

Lemma 12 in [9] details the successful evaluation of an inequality involving square roots using only simple operations. This implies that in some circumstances circuit arguments can be used for evaluating distance and other range spaces involving square roots.

Theorem 5. *Consider values* $a, b, c, d \in \mathbb{R}$ *with* $b, d \geq 0$*. We can compute the truth values of* $a + \sqrt{b} \leq c + \sqrt{d}$ *and* $a + \sqrt{b} \geq c + \sqrt{d}$ *using* $O(1)$ *simple operations.*

As far as we are aware this is the first use of circuit arguments being applied to evaluate expressions with square roots. This theorem is satisfactory for evaluation of a finite number of points or operations involving lines, for instance when working with polygonal curves. Yet, there arises difficulty when applying this theorem to the evaluation of a point and a polynomial as it would require the evaluation of the distances of an infinite number of points.

2.6 Algorithms in Real Algebraic Geometry

2.6.1 Introduction to Algebraic Sets

Though the algebraic geometry has a reputation for being vast, it can generally be thought of as the study of the solutions to polynomial systems. As we are working with polynomials it is reasonable that there are results within the field that could help bound the VC-dimension of range spaces if those range spaces are partly constructed from polynomials.

If $P \in \mathbb{R}[X_1, \ldots, X_d]$, an algebraic set (often called an algebraic variety) is

$$\{x \mid P(x) = 0, \text{ for } x \in \mathbb{R}^d\}$$

With this in mind a semialgebraic set is a finite union of polynomial equalities and polynomial inequalities. For instance:

$$x^2 - y \leq 0 \cup x - y > 0$$

is a semialgebraic set in the plane. Semialgebraic sets are central to our study.

2.6.2 Decidability Theorems and Algorithms

In Basu, Pollack, and Roy's seminal work Algorithms in Real Algebraic Geometry they detail a large number of algorithms on real polynomials. We will use a few results from this work. The first algorithm we will use detailed in [4] is Algorithm 9.5. It is used for counting roots of a univariate polynomial. The citation includes an extra parameter $Q \in \mathbb{R}[X]$ that represents a more general query called a Tarski-Query. By taking $Q = 1$ then a Tarski-query is equivalent to computing the number of roots as given in Sturm's theorem. Sturm's theorem is specifically for univariate polynomials.

Algorithm 1 (Univariate Tarski-Query)**.** *Given the following input we can compute the number of roots of a polynomial in the given complexity:*

- *Input:* $p \in R[X] \setminus \{0\}$.
- *Output: Number of elements in* $\{x \in \mathbb{R} \mid p(x) = 0\}$
- *Complexity:* $O(p + 1)$ *simple operations.*

Next we will use a result from [4] regarding decidability, specifically over the language that is the theory of real closed fields. As detailed in [4] Tarski showed that the theory of the real closed fields is decidable, as a consequence of what is now known as the Tarski–Seidenberg Theorem. Yet it was only until Collins' [7] use of cylindrical algebraic decomposition that a doubly exponential bound was found. Due to this bound it would not seem to give a good bound in practice.

There is a simpler problem which only allows for existential quantifiers. This problem is known as the existential theory of the reals, and [4] details an algorithm for its decidability. The complexity below was originally detailed and shown by James Renegar in [12].

Consider first order logical statements in the following form:

$$\exists X_1, \ldots \exists X_k \; F(X_1, \ldots, X_k)$$

where $F(X_1, \ldots, X_d)$ is a quantifier free $\mathcal{P}$-formula. If that statement is true or false then this is called the decision problem for the existential theory of the reals [4]. The next theorem gives a general and powerful result by showing the existence of an algorithm for deciding the truth such statements.

