Using Existential Theory of the Reals to Bound VC Dimension

Austin Watkins* Jeff M. Phillips†

Abstract

We provide new bounds on the VC dimension of range spaces beyond logical compositions of polynomials and other discrete geometric shapes. Our results address the VC dimension of a seemingly simple class of range spaces we call inflated polynomials, which are defined as the Minkowski sum of a polynomial and a ball; in $\mathbb{R}^2$ with degree p the VC dimension is $\Theta(p)$, and in $\mathbb{R}^d$ the bound is $O(dp^{O(d)})$. This addresses natural questions on learnability in the adversarially-robust setting for polynomial classifiers and of polynomially-defined trajectories. We use a connection between algebraic geometry and classic circuit-based approaches of bounding the VC dimension to derive our results. We believe this connection and our general results may find other applications in learning theory, range searching, and other aspects of computational geometry where the VC dimension plays a key role.

1 Introduction

This paper studies the VC dimension and learnability of regions defined by offsets from polynomial curves and surfaces, which we call *inflated polynomials*. These offsets are no longer polynomial and so little to nothing is known about the learnability of a large family of classes that arise this way. We provide new VC dimension bounds for this family of objects by a connection to the existential theory of the reals. Application of these inflated polynomials are broad and we highlight implications in sweeping out the region around a polynomial curve and in adversarially-robust learning.

The Vapnik-Chervonenkis-dimension (VC dimension) [37] is the central combinatorial complexity score for a range space or a function class. It intricately ties into many aspects of learning theory [1] where it bounds how many data samples are needed to learn over a function class, model theory [2, 4] where it ties into the rich structure of algebraic geometry, big data [15] where it governs the size and runtime for creating coresets, computational geometry [19, 5] where it describes the size of a hitting set, and data structures [8] where it characterizes a class of ranges that admit a near-linear size data structure which allows for sub-linear time range queries. In this paper, we significantly generalize the approaches to analyze function classes defined through non-polynomial and existential formulations.

Inflated polynomials. In particular, in this paper we focus on ranges defined by the Minkowski sum of a Euclidean ball and a polynomial; we call these *inflated polynomials*. A simple example of an inflated parabola in $\mathbb{R}^2$ is shown in Figure 1. In particular, observe that the boundary of this shape is *not* a polynomial, as clearly evidenced by the cusp point, directly above the minimum. Thus, due to this non-polynomial nature, among other complexities, the VC dimension of such shapes have no known bound [9]. Let us highlight two other grounded scenarios where such questions arise.

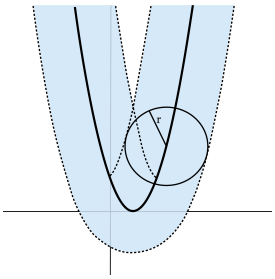

Figure 1: The inflated polynomial, shown in blue, of $(x-1)^2$ with radius r.

First, consider learning a polynomial classifier robustly, in the sense that it should protect against adversarial examples [35]. Typically, the goal is to learn a classifier so no, or few, correctly labeled examples can cross the decision boundary with small perturbations. While this is perhaps most problematic in complex classifiers [35, 18], learnability of robust classifiers has mostly been studied formally [34, 14, 29, 25, 16, 10] for linear (or near-linear) classifiers and/or when data classes have specific and known (uniform, Gaussian, accurate under Gaussian noise) distributions. While polynomial (and other kernel classifiers) can be "linearized" so the inner-product acts as a linear dot-product, this distance no longer measures the amount of perturbation required in the input space needed for a data point to cross the decision boundary. In particular, one goal is to learn a perfect polynomial classifier so that no data points are within a distance r of the decision boundary (measured using Euclidean distance in the input space). As we elaborate in Section 4, the number of samples needed to ensure that such a perfect and r-robust polynomial classifier on the sample will ensure

*School of Computer Science, Johns Hopkins University, awatki29@jhu.edu; This research was supported, in part, by DARPA GARD award HR00112020004 and NSF CAREER award IIS-1943251.

†University of Utah, School of Computing, jeffp@cs.utah.edu; Thanks to NSF IIS-1816149, CCF-2115677, and Visa Research.

at most ε-fraction of all data (with probability $1-\delta$) will be at least a distance r from the polynomial decision boundary is $O(\frac{\nu}{\varepsilon}\log\frac{\nu}{\varepsilon\delta})$, where ν is the VC dimension of the inflated polynomial around the decision boundary.

Second, consider a drone which moves through a neighborhood and transmits malicious computer code to Wi-Fi routers. We do not know the exact drone trajectory, but can model it as a polynomial curve of degree p, and know it is most effective within 30 meters. Thus its affected range is an inflated polynomial curve. How many Wi-Fi routers in the neighborhood do we need to randomly inspect to accurately estimate the drone path (i.e., we can predict the probability of malicious code within ε, with probability $1-\delta$)? The number of samples is $O(\frac{1}{\varepsilon^2}(\nu+\log\frac{1}{\delta}))$ where ν is the VC dimension of this inflated polynomial.

