This is the published version of:

Blackmore, Kim L., Williamson, Robert C., & Mareels, Iven M. Y.  
Decision region approximation by polynomials or neural networks  

© 1997 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.

This material is presented to ensure timely dissemination of scholarly and technical work. Copyright and all rights therein are retained by authors or by other copyright holders. All persons copying this information are expected to adhere to the terms and constraints invoked by each author's copyright. In most cases, these works may not be reposted without the explicit permission of the copyright holder.
Decision Region Approximation by Polynomials or Neural Networks

Kim L. Blackmore, Robert C. Williamson, Member, IEEE, and Iven M. Y. Mareels, Senior Member, IEEE

Abstract—We give degree of approximation results for decision regions which are defined by polynomial and neural network parametrizations. The volume of the misclassified region is used to measure the approximation error, and results for the degree of \( L_1 \), approximation of functions are used. For polynomial parametrizations, we show that the degree of approximation is at least 1, whereas for neural network parametrizations we prove the slightly weaker result that the degree of approximation is at least \( r \), where \( r \) can be any number in the open interval \((0, 1)\).

Index Terms—Classification, decision region, neural networks, polynomials, rate of approximation.

I. INTRODUCTION

D ECISION regions arise in machine learning problems of sorting or classification of data [1]. Points contained in the decision region are positively classified, and points outside the decision region are negatively classified. For a decision region \( D \subset \mathbb{R}^n \), this classification can be described by the discriminant function

\[
y_D(x) = \begin{cases} 
1, & \text{if } x \in D \\
-1, & \text{otherwise.}
\end{cases}
\] (1)

The learning task is to use examples of classified points to be able to correctly classify all possible points.

In neural network learning, decision boundaries are often represented as zero sets of certain functions, with points contained in the decision region yielding positive values of the function, and points outside the decision region yielding negative values [2]. In this case, the learning task is to use examples of correctly classified points to identify a parameter \( a \in \mathbb{R}^m \) for which the set \( \{x : f(a, x) \geq 0\} \), called the positive domain of \( f(a, \cdot) \), matches the true decision region.

For the purposes of analyzing a learning algorithm, it is useful to assume that a suitable value of the parameter exists. However, there is no general reason why such an assumption is satisfied in practice. Even if there is a class of functions \( f(\cdot, \cdot) \) and a parameter \( a \) such that the positive domain of \( f(a, \cdot) \) matches the true decision region, there is usually no way of identifying this class \( a \) priori. It is therefore useful to know how well particular classes of functions can approximate decision regions with prescribed general properties. In particular, it is important to know how fast the approximation error decreases as the approximating class becomes more complicated—e.g., as the degree of a polynomial or the number of nodes of a neural network increases.

The question of approximation of functions has been widely studied. The classical Weierstrass Theorem showed that polynomials are universal approximators [3] (in the sense that they are dense in the space of continuous functions on an interval). Many other classes have been shown to be universal approximators, including those defined by neural networks [4]. Degree of approximation results tell the user how complicated a class of approximating functions must be in order to guarantee a certain degree of accuracy of the best approximation. The classical Jackson Theorem [5] is the first example of this. Hornik [6], Barron [7], Mhaskar and Michelli [8], [9], Mhaskar [10], Darken et al [11], and Hornik et al [12] give degree of approximation results for neural networks.

The problem of approximating sets, rather than functions, has received some attention in the literature. Approximation of (unparametrized) sets and curves has been studied for pattern recognition and computer vision purposes [13]–[15]. The approach is quite different from the approach here. Theoretical work can be grouped according to two basic approaches—namely, explicit and implicit parametrizations. “Explicit parametrization” refers to frameworks where the decision boundary is parametrized. For example, if the decision region is a set in \( \mathbb{R}^n \), the decision boundary might be considered the graph of a function on \( \mathbb{R}^{n-1} \), or a combination of such graphs. “Implicit parametrization” refers to frameworks (as used in this work) where the decision region is the positive domain of some function.

Most existing work is in terms of explicit parametrizations [16]. For instance, Korostelev and Tsybakov [17], [18] consider the estimation (from sample data) of decision regions. Although they consider nonparametric estimation, it is, in fact, the explicit rather than implicit framework as defined above (they reduce the problem to estimating functions whose graphs make up parts of the decision boundary). In a similar vein, Dudley [19] and Scheiblin [20] have determined the metric entropy of certain smooth curves.

