Decision Region Approximation by Polynomials or Neural Networks

Kim L. Blackmore * Robert C. Williamson°
Iven M.Y. Mareels°

Abstract

We give degree of approximation results for decision regions which are defined by polynomial parametrizations and by neural network parametrizations. The volume of the misclassified region is used to measure the approximation error. We use results from the degree of approximation of functions by first constructing an intermediate function which correctly classifies almost all points. For both classes of approximating decision regions, we show that the degree of approximation is at least r, where r can be any number in the open interval $(0, 1)$.

Keywords: Rate of approximation; Decision region; Classification; Polynomials; Neural Networks.

1 Introduction

Decision regions arise in the machine learning problem of sorting or classification of data [24]. 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 fuction

$$y_D(x) = \begin{cases} 1 & \text{if } x \in D \\ -1 & \text{otherwise} \end{cases} \tag{1.1}$$

*Comms Division, TSAS, DSTO, PO Box 1500, Salisbury SA 5108, Australia. Formerly at:
° Department of Engineering, Australian National University, Canberra ACT 0200, Australia

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 [4]. 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 analysing 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 [19] (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 [11]. Other theoretical questions involve determining whether or not best approximations exist and are unique, for approximating functions from particular classes [6, 26]. There are also degree of approximation results, which 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 [6] is the first example of this. Hornik [12], Barron [1], Mhaskar and Michelli [21], Mhaskar [20], Darken et al. [7] and Hornik et al. [13] give degree of approximation results for neural networks. Even more powerful results give the best class of functions to use in approximating particular classes of functions, by showing converses to Jackson type theorems for certain classes of functions [22].

Given that we are representing a decision region as the positive domain of a function, function approximation results do not immediately translate into the decision region context. In order to approximate a given decision region, the first task is

to identify a function g with suitable smoothness properties for which the positive domain closely matches the decision region. Function approximation then establishes the existence of a good approximation f of g, where f belongs to the desired class of functions. "Best" may be measured in terms of any function norm, such as the infinity norm or the 2-norm. However, such measures of good function approximation do not guarantee that the positive domain of the approximate function is close to the original decision region. For instance, there may be arbitrarily many points x where $g(x)$ is small even though x is far from a zero of g. At these points good approximation of g may not ensure that the sign of $g(x)$ equals the sign of the best approximation $f(x)$, so the positive domains of g and f may not match at arbitrarily many points. Restrictions on the function g may guarantee that good function approximation will imply good decision region approximation. However, it is not clear how restrictions on the function describing a decision region translate to restrictions on the decision region itself, which is all that is given in the original problem formulation — the function g is introduced solely for the sake of solving the problem.

The problem of approximating sets, rather than functions, has received some attention in the literature. Approximation of (unparametrised) sets and curves has been studied for pattern recognition and computer vision purposes [2, 15, 27]. The approach is quite different to the approach here. Theoretical work can be grouped according to two basic approaches—namely explicit and implicit parametrisations. "Explicit parametrisation" refers to frameworks where the decision *boundary* is parametrised. 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 parametrisation" refers to frameworks (as used in this thesis) where the decision *region* is the positive domain of some function.

Most existing work is in terms of explicit parametrisations [16]. For instance, Korostelev and Tsybakov [17, 18] consider the *estimation* (from sample data) of decision regions. Although they consider non-parametric 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 [8] and Shchebrina [23] have determined the metric entropy of certain smooth curves.