Results and techniques. In this paper we develop a family of techniques to bound the VC dimension of complex range spaces and apply them to the inflated polynomials and existentially defined sets. We build on traditional techniques for bounding the VC dimension [1, 17], which prior-to-this-work were restricted to polynomially defined sets, a few other specific options like sigmoid, and their compositions. This approach provides a set of simple operations and bounds the VC dimension by the number of such operations needed to determine inclusion in the set in question. For our work, as in [17], we combine this approach with a distinct set of tools from decision algorithms in logic, algebraic geometry, and the existential theory of the reals. While ultimately the proofs are simple; they rely on an observation that the computation model associated with most algebraic geometry is compatible with the simple operations of [1]. This allows us to bound the VC dimension of inflated polynomials and existentially-defined sets. Our main result is as follows:

Theorem 1 *The VC dimension of inflated polynomials in $\mathbb{R}^d$ of degree p is $O(dp^{O(d)})$, and for univariate polynomials, the bound is $\Theta(p)$.*

This provides the specific bound needed to address the two applications (polynomial path learning for detecting Wi-Fi manipulation and adversarially robust polynomial separators) highlighted above. In particular, for adversarially robust learning, we view this as an essential step in how to link the geometry of the decision boundary to the input space.

2 Background, Definitions, and Prior Work

As this paper unites several technical areas, we start with a fair number of definitions.

Polynomials. Central to our study are *real polynomials*, that is polynomials with real coefficients. When there are d variables $x_1,\ldots,x_d$, we denote these as $\mathbb{R}[x_1,\ldots,x_d]$. The *degree* of the polynomial is the maximum sum of exponents of the variables in any monomial. Such polynomials define functions f from $\mathbb{R}^d\to\mathbb{R}$. Hence they can also be viewed as d-dimensional objects in $\mathbb{R}^{d+1}$ which divide $\mathbb{R}^{d+1}$ into 3 sets. For $(x_1,\ldots,x_d,y)\in\mathbb{R}^{d+1}$ with $x=(x_1,\ldots,x_d)$, then it can be "below" if $y<f(x)$, "above" if $y>f(x)$, or "on" if $y=f(x)$.

The *Minkowski sum* between two sets $A,B\subset\mathbb{R}^{d+1}$ is the set of all pairwise additions between A and B, $\{a+b\mid a\in A,b\in B\}$, and is denoted $A\oplus B$. Let $\mathcal{M}_p^d$ be the set of all these *inflated polynomials* constructed as the Minkowski sum of the points "on" a polynomial (the set A) and a ball (the set B). Let $\mathcal{B}_r^d$ be the set of d-dimensional balls with radius r. That is, $\mathcal{M}_p^d=\{P\oplus B\mid B\in\mathcal{B}_r^{d+1},r\in\mathbb{R},P\in\mathbb{R}[x_1,\ldots,x_d]$ of degree at most $p\}$.

Range spaces and VC dimension. 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. $\mathcal{R}$ is often defined in terms of geometric objects. $\mathcal{R}$ could be the set of disks for $X=\mathbb{R}^2$, intervals on $X=\mathbb{R}$, linear halfspaces on $X=\mathbb{R}^d$, or as points below (or on) polynomials in $X=\mathbb{R}^d$. When $X\subset\mathbb{R}^d$ is set of points, then these example ranges $\mathcal{R}$ are the induced subset of X contained in some such shape.

Similar to a restriction over a family of functions to a subset of the domain, we will define the projection of range space $\mathcal{R}$ onto $Y\subset X$ as $\mathcal{R}_{|Y}:=\{R\cap Y\mid R\in\mathcal{R}\}$. For a range space $(X,\mathcal{R})$, if the projection $\mathcal{R}_{|Y}$ contains all subsets of Y, then $\mathcal{R}$ *shatters* Y. The *VC dimension* of $(X,\mathcal{R})$ is the maximum cardinality of any shattered subset of X.

In this paper we mostly consider real ground sets $X=\mathbb{R}^d$ or $X\subset\mathbb{R}^d$, in which case a range space is defined by its range sets, and thus for simplicity we refer to the VC dimension of range sets, where the real ground set and corresponding range space are implicit. Simple examples of VC dimension ν include: for disks in $\mathbb{R}^2$ then $\nu=3$, for intervals in $\mathbb{R}^2$ then $\nu=2$, for linear halfspaces in $\mathbb{R}^d$ then $\nu=d+1$, and for polynomials of degree p in $\mathbb{R}^d$ then $\nu=O(d^p)$. For polynomials of *any* degree in $\mathbb{R}^d$ then ν is unbounded – it is infinite.

2.1 Sample Complexity

For a domain X consider a classifier function $h:X\to\{0,1\}$, it maps any element $x\in X$ to either 0 or 1. Then given a probability distribution μ on $Z=X\times\{0,1\}$, the *error of* h with respect to μ, written $\mathrm{er}_\mu(h)$, is the probability that $(x,y)\sim\mu$ such that $h(x)\neq y$. The goal of classification, for some family of classifiers $\mathcal{H}$ and

some μ, is to find an $h \in \mathcal{H}$ with $\text{er}_\mu(h)$ as small as possible.