Regarding the implicit problem, Mhaskar [10] gives a universal approximation type result for approximation by positive domains of certain neural network functions. Ivanov [21] summarizes many problems in algebraic geometry con-
cerned with the question of when a smooth manifold can be approximated by a real algebraic set but does not address the degree of approximation question. In work similar to that described in [21], Broglia and Tognoli [22] consider when a $C^\infty$ function can be approximated by certain classes of functions without changing the positive domain.

In this paper, we use function approximation results to determine the degree of approximation of decision regions by positive domains of polynomial functions and neural networks. We consider the $L_A$ approximation of the discriminant function $y_D(x)$. This implies a bound on the $L_A$ distance between $y_D(x)$ and $\sgn(f(x))$, where $f(x)$ is the approximating polynomial or neural network function. We use a result from differential geometry to link this distance with the size of the misclassified volume. Since most learning problems can be analyzed probabilistically, the volume of the misclassified region has a natural interpretation as the probability of misclassification by the approximate decision region when the data are drawn from a uniform distribution over the input space.

The next section of this paper contains a formal statement of the degree of approximation problem for decision regions. In Section III we define a corridor around the decision boundary, and give a result concerning its volume, which is used in the later sections. Section IV contains the polynomial approximation results. Our main result is Theorem 8, which says that the volume of the misclassified region when a decision region with smooth boundary is approximated by the positive domain of a polynomial of degree $d$, goes to zero at least as fast as $d^{-r}$. By “smooth boundary” we mean essentially that the boundary is a finite union of $n-1$-dimensional submanifolds. When both cases are allowed, $M$ is usually called a submanifold with boundary. We allow both cases because our consideration of decision regions confined to a compact domain implies that many interesting decision boundaries are not true submanifolds.

**Definition 1:** A set $M \subset \mathbb{R}^n$ is an $n-1$-dimensional submanifold of $\mathbb{R}^n$ if for every $x \in M$, there exists an open neighborhood $U \subset \mathbb{R}^n$ of $x$ and a function $f: U \to \mathbb{R}^n$ such that $f(U) \subset \mathbb{R}^n$ is open, $f$ is a $C^\infty$ diffeomorphism onto its image and either

1) $f(U \cap M) = f(U) \cap \mathbb{R}^{n-1}$, or
2) $f(U \cap M) = f(U) \cap \{y \subset \mathbb{R}^{n-1}; y(1) \geq 0\}$.

Here $y(1)$ denotes the first component of the vector $y$. The usual definition of a submanifold allows only the first case. When both cases are allowed, $M$ is usually called a submanifold with boundary. We allow both cases because our consideration of decision regions confined to a compact domain implies that many interesting decision boundaries are not true submanifolds.

**Definition 2:** The piecewise-smooth decision regions in $X \subset \mathbb{R}^n$ are the sets in the collection

$$D(X) = \{D \subset X; \partial D \text{ is a finite union of } n-1 \text{-dimensional submanifolds of } \mathbb{R}^n\}$$

where $\partial D$ denotes the boundary of $D$.

Allowing $\partial D$ to be a union of submanifolds rather than a single submanifold means $D$ may have (well-behaved) sharp edges. For instance, if $X = [1,1]^n$ and the decision region is the halfspace $\{x \in X; a^T x \geq 0\}$, then the decision boundary consists of a union of up to $2n$ polygonal faces. Each of these faces is an $n-1$-dimensional submanifold (with boundary).

It is assumed that the approximating decision regions belong to a class $C^d$ of subsets of $X$ which gets progressively larger as $d$ increases. That is, $C^d \subset C^{d_2}$ if $d_1 < d_2$. Typically, $d$ is a nondecreasing function of the parameter space. The true decision region is $D$, then for any particular choice of $d$ the minimum approximation error is $\inf_{\Sigma \in C^d} V(D, \Sigma)$. Clearly, the minimum approximation error is a nonincreasing function of $d$. For some choices of $C^d$, the minimum approximation error goes to zero as $d \to \infty$. In such cases, the classes $C^d$ are said to be uniform approximators. The degree of approximation problem for uniform approximators $C^d$ involves determining how quickly the minimum approximation error decreases.