Regarding the implicit problem, Mhaskar [20] gives a universal approximation type result for approximation by positive domains of certain neural network functions. It appears that the argument in [20] can not be used to determine the degree of

approximation. Ivanov [14] summarises many problems in algebraic geometry concerned 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 [14], Broglia and Tognoli [5] consider when a C^∞ 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. This is achieved by constructing a continuous function from the discriminant function for the decision region, using a convolution process. The continuous function can be approximated by polynomials, to a certain degree of accuracy, and this gives a bound on the distance that the boundary of the approximate decision region can be from the true decision boundary. We then use a result from differential geometry to link this distance with the size of the misclassified volume. Since most learning problems can be analysed 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, and definitions of the different measures of approximation error we use in the paper, along with results showing how the measures can be connected. Section 3 contains the construction of a smooth function g whose positive domain closely matches any (reasonable) decision region, by convolution of the discriminant function for the decision region with a square convolution kernel. Section 4 contains the polynomial approximation results. Our main result is Theorem 4.2, 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 m, goes to zero at least as fast as m^{-r}, where r can be made as close to (but less than) 1 as desired. By "smooth boundary" we mean essentially that the boundary is a finite union of $n-1$ dimensional manifolds. In Section 5 a similar result is given for decision regions defined by neural networks and the two results are compared. Section 6 concludes.

2 Measuring Approximation Error

2.1 The Approximation Problem

We assume that a decision region is a closed subset D of a compact set $X \subset \mathbb{R}^n$, called the sample space. Points in the sample space are classified positively if they are contained in the decision region, and negatively if they are not. We wish to determine how well a decision region can be approximated by the positive domain of functions belonging to a parametrized class of functions, in the sense of minimizing the probability of misclassification. If points to be classified are chosen uniformly throughout the sample space X, the probability of misclassification is equal to the *volume of the misclassified region*, i.e. the volume of the symmetric difference of the two sets. For decision regions $D_1, D_2 \subset X$, the volume of the misclassified region is

$$V(D_1, D_2) := \mathrm{vol}(D_1 \triangle D_2) = \int_{D_1 \triangle D_2} dx.$$

For a decision region $D \subset X$ and an approximate decision region $\Sigma \subset X$, we say that Σ approximates D well if $V(D, \Sigma)$ is small; thus most points in X are correctly classified by Σ.

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

The Degree of Approximation Problem *Let $X \subset \mathbb{R}^n$ be compact, let $\mathcal{D}$ be a set of subsets of X and for each $d > 0$, let $\mathcal{C}^d$ be a set of subsets of X, such that*

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

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

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

where c is constant with respect to d.

The constant R in (2.1) is called the *degree of approximation* for the class $\mathcal{C}^d$ of decision regions. Typically one is interested in solving the degree of approximation question for a class $\mathcal{D}$ of subsets of D. Our results are for $\mathcal{D} := \{D : \partial D$ is a finite union of smooth manifolds$\}$. In Sections 4 and 5 it is shown that the degree of approximation for $\mathcal{D}$ by polynomial and neural network decision regions is bounded below by any $r \in (0, 1)$.

2.2 Corridor Size

Let $B(x, \delta) := \{z \in \mathbb{R}^n : \|x - z\| \leq \delta\}$, the closed ball with centre x and radius δ, where $\|\cdot\|$ denotes the 2 norm (Euclidean distance) in $\mathbb{R}^n$. For any set $D \subset \mathbb{R}^n$, ∂D denotes the boundary of D.

Definition 2.1 *The* δ corridor *around any set $D \subset \mathbb{R}^n$ is the set*

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

The corridor size from D to Σ *is defined to be*

$$\rho(D, \Sigma) := \inf\{\delta : \partial\Sigma \subset \partial D + \delta\}$$

The corridor size is the smallest value of δ such that all points in the boundary of Σ are not further than δ from the boundary of D. The corridor size is not a metric because it is not symmetric; however it is worth noting that $\max\{\rho(D, \Sigma), \rho(\Sigma, D)\}$ is the Hausdorff distance between ∂D and $\partial\Sigma$ (which is a metric).

2.3 Relating V to ρ

The construction in Section 4 of the approximating set Σ gives an upper bound on the minimum corridor size from D to Σ rather than on the minimum volume of the misclassified region. So in order to answer the approximation problem, it is necessary to relate the corridor size from D to Σ to the volume of the misclassified region between D and Σ. The following relationship follows immediately from the definitions of V and ρ.