The other side of this seeks to minimize the number of samples required to achieve a certain error on a learned classifier $h \in \mathcal{H}$ for some family $\mathcal{H}$. For a set of m samples $P = (x_1, y_1), \ldots, (x_m, y_m)$ drawn i.i.d. from μ, let μ_P be the sample distribution induced by this set. Let $\varepsilon, \delta \in (0, 1)$. The *sample complexity* is defined as the smallest m such that $\text{er}_{\mu_P}(h) \leq \text{er}_\mu(h) + \varepsilon$ holds with probability at least $1 - \delta$ for all $h \in \mathcal{H}$. Then for parameters $\varepsilon, \delta \in (0, 1)$ we seek to minimize m so that for all $h \in \mathcal{H}$ we have $\text{er}_{\mu_P}(h) \leq \text{er}_\mu(h) + \varepsilon$, with probability at least $1 - \delta$. Since this condition holds for $h^* := \inf_{h \in \mathcal{H}} \text{er}_{\mu_P}(h)$, we can then "learn" h^* on P and know it will ε-approximately hold (with probability at least $1 - \delta$) on μ.

The family of classifiers $\mathcal{H}$ defines a range set, and when $X \subset \mathbb{R}^d$, the VC dimension ν of this range space $(X, \mathcal{H})$ determines the sample complexity m. Vapnik and Chervonenkis [37] and refined by [23, 1] show that $m = O(\frac{1}{\varepsilon^2}(\nu + \log \frac{1}{\delta}))$ samples are sufficient.

When a perfect classifier h exists, one where $\text{er}_\mu(h) = 0$, the sample complexity is lower; using only $m = O(\frac{\nu}{\varepsilon} \log \frac{\nu}{\varepsilon\delta})$ samples is sufficient [20].

2.2 Methods of Bounding VC Dimension

There are two powerful methods for bounding complex range spaces. The first is via composition arguments, where we break (via unions and intersections) a complex range space into simple ranges for which bounds are known, and then bound the complex range by aggregating the effect of the simple ranges. The second is via circuit arguments, where computing set inclusion within a computational framework is used to derive an upper bound for the range space.

Composition argument. Let $(X, \mathcal{R}_1), \ldots, (X, \mathcal{R}_s)$ be a set of range spaces with VC dimension $\nu_1, \ldots, \nu_s$, respectively. Let $f(r_1, \ldots, r_s)$ be a function defined element-wise over the domain X (i.e., unions and intersections), that maps any s-tuple of sets $r_1 \in \mathcal{R}^1, \ldots, r_s \in \mathcal{R}^s$ into a subset of X. That is, f corresponds with a fixed logical formula (i.e., composed of $\vee$s and $\wedge$s) over s binary values determined by if $x \in X$ is in each range r_i. A element $x \in X$ is in the composite range $f(r_1, \ldots, r_s)$ if the logical function returns 1. This process defines a composition range set $\mathcal{R}^\oplus = \{f(r_1, \ldots, r_s) \mid r_1 \in \mathcal{R}_1, \ldots, r_s \in \mathcal{R}_s\}$. Har-Peled [19][Theorem 5.22] shows for the VC dimension of the associated range space $(X, \mathcal{R}^\oplus)$ is bounded by $O(s\nu(1 + \log s))$ where $\nu = \max\{\nu_i\}_{i=1}^s$.

Circuit argument. Goldberg and Jerrum [17], and slightly generalized to this form [1][Theorem 8.4], uses a circuit of *simple operations*, defined to consist of

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

Then suppose h_a is a function from $\mathbb{R}^d$ to $\{0, 1\}$ parameterized by $a \in \mathbb{R}^k$. Let h_a define a range $H_a = \{x \in X \subset \mathbb{R}^d \mid h_a(x) = 1\}$ from the associated family of ranges $\mathcal{H} = \{R_a \mid a \in \mathbb{R}^k\}$. Suppose that h_a can be computed by an algorithm that takes as input the pair $(x, a) \in \mathbb{R}^d \times \mathbb{R}^k$ and returns $h_a(x)$ after no more than t simple operations. Then the VC dimension of $(\mathbb{R}^d, \mathcal{H})$ is at most $4d(t + 2)$.

While this approach (perhaps in combination with composition arguments) seems like it can be applied to handle most geometrically defined range spaces (say including inflated polynomials), there is an important omission from the simple operations: the square root operation. A square root is needed, for instance, to encode distance in a radius r ball. More importantly, simple composition arguments cannot be made, since an inflated polynomial is a union of an infinite number of these balls. Towards addressing some such goals (with respect to range spaces defined by polynomial curves), [13][Lemma 12] provides a special case where one can handle a square root inside of the circuit argument: Consider values $a, b, c, d \in \mathbb{R}$ with $b, d \geq 0$, then one 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. However, restricting to this use of the square root, to apply this to general range sets in a metric space where the square root is needed, such as inflated polynomials, one would need to perform this comparison at an infinite number of points, or a composition of an infinite number of sets. So we will still require more powerful machinery from the existential theory of the reals in real algebraic geometry.

2.3 Algorithms in Real Algebraic Geometry

We next focus on the interpretation of algebraic geometry through the perspective of solutions to polynomial systems. We will mostly follow notation from [3].