Theorem 6 (Existential Theory of The Reals)**.** *Let* $\mathbb{R}$ *be a real closed field. Let* $\mathcal{P} \subset \mathcal{R}[X_1, \ldots, X_k]$ *be a finite set of* s *polynomials each of degree at most* p*, and let* $\exists X_1, \ldots \exists X_k \; F(X_1, \ldots, X_d)$ *be a sentence, where* $F(X_1, \ldots, X_d)$ *is a quantifier free* $\mathcal{P}$*-formula. There exists an algorithm to decide the truth of the sentence with complexity* $s^{d+1} p^{O(d)}$ *in simple operations.*

CHAPTER 3

MAIN RESULTS

3.1 Problem Statement

In [6], Phillips poses the following problem. "Is there a bound on the VC-dimension of the range spaced defined by shapes formed by the Minkowski sum of a ball and a polynomial curve?" We detail the problem formally in the next problem.

Problem 1. *Let* $(\mathbb{R}^{d+1}, \mathcal{M}_p)$ *be a range space, where* $\mathcal{M}_p$ *is the Minkowski sum of a disk of any radius and a polynomial* P *in* $\mathbb{R}[X_1, \ldots, X_d]$ *where the degree of* P *is bound by* p*. Is the VC-dimension of* $(\mathbb{R}^{d+1}, M_p)$ *finite, and if so, what is its bound?*

Although evaluation of distance is important for geometric applications many of the existing tools for bounding the VC-dimension are unable to accommodate a square root, which is needed for Euclidean distance. It is not obvious how to use general circuit theorems or composition theorems to bound the VC-dimension in such cases.

We could try using Theorem 5, yet attempting to do so leads to performing an infinite number of operations. This is due to needing to evaluate the distance between a point and potentially an infinite number of points on the polynomial. We could try a composition argument yet the boundaries of the inflated polynomials are not polynomials. On the boundaries of the polynomials if the radius is sufficient there can be non-differentiable points due to clear "kinks" in the boundary. Thus it is not clear how to deconstruct the inflated polynomial into component parts. It would seem that other techniques are needed to determine if there exists a finite bound for the VC-dimension of the inflated polynomial range space.

3.2 Our Approach

In fact, the inflated polynomial range space, detailed in Problem 1, does have a finite VC-dimension. Our general approach for forming an upper bound is to perform a reduction to either counting roots of a polynomial or a logical decidability question within a theory that is known to be algebraically decidable with simple operations. Once we have a bound on the number of simple operations we can bound the VC-dimension using a circuit argument.

We will bound the VC-dimension of the inflated polynomial range space in two ways. First, for univariate polynomials we will reduce to real root counting then apply a circuit argument. Second, for the multivariate case we will reduce to existential theory of the reals. We include the two methods because we believe it demonstrates the flexibility of combining the algebraic geometry algorithms with circuit arguments.

We believe there are further results to be found from using such methods. To establish this point we will show that all semialgebraic sets composed of finite bounded degree polynomials have finite VC-dimension. This result, through the use of the circuit argument, establishes a connection between logic, specifically decidability algorithms, and VC-dimension. As VC-dimension is ubiquitous to other fields like probability and model theory perhaps there are further connections to establish. In conclusion we have solved the problem of bounding the VC-dimension of inflated polynomial range space and, in addition, bounded the VC-dimension of semialgebraic sets.

3.3 Proofs of Upper Bound of VC-Dimension of Inflated Range Space

First we will perform a reduction of Problem 1 to a problem that is easier to think about. We will state this result as a lemma.

Lemma 1. *Consider range space* $(\mathbb{R}^{d+1}, \mathcal{M}_p)$. *Given a query point* $w \in \mathbb{R}^{d+1}$ *and inflated polynomial* $P_r \in \mathcal{M}_p$. w *is incident with* P_r *if an only if*

$$\exists x_0 \in P(\mathbb{R}^d)(||w - (x_0, P(x_0))||_2 \leq r)$$

where P *is the inflated polynomial of* P_r, *and* $(x_0, P(x_0))$ *is appending the function value at* x_0 *to the end of* x_0.