Lemma 2.2 *Let $\Sigma, D \subset \mathbb{R}^n$. If $\rho(D, \Sigma) = \delta$ and $\Sigma \subset D + \delta$, then*

$$V(D, \Sigma) \leq \mathrm{vol}(\partial D + \delta).$$

The requirement that $\Sigma \subset D + \delta$ is necessary because the corridor size can be small if either most points are correctly classified, or most points are misclassified. The function classes considered in Sections 4 and 5 contain complements of all of their members, so this is not a restrictive assumption.

Lemma 2.2 shows it is possible to bound the volume of the misclassified region by bounding the volume of the δ corridor around ∂D. This requires some knowledge of the size and smoothness of ∂D. For instance, if ∂D is a space filling curve, then the volume of *any* corridor around ∂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 two times the corridor size multiplied by the surface area of the ball. In order to obtain a general result, we assume that the decision boundary is finite union of hypersurfaces — $(n-1)$ dimensional submanifolds of $\mathbb{R}^n$, and measure the size of the hypersurface using the surface area [25, 3].

Definition 2.3 *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 neighbourhood $U \subset \mathbb{R}^n$ 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^∞ 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. Moreover, allowing ∂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 $\{x \in X : a^\top x \geq 0\}$, 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).

Definition 2.4 *Let M be a union of finitely many $n-1$ dimensional submanifolds of $\mathbb{R}^n$. Let the points $u \in M$ 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 $v(1), \ldots, v(n-1)$. The* surface area *of M is defined as*

$$\text{area}(M) := \int_M \det(R) du(1) \ldots du(n-1),$$

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

If $n = 2$, then M is a curve in the plane, and area(M) is the length of M.

Using these definitions the volume of the corridor around a decision region can be bounded as follows:

Lemma 2.5 *Let $D \subset X \subset \mathbb{R}^n$, X compact. If ∂D is a union of finitely many $n-1$ dimensional submanifolds of $\mathbb{R}^n$ then there exists $\Delta = \Delta(D) > 0$ such that*

$$\text{vol}(\partial D + \delta) \leq 4\delta \, \text{area}(\partial D)$$

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

This result is intuitively obvious, since ∂D can be locally approximated by an $n-1$ dimensional hyper-plane, and the volume of the δ corridor around a piece of an $n-1$ dimensional hyper-plane with area a is $2\delta a + O(\delta^2)$. A rigorous proof of Lemma 2.5 can be given using a result by Weyl that appears in [25].

If the decision boundary is a union of $n-1$ dimensional submanifolds of $\mathbb{R}^n$, Lemmas 2.2 and 2.5 can be employed to translate an upper bound on the minimum corridor distance into an upper bound on the minimum misclassified volume. Using this method, the surface area of the decision boundary does not affect the degree of approximation of decision regions, but only the constant in the approximation bound. The smoothness properties of the decision boundary, such as the curvature, do not even affect the constant in the approximation bound, according to the result in Weyl [25], they determine constants multiplying higher order terms in δ. This is in contrast with the function approximation results, where higher order smoothness of the original function does give higher degree of approximation (see Theorem 4.1). It appears unknown whether such a relationship exists for approximation of decision regions by positive domains of parametrised functions.

3 Construction of a Smooth Discriminant

In the following lemma we construct a Lipschitz continuous approximation to the discriminant function by convolution of the discriminant function with a function of compact support. This new function satisfies $\operatorname{sgn} g(x) = y_D(x)$ for all x sufficiently far from the decision boundary. In Sections 4 and 5 we use this constructed function to apply a bound on the rate of function approximation to our problem of bounding the rate of decision region approximation.

In the following, the i-th component of any vector $x \in \mathbb{R}^n$ is denoted $x(i)$, and $I(x,s) := \{z \in \mathbb{R}^n : |z(i) - x(i)| < \frac{s}{2}, i = 1, \ldots, n\}$ is the open n-cube with centre x and side s.