P-atoms for P-formulas. For our purposes (specifying the field to be $\mathbb{R}$), a $\mathcal{P}$*-atom* is a polynomial equality or inequality; if $P \in \mathbb{R}[x_1, \ldots, x_d]$ then the options are $P = 0$, $P \neq 0$, $P > 0$, or $P < 0$. Similarly, a $\mathcal{P}$*-formula* is a combination of $\wedge, \vee, \neg, \forall, \exists$ with $\mathcal{P}$-atoms to form a logical statement. For example a $\mathcal{P}$-formula could be $\forall x \exists y (x^2 y + 2 > 0 \wedge y \leq 0)$. 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 ($\mathbb{R}^2$). In [3] they detail a large number of algorithms on real polynomials. We will

use two key results from this work: Tarski queries and existential decidability over a subset of $\mathcal{P}$-formulas.

Arithmetic operations for algebraic geometry. As detailed in [4], the complexity of algorithms within algebraic geometry is given in terms of specified allowable arithmetic operations between elements of a chosen set. These operations and a chosen set define the structure of an algorithm. The structures which concern us are a ring, an integral domain, and an ordered integral domain. A *ring* structure defines the allowable operations to be $+, -, \times,$ and $= 0$, where $= 0$ is the unary operation of deciding if an element in the ring is zero. An *integral domain* structure defines, in addition to the ring structure, exact division $/$, between two elements given that division will be in the integral domain. An *ordered integral domain* structure defines, in addition to the integral domain structure, comparison between elements with $>, =,$ and $<$ operations.

Importantly, $\mathbb{R}$ is an ordered integral domain, and the "simple operations" in the circuit argument [17, 1] include all allowable arithmetic operations for a ordered integral domain. Hence a bound on arithmetic operations provides a bound on simple operations.

Univariate Tarski queries. The first real algebraic geometry result we use is Pollack's [3] Algorithm 9.5 for counting roots of a univariate polynomial. The cited form 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,[1] which is specifically for univariate polynomials. Ultimately, a *univariate Tarski query* can take in a univariate polynomial $P \in \mathbb{R}[x] \setminus \{0\}$ (that is, not including the trivial 0 polynomial), it outputs the number of elements in $\{x \in \mathbb{R} \mid P(x) = 0\}$ using $O(p + 1)$ simple operations.

Decidability. Next we will use a result regarding decidability, specifically over the language that is the theory of real closed fields. The Tarski–Seidenberg Theorem implies that the theory of the real closed fields is decidable. Yet it was only with Collins's [11] use of cylindrical algebraic decomposition that a doubly exponential bound was found. There is a simpler problem which only allows for existential quantifiers. This problem is known as the existential theory of the reals, with the first singly exponential complexity provided by Renegar [31].

Consider first-order logical statements in the following form: $\exists x_1, \ldots \exists x_d\ F(x_1, \ldots, x_d)$ where $F(x_1, \ldots, x_d)$ is a quantifier free $\mathcal{P}$-formula. Determining if that statement is true or false is called the *decision problem for the existential theory of the reals*. When $\mathcal{P} \subset \mathbb{R}[x_1, \ldots, x_d]$ is a finite set of s polynomials each of degree at most p, then there is an algorithm to decide the truth of $\exists x_1, \ldots, \exists x_d\ F(x_1, \ldots, x_d)$ using $s^{d+1}p^{O(d)}$ simple operations.

[1] see Theorem 2.50 and Theorem 2.61 in [3]

3 New VC Dimension Bounds

We begin with a two-dimensional bound for univariate inflated polynomials, based on Tarski queries. Then we generalize to d-dimensional inflated polynomials using Renegar's algorithm. We provide a lower bound of the same order, which matches when $d = 1$.

3.1 Upper Bound of VC Dimension for Inflated Polynomials

We first translate inflated polynomials into the language of existential algebraic geometry. Consider range space $(\mathbb{R}^{d+1}, \mathcal{M}_p^d)$, and a query point $w \in \mathbb{R}^{d+1}$, and inflated polynomial $P_r \in \mathcal{M}_p^d$. Then w is in P_r if and only if $\exists x_0 \in P(\mathbb{R}^d)(\|w - (x_0, P(x_0))\|_2 \leq r)$ where P is the polynomial of the inflated polynomial P_r, and $(x_0, P(x_0))$ is a point on that polynomial in $\mathbb{R}^{d+1}$. A univariate degree-p polynomial curve in $\mathbb{R}^2$ is an element of $(\mathbb{R}^2, \mathcal{M}_p^1)$, which is the domain of our first upper bound.

Theorem 2 *The range space $(\mathbb{R}^2, \mathcal{M}_p^1)$, where $\mathcal{M}_p^1$ is composed of only univariate inflated polynomials, has VC dimension $O(p)$.*

Proof. We must find a point on the polynomial close enough to $w = (w_1, w_2)$. And $\|w - (x, P(x))\|_2 \leq r$ implies $(w_1 - x)^2 + (w_2 - P(x))^2 - r^2 \leq 0$. Notice that this is a polynomial inequality. As P is defined for all $\mathbb{R}$, the distance is unbounded from above and, due to the squared terms, that the final polynomial has even degree. Therefore, to determine if there exists an x that satisfies the inequality (≤ 0) above it is sufficient to count roots of the polynomial. That is, since the number of roots is the number of times a set satisfies $= 0$, and it is $+\infty$ as $x \to \{-\infty, +\infty\}$, then if the number of roots is non-zero, there must exist a point x where the ≤ 0 condition is satisfied. Using univariate Tarski queries, we can count the real roots of univariate polynomials in $O(p + 1)$ simple operations, with p the degree of P.