Proof. If w is incident with inflated polynomial P_r, then there exists a point $x_0 \in \mathbb{R}^d$ where $(x_0, P(x_0))$ is on the curve P. w is incident with disk $D_r((x_0, P(x_0)))$. As w is within this disk then the distance $||w - (x_0, P(x_0))||_2$ must be less than or equal to r. Conversely, if there exists a point $x_0 \in P(\mathbb{R}^d)(||w - (x_0, P(x_0))|| \leq r)$ then w is incident with $D_r((x_0, P(x_0)))$ and thus the inflated polynomial P_r. □

The above lemma provides us the benefit of only searching for a point on the polynomial that is within a distance r. This could be viewed as an optimization problem. However finding a point on the polynomial that minimizes the distance is not needed. With this lemma we will now prove an upper bound on the VC-dimension using a Tarski-query.

3.3.1 Upper Bound of $(\mathbb{R}^2, \mathcal{M}_p)$ with Tarski-Query

Theorem 7. *Consider range space $(\mathbb{R}^2, \mathcal{M}_p)$ where $\mathcal{M}_p$ is composed of only univariate inflated polynomials. The VC-dimension of this space is $O(p)$*

Proof. Due to Lemma 1 we must find a point on the polynomial close enough to w.

$$||w - (x, P(x))||_2 \leq r$$

$$\sqrt{(w_1 - x_1)^2 + (w_2 - P(x))^2} \leq r$$

$$(w_1 - x_1)^2 + (w_2 - P(x))^2 - r^2 \leq 0$$

Notice that this is a polynomial inequality. As P is defined for all $\mathbb{R}$ the distance is unbounded. Notice, due to the squared terms, that the final polynomial has even degree. Therefore, to determine if there exists an x that satisfies the inequality above it is sufficient to count for roots of the polynomial. By Algorithm 1 we an count roots of univariate polynomials in $O(p + 1))$ simple operations, with p the degree of P. Now we will use a circuit argument, Theorem 4. Recall this bound is established by the number of simple operations and the number of free variables of the polynomial. In our case the number of free variables is just 1 and we have $p + 1$ simple operations. Thus the bound on the VC-dimension is $O(4 \cdot 1((p + 1) + 1) = O(p)$.

□

3.3.2 Upper Bound of $(\mathbb{R}^{d+1}, \mathcal{M}_p)$ with Existential Theory of The Reals

Now we will generalize to multivariate polynomials by using a decidability algorithm.

Theorem 8. *Consider range space* $(\mathbb{R}^{d+1}, \mathcal{M}_p)$.

The VC-dimension of this space is

$$O(dp^{O(d)})$$

Proof. Consider inflated polynomial P_r of degree p and fix $w \in \mathbb{R}^{d+1}$. Due to lemma 1 we must find a point on the polynomial close enough to w.

$$||w - (x, P(x))||_2 \leq r$$

$$\sqrt{(w_1 - x_1)^2 + \ldots + (w_d - x_d)^2 + (w_{d+1} - P(x))^2} \leq r$$

$$(w_1 - x_1)^2 + \ldots + (w_d - x_d)^2 + (w_{d+1} - P(x))^2 - r^2 \leq 0$$

As before this is a polynomial inequality only with more free variables. Now we will invoke the existential theory of the reals decidability algorithm with Theorem 3.6. To do this we need to put the final inequality into the logical structure desired by the algorithm.

$$\begin{aligned}
&(\exists x_1) \ldots (\exists x_d)(\\
&\qquad (w_1 - x_1)^2 + \ldots + (w_d - x_d)^2 + (w_{d+1} - P(x))^2 - r^2 = 0 \\
&\qquad \vee (w_1 - x_1)^2 + \ldots + (w_d - x_d)^2 + (w_{d+1} - P(x))^2 - r^2 < 0 \\
&\quad) \\
&\equiv \\
&\quad (\exists x_1) \ldots (\exists x_d)((w_1 - x_1)^2 + \ldots + (w_d - x_d)^2 + (w_{d+1} - P(x))^2 - r^2 = 0) \\
&\quad \vee \\
&\quad (\exists x_1) \ldots (\exists x_d)((w_1 - x_1)^2 + \ldots + (w_d - x_d)^2 + (w_{d+1} - P(x))^2 - r^2 < 0)
\end{aligned}$$