**The Degree of Approximation Problem:** Let $X \subset \mathbb{R}^n$ be compact and for each $d \geq 1$ let $C^d$ be a set of subsets of $X$ such that

$$\lim_{d \to \infty} \sup_{D \in D(X)} \inf_{\Sigma \in C^d} V(D, \Sigma) = 0.$$

Find the largest $R \geq 0$ such that, for all sufficiently large $d$

$$\sup_{D \in D(X)} \inf_{\Sigma \in C^d} V(D, \Sigma) \leq \frac{c}{d^R} \quad (2)$$

where $c$ is constant with respect to $d$. For a decision region $D \subset X$ and an approximate decision region $\Sigma \subset X$, we say that $\Sigma$ approximates $D$ well if $V(D, \Sigma)$ is small; thus most points in $X$ are correctly classified by $\Sigma$.
The constant $R$ in (2) is called the degree of approximation for the class $C^d$ of decision regions.

III. THE DELTA CORRIDOR

Let

$$B(x, \delta) := \{ z \in \mathbb{R}^n : \| x - z \| \leq \delta \}$$

the closed-ball with center $x$ and radius $\delta$, where $\| \cdot \|$ denotes the 2-norm (Euclidean distance) in $\mathbb{R}^n$.

**Definition 3:** The $\delta$ corridor around the decision boundary is the set

$$\partial D + \delta := \bigcup_{x \in \partial D} B(x, \delta).$$

For any $y \in \mathbb{R}^n$, $y \leq \delta$

$$\partial D + y = \{ x + y : x \in \partial D \} \subset \partial D + \delta.$$

The construction in Section IV of the approximating set $\Sigma$ bounds the misclassified region by the volume of a $\delta$ corridor around the decision boundary. Thus in order to answer the approximation problem, it is necessary to determine the volume of the decision boundary. This requires some knowledge of the size and smoothness of $\partial D$. For instance, if $\partial D$ is a space-filling curve, then the volume of any corridor around $\partial D$ will be equal to the volume of $X$, and knowledge of the size of the corridor offers no advantage. On the other hand, if $D$ is a ball with radius greater than the corridor size, then the volume is equal to twice the times the corridor size multiplied by the surface area of the ball. In order to obtain a general result for decision regions in $D(X)$, we use the following definition for the area of a hypersurface [23, 24].

**Definition 4:** Let $\partial D \subset D(X)$, and let the points $u \in \partial D$ be locally referred to parameters $u(1), \ldots, u(n-1)$, which are mapped to the Euclidean space $\mathbb{R}^{n-1}$ with the coordinates $u(1), \ldots, u(n-1)$. The surface area of $\partial D$ is defined as

$$\text{area}(\partial D) := \int_{\partial D} \det(R) \, du(1) \cdots du(n-1)$$

where $R = [R_{ij}], R_{ij} = \partial u(i)/\partial u(j)$. Thus area($\partial D$) is the volume of the image of $\partial D$ in $\mathbb{R}^{n-1}$.

If $n = 2$, then $\partial D$ is a curve in the plane, and area($\partial D$) is the length of $\partial D$. The size of area($\partial D$) increases rapidly with $n$, for instance, if $D = [-1, 1]^n$ then area($\partial D$) $\approx 2\pi^{n/2}n$. Using this definition, the volume of the corridor around a decision boundary can be bounded as follows:

**Lemma 5:** Let $X \subset \mathbb{R}^n$ be compact. For any $D \in D(X)$ there exists $\Delta = \Delta(D) > 0$ such that

$$\text{vol}(\partial D + \delta) \leq c\delta \text{ area}(\partial D)$$

for all $\delta$ such that $0 < \delta < \Delta$.

This result is intuitively obvious, since $\partial D$ can be locally approximated by an $n-1$-dimensional hyperplane, and the volume of the $\delta$ corridor around a piece of an $n-1$-dimensional hyperplane with area $\alpha$ is $2\alpha\delta + O(\delta^2)$. A rigorous proof of Lemma 5 can be given using a result by Weyl that appears in [23].

IV. POLYNOMIAL DECISION REGIONS

**Definition 6:** $P_d^m$ is the space of polynomials of degree at most $d$ in each of $n$ variables. That is, $P_d^m$ is the space of all linear combinations of $x(1)^{s_1}x(2)^{s_2}\cdots x(n)^{s_n}$ with $s_i \leq d, s_i \in \mathbb{N}_0$. The number of parameters necessary to identify elements in $P_d^m$ is $(d+1)^n$.