Then we can use the circuit argument with a bound on the number of free variables $d = 1$ and depth in simple operations of the circuit as $t = O(p + 1)$. Hence the VC dimension is $4d(t + 2) = O(p)$. □

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

Theorem 3 *The range space $(\mathbb{R}^{d+1}, \mathcal{M}_p^d)$ of inflated polynomials in $\mathbb{R}^{d+1}$ has VC dimension $O(dp^{O(d)})$.*

Proof. Consider an inflated polynomial P_r of degree p and fix $w \in \mathbb{R}^{d+1}$. We must find a point $x \in \mathbb{R}^d$

on the polynomial close enough to w, satisfying $\|w - (x, P(x))\|_2 \leq r$, and equivalently $(w_1 - x_1)^2 + \ldots + (w_d - x_d)^2 + (w_{d+1} - P(x))^2 - r^2 \leq 0$. As before this is a polynomial inequality only with more free variables. Now we will invoke the existential theory of the reals decidability result of Renegar [31]. To do this we need to write the inequality into the logical structure desired by the algorithm.

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

Thus we have two d-variate polynomials we must evaluate. The existential theory of the reals algorithm takes $O(p^{O(d)})$ simple operations to evaluate for each $\mathcal{P}$-atom. Now we will use a circuit argument with d free variables/dimensions and $t = p^{O(d)}$ simple operations. Thus the VC dimension for one $\mathcal{P}$-atom is $O(dp^{O(d)})$. Then using a composition argument, we can combine these together increasing the bound only a constant factor. □

3.2 Lower Bound of VC Dimension for Inflated Polynomials via Interpolation

We show a 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. The proof uses that a polynomial $P \in \mathbb{R}[x_1, \ldots, x_d]$ can uniquely interpolate $\binom{d+p}{p}$ points in $\mathbb{R}^{d+1}$. With some perturbation, we can always shatter sets of this size.

Theorem 4 *The lower bound of the VC dimension of* $(\mathbb{R}^{d+1}, \mathcal{M}_p^d)$ *is* $\binom{d+p}{p}$.

Proof. Given $(\mathbb{R}^{d+1}, \mathcal{M}_p^d)$ 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 [33], 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 coordinate of the points in $X \setminus Z$. Let $P_r \in \mathcal{M}_p^d$ and P be the polynomial at the center of 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 [33] 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 perturbing these points does not affect the existence of the interpolant since changing the final coordinate of our set does not change the determinant of the sample matrix. We can then take r sufficiently small so that P_r does not contain any element of $X \setminus Z$. 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 $\binom{d+p}{p}$. □

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

Comment on tightness. We have an upper bound and a lower bound on the VC dimension of the inflated polynomial range space $(\mathbb{R}^{d+1}, \mathcal{M}_p^d)$. When p is constant then $\binom{d+p}{p} = \Theta(d^p)$ and when d is constant, then $\binom{d+p}{p} = \binom{d+p}{d} = \Theta(p^d)$. So for constant d, we have upper bound of $O(p^{O(d)})$ and lower bound of $\Omega(p^d)$. For $d = 1$, we have established $\Theta(p)$ VC dimension.

4 Application in Robust Adversarial Learning

We highlight an application in robust adversarial learning. Others implications can be found in Appendix A and by connecting to results in coresets [15], hitting sets [5], and range searching [8].

Adversarial attacks on classifiers refers to when someone makes small perturbations to input data so it fools a classifier. This phenomenon has been demonstrated in images, question answering, voice recognition, among other areas [35, 7, 34, 18]. Current defenses against adversarial robustness [24, 6, 12, 26, 27, 29] may have undesirable consequences, such as decreasing test accuracy, leading some to investigate a potential trade-off between accuracy and robustness [39, 36]. Yet, further investigation on robustness prevention methods and the separability of image datasets show that accuracy and robustness are obtainable for real-life data [38, 30]. Also, random smoothing of a classifier, a defense in which you randomly sample around points within the data to build robustness [10, 32, 22] has been effective in low dimensions yet may be untenable in high dimension [21].

To formalize this problem, we need to consider a classifier $h : \mathbb{R}^d \to \{-1, 1\}$. Let $B_\gamma(x) = \{x' \in \mathbb{R}^d \mid \|x' - x\|_2 \leq \gamma\}$ be the l_2 ball of radius γ around x, which describes the allowable perturbations around a data point $x \in \mathbb{R}^d$. We say a point $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$ is γ-*safe* from h if all $x' \in B_\gamma(x)$ has that $h(x') = y$; this implies it is sufficiently far from the decision boundary.

The γ-*error* can be measured on a distribution μ as the probability a sample $(x, y) \sim \mu$ is not γ-safe.

Prior work has defined a few notions of adversarial robustness. [14] considers the expected minimum Euclidean distance γ of a point x to decision boundary of h, formally: $\mathbb{E}_{(x,y)\sim\mu}[\min_{x'\in\mathbb{R}^d} \|x' - x\|_2$ such that $h(x') \neq y$.] This line of work uses specific function classes (some linear and quadratic classifiers) $\mathcal{H}$, which can use the value $h(x)$ to upper bound the expected perturbation radius γ for specific distributions (e.g., μ is Gaussian or uniform for each class). [34] defines *robust classification error* as the probability of drawing a γ-safe point from μ, mostly focusing on l_∞ perturbations. They show for linear models on Gaussian mixture distributions μ that more samples are needed to generalize wrt robust classifiers error than just classification error.