Thus we have 2 d-variate polynomials we must evaluate. The existential theory of the reals algorithm takes $O(p^d)$ simple operations to do evaluate for each $\mathcal{P}$-atom. Now we will use a circuit argument, Theorem 4. Recall this bound is established by the number of simple operations and the number of free variables of the polynomial. In our case the

number of free variables is d and we have $O(p^d)$ simple operations. Thus the bound on the VC-dimension for only one of our $\mathcal{P}$-atoms is

$$O(4 \cdot d(p^d + 1)) = O(dp^d)$$

Finally we use a composition argument, but we only have 2 pieces so it only increases the VC-dimension by a constant factor. Thus our final VC-dimension bound for $(\mathbb{R}^{d+1}, \mathcal{M}_p)$ is

$$O(2dp^d \log 2) = O(dp^d)$$

□

3.4 Lower Bound of $(\mathbb{R}^{d+1}, \mathcal{M}_p)$ with Interpolation

We show a tight lower bound of $\binom{d+p}{p}$ where p is the degree of the polynomial and d is the number of variables in the polynomial. This is not surprising as this is known how many points a multivariate polynomial can interpolate. In fact, we will shatter a set using interpolation in the proof below. For this proof we use a multivariate Lagrangian interpolation detailed by [13].

Theorem 9. *Let $(\mathbb{R}^{d+1}, \mathcal{M}_p)$ be a range space, where M_p is the Minkowski sum of a ball over polynomials in $\mathbb{R}[X_1, \ldots, X_d]$ of degree p. The lower bound of the VC-dimension of $(\mathbb{R}^{d+1}, \mathcal{M}_p)$ is $\binom{d+p}{p}$.*

Proof. Given $(\mathbb{R}^{d+1}, \mathcal{M}_p)$ consider X, a set of points in $\mathbb{R}^{d+1}$ where $|X| = \binom{d+p}{p}$ points such that the sample matrix's determinant as in [13] is nonzero. Let Z be a non empty element of the power set of X. To intersect all points in Z and none in $X \setminus Z$ we interpolate over Z and $\binom{d+p}{p} - |Z|$ *perturbed* points in $X \setminus Z$. We will perturb these points by adding ε to the final component of the points in $X \setminus Z$. Let $P_r \in \mathcal{P}_p$ and P be the polynomial from P_r. Recall that polynomial P is a function from $\mathbb{R}^d \to \mathbb{R}$. If we then interpolate using Lagrange interpolation detailed in [13] over the Z and the perturbed points of $X \setminus Z$ the function will not interpolate the original points of $X \setminus Z$. We know that this perturbing these points does not affect existence of the interpolant since changing the final component of our set does not change the determinant of the sample matrix. Therefore as we can interpolate any subset of $\binom{d+p}{p}$ points in this way the VC-dimension of the range space must be at least $\binom{d+p}{p}$. □

Notice that if we are dealing with univariate polynomials then the curve lives in $\mathbb{R}^2$ and can shatter according to the above theorem $\binom{1+p}{p} = p + 1$. Note that this is a lower bound due to the fact that we are not using the expressiveness of radius of the inflated polynomial to our advantage. Yet as the modification of the radius affects the inflated polynomial globally not locally its expressiveness is limited.

3.5 Tight VC-Dimension Bound for Inflated Polynomial Range Space

We have an upper bound and a lower bound on the VC-dimension of the inflated polynomial range space. Notice that as the dimension increases the bounds become tighter. Also notice that we have established $\theta(p)$ VC-dimension for univariate inflated polynomials.