Lemma 3.1 *Let $D \subset \mathbb{R}^n$, and $0 < \delta < \frac{1}{\sqrt{n}}$. Define functions $h, g : \mathbb{R}^n \to \mathbb{R}$ as follows:*

$$h(x) := \frac{y_{I(0,s)} + 1}{2s^n} \tag{3.1}$$

where $s = \frac{2\delta}{\sqrt{n}}$, and

$$g(x) := (h * y_D)(x) = \int_{\mathbb{R}^n} h(x-t) y_D(t) dt. \tag{3.2}$$

Then

1. *For all $x \notin \partial D + \delta$, $g(x) = y_D(x)$.*
2. *g is Lipschitz continuous, with Lipschitz constant $\frac{n^{\frac{3}{2}}}{\delta}$.*

From equations 3.1 and 3.2 it can be seen that

$$\begin{aligned} g(x) &= s^{-n} \int_{I(x,s)} y_D(t) dt \\ &= \frac{\operatorname{vol}(I(x,s) \cap D) - \operatorname{vol}(I(x,s) \setminus D)}{\operatorname{vol}(I(x,s))}, \end{aligned} \tag{3.3}$$

Therefore $g(x) \in [-1, 1]$.

Proof

1. Let $x \in D \setminus (\partial D + \delta)$. Since the greatest distance from x to any point in $I(x,s)$ is

δ, $I(x,s) \subset D$. Thus the second volume in the numerator of (3.3) is zero. Therefore $g(x) = 1 = y_D(x)$. A similar argument gives the result if $x \notin D$ and $x \notin (\partial D + \delta)$.

2. Continuity of g follows from the definition of g as a convolution of two bounded, piecewise constant, functions. For Lipschitz continuity, we need to show that

$$|g(x_1) - g(x_2)| \leq \frac{n^{\frac{3}{2}}}{\delta}\|x_1 - x_2\| \tag{3.4}$$

for any $x_1, x_2 \in \mathbb{R}^n$. First, note that for any $x_1, x_2 \in \mathbb{R}^n$, $|g(x_1) - g(x_2)| \leq 2 \leq 2n = \frac{n^{\frac{3}{2}}}{\delta}s$. So if $\|x_1 - x_2\| \geq s$ then (3.4) holds.

Now assume that $\|x_1 - x_2\| < s$. Then $I_1 \cap I_2 \neq \emptyset$, where $I_1 = I(x_1, s)$ and $I_2 = I(x_2, s)$. From (3.3),

$$\begin{aligned}
|g(x_1) - g(x_2)| &= s^{-n}\Big|\operatorname{vol}(I_1 \cap D) - \operatorname{vol}(I_1 \backslash D) - \operatorname{vol}(I_2 \cap D) + \operatorname{vol}(I_2 \backslash D)\Big| \\
&= s^{-n}\Big|\operatorname{vol}(I_1 \backslash I_2 \cap D) - \operatorname{vol}((I_1 \backslash I_2)\backslash D) \\
&\qquad - \operatorname{vol}(I_2 \backslash I_1 \cap D) + \operatorname{vol}((I_2 \backslash I_1)\backslash D)\Big| \\
&\leq s^{-n}(\operatorname{vol}(I_1 \backslash I_2) + \operatorname{vol}(I_2 \backslash I_1))
\end{aligned}$$

which is the volume of the symmetric difference between I_1 and I_2, divided by the volume of the n-cubes I_1 and I_2. That is,

$$|g(x_1) - g(x_2)| \leq 2 - 2s^{-n}\operatorname{vol}(I_1 \cap I_2),$$

The intersection $I_1 \cap I_2$ is a rectangular region in $\mathbb{R}^n$, with side of length $s - |x_1(i) - x_2(i)|$ in the direction of the i-th axis. Thus

$$\begin{aligned}
|g(x_1) - g(x_2)| &\leq 2 - 2s^{-n}\prod_{i=1}^{n}(s - |x_1(i) - x_2(i)|) \\
&\leq 2\left\{1 - \left(1 - \frac{\|x_1 - x_2\|_\infty}{s}\right)^n\right\},
\end{aligned}$$

where $\|x\|_\infty = \max_{i=1,\ldots,n}|x(i)|$.