Our work, extends this to more general distributions, more complex (polynomial) classifiers, and to Euclidean perturbations. An important point to make is that while polynomials can be linearized to a higher-dimensional space so whether a point is classified correctly by the polynomial is preserved, this does *not* preserve the *distance* to the decision boundary, and so such techniques cannot be directly applied to understand the learnability of these polynomial classification problems.

Let $\mathcal{H}_p = \{\text{sgn} \circ h \mid h \in \mathcal{P}_p\}$ where $\mathcal{P}_p = \{f \in \mathbb{R}[X_1, \ldots, X_d] : \deg(f) \leq p\}$. The key insight is to describe a range space $(\mathbb{R}^d \times \{-1, 1\}, \mathcal{R}_p)$ derived from $\mathcal{P}_p$ and a robustness parameter $\gamma > 0$. Each function $h \in \mathcal{H}_p$ maps to a function $g : \mathbb{R}^d \times \{-1, 1\} \to \{-1, 1\}$, where $g(x, y) = 1$ if and only if (x, y) is γ-safe with respect to h. This takes on two cases, if $y = +1$, then $h(x)$ must be positive and x not in the γ-inflated polynomial around the decision boundary. Similarly, if $y = -1$, then $h(x)$ must be negative and x not in the γ-inflated polynomial.

Lemma 5 *The VC dimension of* $(\mathbb{R}^d \times \{-1, 1\}, \mathcal{R}_p)$ *is* $O(p)$ *for* $d = 1$ *and* $O(dp^{O(d)})$ *for* $d > 1$.

Proof. We can apply the composition argument detailed in Section 2.2 to the two d-dimensional ranges considered: at $y = +1$ the complement of an inflated polynomial and a polynomial, and at $y = -1$ the complement of an inflated polynomial and a polynomial, all of degree p; see example in Figure 2. All of these ranges are derived from the same polynomial $f \in \mathbb{R}[X_1, \ldots, X_d]$, but this only restricts the range space and does not increase the VC dimension. For the composition of a constant number of range spaces, the VC dimension is asymptotically the max of them. The stated bounds follow from the inflated polynomial bounds in Theorem 2 and Theorem 3. □

Next we analyze the learnability of polynomial classifiers which are γ-robust; those deemed successful on data which is γ-safe. The previous lemma demonstrated that such classifiers can be characterized with range spaces with bounded VC dimension, and directly linked to that for inflated polynomials.

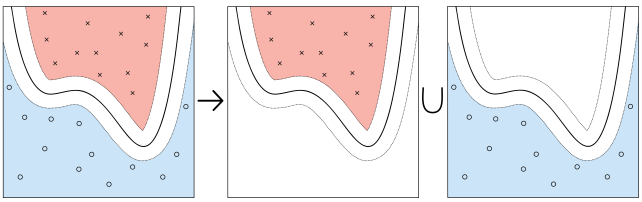

Figure 2: Decomposition of robust polynomial classification into ranges.

We first focus on non-agnostic learning, where 0 error can be achieved on a sample from family $\mathcal{H}$. The *non-agnostic robust sample complexity* of a family $\mathcal{H}$, a parameter $\gamma > 0$, and a distribution μ is the size of the smallest iid sample $P = \{(x_i, y_i)\} \subset \mu$ so that for any $h \in \mathcal{H}$ with γ-error of zero on μ_P, then with probability at least $1 - \delta$, it has at most γ-error of ε on μ.

Theorem 6 *For any* $\gamma > 0$*, the non-agnostic robust sample complexity is* $O(\frac{p}{\varepsilon} \log \frac{p}{\varepsilon\delta})$ *for univariate polynomials of degree at most* p *and* $O(\frac{pd^{O(d)}}{\varepsilon} \log \frac{pd^{O(d)}}{\varepsilon\delta})$ *for* d*-variate polynomials of degree at most* p.

Proof. Let any function $g \in \mathcal{R}_p$ have $g(x, y) = 1$ iff the point $(x, y) \in \mathbb{R}^d \times \{-1, 1\}$ is γ-safe. By assumption of the theorem there is a function $g \in \mathcal{R}_p$ with $\text{er}_{\mu_P}(g) = 0$ on a sample P. Then by bounding the VC dimension in Lemma 5 and applying the non-agnostic bound of [20], we obtain the claimed result. □

We can also apply this to agnostic settings, where we cannot guarantee a perfect classifier. The *agnostic robust sample complexity* of a family $\mathcal{H}$, a parameter $\gamma > 0$, and a distribution μ is the size of the smallest iid sample $P\{(x_i, y_i)\} \subset \mu$ so that for any $h \in \mathcal{H}$ with γ-error of η on μ_P, then with probability at least $1 - \delta$, it has at most γ-error of $\eta + \varepsilon$ on μ. By the same argument as in Theorem 6 but applying the more general bound of [23], we obtain the following result which has no assumptions on the distribution μ.