3.6 Proofs of Upper Bound of VC-Dimension of Semialgebraic Sets

Besides the dimension of the space there are two factors on the VC-dimension bound for semialgebraic sets, how many polynomials we use and the maximum degree of those polynomials. Let us fix $s, p \in \mathbb{N}$ for the number of polynomials and bound on the degree, respectively. As mentioned in the prior chapter all semialgebraic sets are a union of polynomial equalities and inequalities. The proof technique is to represent the set as a disjunction of polynomial equalities and inequalities. Then we will use the existential theory of the reals algorithm on each of these equalities to test satisfiability. We then use a composition argument to combine all equalities and inequalities.

Theorem 10. *The VC-dimension of a semialgebraic set in $\mathbb{R}^{d+1}$ composed of s polynomials of bounded degree p is $O(sdp^d \log s)$.*

Proof. Consider $P_1, \ldots, P_s$ $\mathcal{P}$-atoms, Definition 8, composed of d-variate polynomials and all bounded by degree p. The set $P_1 \vee \ldots \vee P_s$ is semialgebraic. We will ask if the following sentence is satisfied.

$$\exists X_1, \ldots \exists X_d (P_1, \ldots, P_s)$$

Use the existential theory of the reals Algorithm on each P_i. This takes $O(p^d)$ simple operations. Therefore, via a circuit argument, the VC-dimension of each $\mathcal{P}$-atom is $O(dp^d)$.

We will then use the composition argument, theorem , to get bounds

$$O(sdp^d \log s)$$

□

CHAPTER 4

APPLICATIONS

Now we will briefly detail a few applications of this work. First, we will talk about learnability of a polynomial with margin. Second, we will talk about splines and spline learnability. Finally, we will cover learnability of semialgebraic sets.

4.1 Inflated Polynomial Classification

The VC-dimension bounds the sample complexity for learning a binary classifier [1]. Polynomials are often used in modeling problems due to their well behaved nature. If we have a ground function that is approximately polynomial but it has a "margin" similar to an inflated polynomial then we now have sample complexity bounds for binary classifier. In particular, here classification is defined for a point as labeling that point 1 if within the inflated polynomial otherwise 0. By Theorem 1 we know that the sample complexity for (ε, δ) learning is bounded by

$$m_L(\varepsilon, \delta) \leq O(\frac{1}{\varepsilon^2}(dp^d + \ln \frac{1}{\delta}))$$

This also bounds the sample complexity of "smoothed range spaces" [11] with polynomial boundaries; these allow points near the margin to be given a loss in an intuitive geometric fashion. Previously, the complexity of this problem was unknown.

4.2 Inflated Univariate Spline Classification

An inflated spline is a polynomial spline that has been inflated with radius r. A spline is a piecewise polynomial. To derive the VC-dimension for univariate splines we must add two inequalities for each "piece," where a "piece" in this context is the interval in which a specific polynomial is defined. Thus for each polynomial the number of inequalities we need to consider for the existential theory of the reals is three. Thus the complexity remains the same. Now we must only apply a composition argument over each piece to

get the VC-dimension. Therefore we find the following bound

$$O(np \lg n)$$

where n is the number of polynomial pieces used. In addition, like above we have the complexity for learning inflated polynomial splines.

$$m_L(\varepsilon, \delta) \leq O(\frac{1}{\varepsilon^2}(np \lg n + \ln \frac{1}{\delta}))$$

This applies immediately to induction of a trajectory. Suppose we are unaware of a person's location over time and that we make the modeling assumption that he or she traced a piecewise polynomial path. A piecewise polynomial curve, perhaps a natural cubic spline, is a more natural assumption than a piecewise polygonal curve. For example, we may want to know the path taken by a super-spreader of COVID-19 traveling within a city containing static points (positions of people) and we believe they traced a spline path. If someone is within 6 feet, radius of the inflated spline, the bystander is at risk of infection of the virus. How many people do we need to test (binary classification) to up to $1 - \varepsilon$ accuracy to induce the path the super-spreader took, with probability $1 - \delta$. It was previously unknown how many people are required to be tested, yet in $\mathbb{R}^2$ with n polynomial pieces each with bounded degree p we know now the bound is

$$m_L(\varepsilon, \delta) \leq O(\frac{1}{\varepsilon^2}(np \lg n + \ln \frac{1}{\delta}))$$