$C_P^m_d$ is the class of polynomial decision regions. Each decision region in $C_P^m_d$ is the positive domain of a polynomial in $P_d^m$. Specifically,

$$C_P^m_d := \{ \Sigma \subset X : \exists f \in P_d^m \text{ satisfying } f(x) \geq 0, \text{ if } x \in \Sigma \}$$

In this section and in Section V, $c \in \mathbb{R}$ denotes a quantity which is independent of $d$. Dependence of $c$ on other variables will be indicated by, for instance, $c = c(n)$. If no such indication is given, $c$ is an absolute constant. The exact value of $c$ will change without notice, even in a single expression.

The following Theorem is derived from Timan [25, result 5.3.2].

**Theorem 7 (Timan):** Let $X = [ -1, 1]^n$ and $g : X \rightarrow \mathbb{R}$. If $g$ is $L_1$-integrable on $X$ then there exists a constant $c$ such that

$$\inf_{f \in P_d^m} \int_X |g(x) - f(x)| \, dx \leq cn \sup_{\| g \| \leq 1} \frac{1}{d+1} \int_X |g(x) - g(x + y)| \, dx.$$ 

In the following, Theorem 7 is used to determine the degree of approximation of decision regions possessing a smooth boundary by polynomial decision regions.

**Theorem 8:** Let $X \subset \mathbb{R}^n$ be compact. If $D \in D(X)$ then there exists a constant $c$ such that

$$\inf_{\Sigma \subset C_P^m_d} V(D, \Sigma) \leq \frac{cn \text{ area}(\partial D)}{d+1}.$$

**Proof:** From the definition of $V(D, \Sigma)$

$$\inf_{\Sigma \subset C_P^m_d} V(D, \Sigma) = \inf_{f \in P_d^m} \int_X \| y_D(x) - \text{sgn}(g(x)) \| \, dx$$

$$= \inf_{f \in P_d^m} \int_{\{ |y_D(x) - g(x)| > 1 \} \cdot \| y_D(x) - g(x) \| \, dx \leq 2 \inf_{f \in P_d^m} \int_X |y_D(x) - g(x)| \, dx.$$ 

Now $y_D$ is $L_1$-integrable, so Theorem 3 applies with $g = y_D$. From the definition of the $\delta$-corridor

$$\sup_{\| g \| \leq 1} \int_X |y_D(x) - y_D(x + y)| \, dx \leq \text{vol}(\partial D + \delta).$$

Letting $\delta = 1/(d+1)$ and combining with Lemma 5 gives the result.

V. NEURAL NETWORK DECISION REGIONS

**Definition 9:** $N_d^n$ is the space of functions defined by single hidden layer feedforward neural networks with $n$ inputs, and
nodes in the hidden layer. That is, $\mathcal{N}_d^\alpha$ is the space of all linear combinations
\[
\phi_0 + \sum_{i=1}^d \alpha_i \phi_i(\beta_i^T x + \delta_i)
\]
where $x, \beta_i \in \mathbb{R}^n, \alpha_i, \delta_i \in \mathbb{R}$, and $\phi(x) = (1 + e^{-x})^{-1}$. In order to identify elements in $\mathcal{N}_d^\alpha$, one must specify $1 + d(n+2)$ real numbers.

$\mathcal{CN}_d^\alpha$ is the class of neural network decision regions. Each decision region in $\mathcal{CN}_d^\alpha$ is the positive domain of a function in $\mathcal{N}_d^\alpha$. Specifically
\[
\mathcal{CN}_d^\alpha := \{ \Sigma \subset X : \exists f \in \mathcal{N}_d^\alpha \text{ satisfying } f(x) \geq 0 \text{ if } x \in \Sigma, f(x) < 0 \text{ if } x \notin \Sigma \}.
\]

Theorem 10 is derived from [9, Corollary 5.2], since the degree of approximation by polynomials is the same as the degree of approximation by trigonometric polynomials in this case.