Theorem 7 *For any* $\gamma > 0$*, the agnostic robust sample complexity is* $O(\frac{1}{\varepsilon^2}(p + \log \frac{1}{\delta}))$ *for univariate polynomials of degree at most* p *and* $O(\frac{1}{\varepsilon^2}(pd^{O(d)} + \log \frac{1}{\delta}))$ *for* d*-variate polynomials of degree at most* p.

5 Conclusion & Discussion

This paper uses a combination of traditional techniques of bounding VC dimension and algorithms in algebraic geometry to bound the VC dimension of complex range spaces. These techniques are useful for ranges defined

with a combination of polynomials and existential quantifiers, such as geometric ranges of all points within a fixed Euclidean distance from an object. These apply as long as the geometric object can be described as a polynomial, or by n polynomial pieces. A key example is the class of inflated polynomials; for one such range, a point x_0 is inside if *there exists* a ball, centered on the defining polynomial, which contains x_0. These results have implications in range searching, hitting sets, and learning on swept out polynomial curves, as well as in adversarial learning.

ℓ_∞ perturbations. The applications to adversarially-robust sample complexity we develop focus on how inflated polynomials correspond with robust classifiers, which allow any ℓ_2 perturbation of data and still have the correct classification. Other work in this subarea has considered ℓ_∞ perturbations. We remark here that the VC-dimension of a polynomial of degree p under ℓ_∞ perturbation may not require analysis with existential theory of the reals. We claim that the Minkowski sum of an ℓ_∞ ball with a polynomial of degree p in $\mathbb{R}^2$ can be described as the composition of 4 polynomial classifiers of degree p, and $O(p)$ linear segments. Thus, since the VC dimension of any one of the polynomial parts is $O(p)$, the composition of the $O(p)$ linear parts is $O(\log p)$, and the composition of these two aspects is $O(p)$.

References

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

[2] Matthias Aschenbrenner et al. "Vapnik-Chervonenkis density in some theories without the independence property, I". In: *Trans. American Mathematical Society* 368 (Sept. 2011).

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

[4] Saugata Basu. "Combinatorial complexity in O-minimal geometry". In: *Proceedings of The London Mathematical Society* 100 (Dec. 2006).

[5] Hervé Brönnimann and Michael T. Goodrich. "Almost Optimal Set Covers in Finite VC-Dimension". In: *Disc. & Comp. Geom.* (1995).

[6] J. Buckman et al. "Thermometer Encoding: One Hot Way To Resist Adversarial Examples". In: *ICLR*. 2018.

[7] Krzysztof Chalupka, P. Perona, and F. Eberhardt. "Visual Causal Feature Learning". In: *UAI*. 2015.

[8] Bernard Chazelle and Emo Welzl. "Quasi-optimal range searching in spaces of finite VC-dimension". In: *Disc. & Comp. Geom.* 4.5 (1989), pp. 467–489.

[9] Orfried Cheong, Anne Driemel, and Jeff Erickson. "Computational Geometry (Dagstuhl Seminar 17171)". In: *Dagstuhl Reports* 7.4 (2017). Ed. by Otfried Cheong, Anne Driemel, and Jeff Erickson, pp. 107–127. ISSN: 2192-5283.

[10] Jeremy M. Cohen, Elan Rosenfeld, and J. Z. Kolter. "Certified Adversarial Robustness via Randomized Smoothing". In: *ICML*. 2019.

[11] George E. Collins. "Quantifier elimination for real closed fields by cylindrical algebraic decompostion". In: *Automata Theory and Formal Languages*. Ed. by H. Brakhage. Springer, Berlin Heidelberg, 1975, pp. 134–183.

[12] Guneet S. Dhillon et al. "Stochastic Activation Pruning for Robust Adversarial Defense". In: *ArXiv* abs/1803.01442 (2018).

[13] Anne Driemel et al. "The VC dimension of metric balls under Fréchet and Hausdorff distances". In: *Disc. & Comp. Geom.* 66.4 (2021), pp. 1351–1381.

[14] Alhussein Fawzi, Omar Fawzi, and P. Frossard. "Analysis of classifiers' robustness to adversarial perturbations". In: *Machine Learning* 107 (2017), pp. 481–508.

[15] Dan Feldman and Michael Langberg. "A unified framework for approximating and clustering data". In: *STOC*. 2011, pp. 569–578.

[16] Jean-Yves Franceschi, Alhussein Fawzi, and Omar Fawzi. "Robustness of classifiers to uniform ℓ_p and Gaussian noise". In: *AISTATS*. 2018.

[17] Paul W. Goldberg and Mark R. Jerrum. "Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers". In: *Machine Learning* 18 (1995), pp. 131–148.

[18] I. Goodfellow, Jonathon Shlens, and Christian Szegedy. "Explaining and Harnessing Adversarial Examples". In: *CoRR* abs/1412.6572 (2015).

[19] Sariel Har-peled. *Geometric Approximation Algorithms*. USA: American Mathematical Society, Providence, Rhode Island, 2011. ISBN: 0821849115.

[20] David Haussler and Emo Welzl. "epsilon-Nets and Simplex Range Queries." In: *Disc. & Comp. Geom.* 2 (1987), pp. 127–151.

[21] Aounon Kumar et al. "Curse of Dimensionality on Randomized Smoothing for Certifiable Robustness". In: *ArXiv* abs/2002.03239 (2020).