Writing $z = 1 - \frac{\|x_1 - x_2\|_\infty}{s}$, we use the fact that

$$1 - z^n \leq n(1 - z)$$

whenever $0 < z < 1$ and $n > 1$ (see Theorem 42 of [10]), to give

$$\begin{aligned}
|g(x_1) - g(x_2)| &\leq \frac{2n}{s}\|x_1 - x_2\|_\infty \\
&= \frac{n^{\frac{3}{2}}}{\delta}\|x_1 - x_2\|_\infty \\
&\leq \frac{n^{\frac{3}{2}}}{\delta}\|x_1 - x_2\|.
\end{aligned}$$

Thus (3.4) holds for all $x_1, x_2 \in \mathbb{R}^n$. ∎

The function $g : \mathbb{R}^n \to \mathbb{R}$ is continuous and $\operatorname{sgn} g$ correctly classifes all points in $\mathbb{R}^n$, except possibly points in $\partial D + \delta$. If h is replaced with a smoother convolution kernel, the resulting function g will be smoother. For instance, if h has p continuous derivatives, and the p-th derivative is bounded, then g will have p continuous derivatives, and the p-th derviative will be bounded [28]. At the end of Section 4 we show that this apparent improvement does not actually affect the order of the approximation result achievable by our argument.

4 Polynomial Decision Regions

First, some notation from polynomial function approximation:

1. For any function $f : \mathbb{R}^n \to \mathbb{R}$ and any vector $\alpha \in \mathbb{N}_0^n$, we say $D^\alpha f = \frac{\partial f}{\partial x(1)^{\alpha(1)} \partial x(2)^{\alpha(2)} \cdots \partial x(n)^{\alpha(n)}}$, where $\sum_{i=1}^n \alpha(i) = p,\ \alpha(i) \geq 0$ is a p-th order derivative of f.

2. $\mathcal{P}_d^n$ is the space of polynomials of degree at most d in $x(1), \ldots, x(n)$. That is, $\mathcal{P}_d^n$ is the space of all linear combinations of $x(1)^{s_1} x(2)^{s_2} \cdots x(n)^{s_n}$ with $\sum_{i=1}^n s_i \leq d,\ s_i \in \mathbb{N}_0$. The number of parameters necessary to identify elements in $\mathcal{P}_d^n$ is the number of ways of choosing n nonnegative integers s_i so that $\sum_{i=1}^n s_i \leq d$, since the parameters are the coefficients in the linear combination. This number is less than $(d+1)^n$.

3. $\mathcal{CP}_d^n$ is the class of *polynomial decision regions*. Each decision region in $\mathcal{CP}_d^n$ is the positive domain of a polynomial in $\mathcal{P}_d^n$. Specifically,

$$\mathcal{CP}_d^n := \left\{ \Sigma \subset X : \exists f \in \mathcal{P}_d^n \text{ satisfying } \begin{array}{lll} f(x) \geq 0 & \text{if} & x \in \Sigma \\ f(x) < 0 & \text{if} & x \notin \Sigma \end{array} \right\}.$$

In this section and in Section 5, $c \in \mathbb{R}$ 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 result is an n-dimensional generalisation of Jackson's Theorem. It can be derived from Theorem 9.10 of Feinerman and Newman [9]. The derivation closely mimics the derivation of Theorem 4.5 in [9] from Theorem 4.2 in [9].

Theorem 4.1 *Let $X = [-1,1]^n$ and $g : X \to \mathbb{R}$. If $D^\alpha g$ is Lipschitz continuous with Lipschitz constant L for all $\alpha \in \mathbb{N}_0^n$ such that $\sum_{i=1}^n \alpha(i) = p$, then there exists $c(p) > 1$ such that*

$$\inf_{f \in \mathcal{P}_d^n} \sup_{x \in X} |g(x) - f(x)| \leq c(p)\ L \frac{n^{\frac{3}{2}(p+1)}}{(d+n)^{p+1}},$$

if $d + n \geq p + 1$.