4.3 Semialgebraic Set learnability

All semialgebraic sets form a large function class, for which we now can bound sample complexity. If attempting to (ε, δ)-learn with semialgebraic sets the sample complexity is the following:

$$m_L(\varepsilon, \delta) \leq O(\frac{1}{\varepsilon^2}(sdp^d \log s + \ln \frac{1}{\delta}))$$

This provides a general result for the union of all polynomial equalities and inequalities.

CHAPTER 5

CONCLUSION

In conclusion, we have developed a new mechanism to bound range spaces involving polynomials. We solved the open problem of bounding the VC-dimension of inflated polynomials. We have provided an upper bound for the VC-dimension of semialgebraic sets. Finally, we established a connection between more abstract fields of logic and computational learning theory.

REFERENCES

[1] M. Anthony and P. L. Bartlett, *Neural Network Learning: Theoretical Foundations*, Cambridge University Press, Cambridge, 1999.

[2] M. Aschenbrenner, A. Dolich, D. Haskell, D. Macpherson, and S. Starchenko, *Vapnik-chervonenkis density in some theories without the independence property, i*, Transactions of the American Mathematical Society, 368 (2011).

[3] S. Basu, *Combinatorial complexity in o-minimal geometry*, Proceedings of The London Mathematical Society - PROC LONDON MATH SOC, 100 (2006).

[4] S. Basu, R. Pollack, and R. M.F., *Algorithms in Real Algebraic Geometry*, vol. 10, Springer-Verlag, Berlin Heidelberg, 2 ed., 2006.

[5] M. Berg, M. Kreveld, M. Overmars, and O. Cheong, *Computational Geometry: Algorithms and Applications*, Springer-Verlag, Berlin Heidelberg, January 2008.

[6] O. Cheong, A. Driemel, and J. Erickson, *Computational Geometry (Dagstuhl Seminar 17171)*, Dagstuhl Reports, 7 (2017), pp. 107–127.

[7] G. E. Collins, *Quantifier elimination for real closed fields by cylindrical algebraic decompostion*, in Automata Theory and Formal Languages, H. Brakhage, ed., Berlin, Heidelberg, 1975, Springer, Berlin Heidelberg, pp. 134–183.

[8] A. Driemel, J. M. Phillips, and I. Psarros, *The vc dimension of metric balls under fréchet and hausdorff distances*, Symposium on Computational Geometry, (2019).

[9] A. Driemel, J. M. Phillips, I. Psarros, and A. Nusser, *The VC dimension of metric balls under fréchet and hausdorff distances*, CoRR, abs/1903.03211 (2019).

[10] S. Har-peled, *Geometric Approximation Algorithms*, American Mathematical Society, Providence, Rhode Island, USA, 2011.

[11] J. M. Phillips and Y. Zheng, *Subsampling in smoothed range spaces*, in Algorithmic Learning Theory, K. Chaudhuri, C. GENTILE, and S. Zilles, eds., Cham, 2015, Springer International Publishing, pp. 224–238.

[12] J. Renegar, *On the computational complexity and geometry of the first-order theory of the reals, part i: Introduction. preliminaries. the geometry of semi-algebraic sets. the decision problem for the existential theory of the reals*, J. Symb. Comput., 13 (1992), pp. 255–300.

[13] K. Saniee, *A simple expression for multivariate lagrange interpolation*, SIAM Undergraduate Research Online, 1 (2007).

[14] V. N. Vapnik and A. Y. Chervonenkis, *On the Uniform Convergence of Relative Frequencies of Events to Their Probabilities*, Springer International Publishing, Cham, 2015, pp. 11–30.