[22] Guang-He Lee et al. "Tight Certificates of Adversarial Robustness for Randomly Smoothed Classifiers". In: *NeurIPS*. 2019.

[23] Yi Li, Philip M. Long, and Aravind Srinivasan. "Improved Bounds on the Samples Complexity of Learning". In: *J. Comp. and Sys. Sci.* 62 (2001), pp. 516–527.

[24] A. Madry et al. "Towards Deep Learning Models Resistant to Adversarial Attacks". In: *ArXiv* abs/1706.06083 (2018).

[25] Seyed-Mohsen Moosavi-Dezfooli et al. "Robustness of Classifiers to Universal Perturbations: A Geometric Perspective". In: *International Conference on Learning Representations*. 2018.

[26] Seyed-Mohsen Moosavi-Dezfooli et al. "Robustness via Curvature Regularization, and Vice Versa". In: *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2019).

[27] Nicolas Papernot et al. "Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks". In: *2016 IEEE Symposium on Security and Privacy (SP)* (2016), pp. 582–597.

[28] Jeff M. Phillips and Yan Zheng. "Subsampling in Smoothed Range Spaces". In: *Algorithmic Learning Theory*. 2015.

[29] Chongli Qin et al. "Adversarial Robustness through Local Linearization". In: *NeurIPS*. 2019.

[30] Aditi Raghunathan et al. "Understanding and Mitigating the Tradeoff Between Robustness and Accuracy". In: *ArXiv* abs/2002.10716 (2020).

[31] James 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". In: *J. Symb. Comput.* 13 (1992), pp. 255–300.

[32] Hadi Salman et al. "Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers". In: *NeurIPS*. 2019.

[33] Kamron Saniee. "A Simple Expression for Multivariate Lagrange Interpolation". In: *SIAM Undergraduate Research Online* (Jan. 2007).

[34] L. Schmidt et al. "Adversarially Robust Generalization Requires More Data". In: *NeurIPS*. 2018.

[35] Christian Szegedy et al. "Intriguing properties of neural networks". In: *CoRR* abs/1312.6199 (2014).

[36] D. Tsipras et al. "Robustness May Be at Odds with Accuracy". In: *arXiv: Machine Learning* (2019).

[37] Vladimir Vapnik and Alexey Chervonenkis. "On the Uniform Convergence of Relative Frequencies of Events to their Probabilities". In: *Theory of Probability and its Applications* 16 (1971).

[38] Y. Yang et al. "Adversarial Robustness Through Local Lipschitzness". In: *ArXiv* abs/2003.02460 (2020).

[39] Hongyang Zhang et al. "Theoretically Principled Trade-off between Robustness and Accuracy". In: *ICML*. 2019.

A Additional Implications

Smoothed range spaces. Another application related to robust adversarial learning is the idea of a "smoothed range space" [28], where the misclassification error is replaced around a binary decision boundary with a continuous function, where significantly misclassified points are given a penalty of 1, but points close to the boundary (under Euclidean distance) are given a penalty between 0 and 1 according to a continuous rate based on how close. Zheng and Phillips [28] showed that the VC dimension of the decision boundary expanded by a Euclidean distance of r in all directions (i.e., inflated the decision boundary) governs the sample complexity of this task. However, this bound was unknown for polynomial decision boundaries [9] until this paper. The relevant VC dimension is that of an inflated polynomial.

Inflated univariate spline classification. An inflated spline is a polynomial spline that has been inflated with radius r. A spline is a piecewise polynomial that preserves stronger continuity between pieces. Suppose we are unaware of an object's (perhaps a person or vehicle) location over time and that we make a modeling assumption that the object is traveling along a piecewise polynomial path. A piecewise polynomial curve, perhaps a natural cubic spline, could be a more natural assumption than a piecewise polygonal curve. Suppose there is a low-flying unmanned aerial vehicle (UAV) with a radio jamming device which is disrupting cellular and GPS signals within r meters. We would like to approximate the UAV's trajectory over time. How many devices with radio sensors (cell towers, GPS, etc.) do we need to test (build a binary classifier) with up to $1-\varepsilon$ accuracy, to induce the path the object took, with probability $1-\delta$. It was previously unknown how many radio sensors 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 $m = O(\frac{1}{\varepsilon^2}(np\log n + \log\frac{1}{\delta}))$. The specific application described in the Introduction with a polynomial curve, has $n = 1$, so the specific bound in that case is $m = O(\frac{1}{\varepsilon^2}(p + \log\frac{1}{\delta}))$.

Theorem 8 *If points $x \in \mathbb{R}^2$ within r distance of a univariate polynomial spline are classified as* 1 *and points outside r are classified as -1, then to induce a trajectory with ε error, $m = O(\frac{1}{\varepsilon^2}(np\log n + \log\frac{1}{\delta}))$ points randomly chosen are sufficient, with probability $1-\delta$.*

Proof. In $\mathbb{R}^2$, by Theorem 2, the VC dimension associated with each piece is $O(p)$, if its degree is bounded by p. Now we must only apply a composition argument over each piece to get the VC dimension. Therefore we find the following bound $O(np\log n)$ where n is the number of polynomial pieces used. Hence, the sample complexity for learning on inflated polynomial splines is $m = O(\frac{1}{\varepsilon^2}(np\log n + \log\frac{1}{\delta}))$. □