Theorem 10 (Mhaskar and Micchelli): Let $X \subset \mathbb{R}^n$ be compact and $g : X \rightarrow \mathbb{R}$. There exists a function $\mathcal{C} \mathcal{P}_d^\alpha \ni \mathcal{N}_d^\alpha \rightarrow \mathbb{N}$ such that if $g$ is $L_1$ integrable on $X$ then for any $p, q \in \mathbb{N}$, there exists a constant $c$ such that
\[
\inf_{f \in \mathcal{N}_d^\alpha} \int_X |g(x) - f(x)| \, dx 
\leq c \left( \inf_{p \geq q} \int_X |g(x) - f(x)| \, dx + p^n e^{-cq} \right) \int_X |g(x)| \, dx.
\]
Moreover, there exists $c > 0$ such that $d(p, q) \leq c p^{-n} q^2$ for sufficiently large $p$ and $q$.

The degree of approximation by neural net functions is related to the degree of approximation by polynomial functions. The upper bound in Theorem 10 is a monotonic decreasing function if both $p$ and $q$ are increasing.

Next we use the technique in Section IV to obtain a degree of approximation result for neural network decision regions. In order to get a fair comparison with the polynomial decision regions, we consider degree of approximation by $\mathcal{C} \mathcal{P}_d^\alpha$, since the number of parameters necessary to specify elements in either $\mathcal{C} \mathcal{P}_d^\alpha$ or $\mathcal{N}_d^\alpha$ is approximately $c(n)d^p$.

Theorem 11: Let $X \subset \mathbb{R}^n$ be compact. If $D \in \mathcal{D}(X)$ then for any $r \in (0, 1)$ there exist constants $c, c(n, r)$ such that
\[
\inf_{\Sigma \in \mathcal{CN}_d^\alpha} V(D, \Sigma) < \frac{c \text{ area}(\partial D)}{d^r}
\]
for all $d \geq c(n, r)$.

Proof: Assume $d^p = q^n q^2 \geq d(p, q)$, From Theorem 10, the minimum misclassified region satisfies
\[
\inf_{\Sigma \in \mathcal{CN}_d^\alpha} V(\partial D) \leq \inf_{\Sigma \in \mathcal{CN}_d^\alpha} \left( \frac{n \text{ area}(\partial D)}{p+1} + p^n e^{-cq} \right).
\]
Now choose $p$ and $q$ such that the upper bound in (3) decreases to zero as quickly as possible as $p$ and $q$ increase. Let $q = p^n$, so that $d = p^{1+\alpha}$. For any $\alpha > 0$, $p^n e^{-cq} \leq p^{-1}$ is decreasing for sufficiently large $p$. How large $p$ must be depends on $\alpha$. Now for any $r \in (0, 1)$ there exists $\alpha > 0$ such that $r = (1 + \alpha)^{-1}$, so there exists a choice of $p, q$ such that
\[
\frac{n \text{ area}(\partial D)}{p+1} + p^n e^{-cq} \leq \frac{cn \text{ area}(\partial D)}{d^r},
\]
The result follows.

VI. CONCLUDING REMARKS

We have given degree of approximation results for implicit decision region approximation which are similar to Jackson’s Theorem for polynomial function approximation. The approximating decision regions are defined by the positive domains of polynomial functions or feedforward neural networks. These results support our intuition that classes of functions which are good function approximators tend to be good implicit decision region approximators.

Many open problems remain—the most pressing being “What conditions give better degree of approximation?” In function approximation, higher order smoothness of the approximated function gives a better degree of approximation. For instance, in Theorem 7 if the $p$th derivative is Lipschitz-continuous, then the degree of approximation is at least $p+1$. We would expect that there exist restrictions on the decision region to be approximated, $D$, which will guarantee a better degree of approximation than our results suggest. Moreover, we would expect that there would be a series of successively tighter restrictions on $D$ which would guarantee successively better degree of approximation results.

However, it is not clear what the right conditions are. Bounding the curvature of the boundary of $D$ will not affect the degree of approximation using our argument, since all information about the decision boundary other than its area affects only higher order terms in the approximation bound, not the degree of approximation obtained in Theorem 8. Perhaps the number of connected components in $D$ is the condition we need. Or perhaps the curvature properties of the decision boundary are important, but a tighter method of bounding $V(D, \Sigma)$ than the volume of the corridor size is needed. Maybe a completely different proof technique is needed to get higher degree of approximation results.

ACKNOWLEDGMENT

The authors wish to thank the reviewers and associate editor of this journal for their comments, and P. Bartlett for helpful discussions motivating this research. In particular, H. Mhaskar pointed out in reviews that the use of $L_1$-function
approximation results would give a much simpler derivation and a tighter bound than the $L_\infty$ technique we originally proposed.

REFERENCES