In the following Theorem 4.1 is used to determine the degree of approximation of decision regions possessing a smooth boundary by polynomial decision regions. First it is shown that the minimum corridor distance between the true decision region and the approximating decision regions goes to zero at least as fast as d^{-r}, where $0 < r < 1$. This bound is then used in order to obtain a bound on the misclassified volume.

Theorem 4.2 *Let $D \subset X = [-1,1]^n$. If ∂D is a union of finitely many $n-1$ dimensional submanifolds of $\mathbb{R}^n$ then for any $r \in (0,1)$ there exist constants c, $M(r, D, n) > 1$ such that*

$$\inf_{\Sigma \in \mathcal{CP}_d^n} V(D, \Sigma) < \frac{c\ \text{area}(\partial D)}{d^r}$$

for all $d \geq c(r, D, n)$.

Proof Choose $d \geq 1$ and $r \in (0,1)$. Define $\delta_d = \frac{1}{(d+n)^r}$.

Define g_d as in (3.2), using $\delta = \delta_d$. Then g_d is C^0, with Lipschitz constant $\frac{n^{\frac{3}{2}}}{\delta_d}$. According to Theorem 4.1, there exists $f_d \in \mathcal{P}_d^n$ satisfying

$$\begin{aligned} \Omega_d := \sup_{x \in X} |g_d(x) - f_d(x)| &\leq c \frac{n^{\frac{3}{2}}}{\delta_d} \frac{n^{\frac{3}{2}}}{d+n} \\ &\leq \frac{cn^3}{(d+n)^{1-r}}. \end{aligned}$$

If $d \geq d_r = (cn^3)^{\frac{1}{1-r}} - n$ then $\Omega_d < 1$.

By Lemma 3.1, $|g_d(x)| = 1$ for all $x \notin \partial D + \delta_d$. So if $d > d_r$, for all points outside of the δ_d corridor of ∂D, $f_d(x) = g_d(x) - (g_d(x) - f_d(x))$ has the same sign as $g_d(x)$. That is, points outside of the δ_d corridor of ∂D are correctly classified. The corridor size from D to the positive domain of f_d is thus bounded above by δ_d. Set $c = 4$ and $c(r, D, n) = \max\{d_r, \Delta^{-1} - n\}$, where Δ is defined in Lemma 2.5. The result follows from Lemmas 2.2 and 2.5. ∎

The requirement that $r < 1$ in Theorem 4.2 is essential since the constant d_r increases exponentially with $\frac{1}{1-r}$. Obviously larger r gives a stronger order of approximation result in the limit $d \to \infty$, but $c(r, D, n)$, the lower bound on d for which the result holds, will be larger.

It might seem that Theorem 4.2 could be improved upon by making the intermediate function g smoother, since this gives great improvement in the result of Theorem 4.1. However, this improvement cannot be achieved using the obvious generalisation of our convolution technique, as we now show.

Assume $g = h * y_D$ where h is any nonnegative function with support $I(0, 2\delta n^{-1/2})$ which is differentiable to $(p+1)$-th order such that $|D^\alpha h| \leq B$ for all $\alpha \in \mathbb{N}_0^n$ such that $\sum_{i=1}^n \alpha(i) = p + 1$. Then g is differentiable to $(p+1)$-th order and $|D^\alpha g| \leq c(n)\delta^n B$ for all $\alpha \in \mathbb{N}_0^n$ such that $\sum_{i=1}^n \alpha(i) = p + 1$. This bound is found by using the fact that $\sum_{i=1}^n \alpha(i) = p + 1$. $D^\alpha g = y * D^\alpha h$ (Theorem 18.4 in [28]), and that convolution involves integration over an n-cube of side $2\delta n^{-1/2}$. On the other hand the fact that h has compact support and is differentiable to $p + 1$-th order implies that $|h(x)| \leq c(n, p)\delta^{p+1} M$, so for all $x \notin \partial D + \delta$, $|g(x)| \leq c(n, p)\delta^{p+1+n} M$.

Now applying Theorem 4.1 gives

$$\inf_{g \in \mathcal{P}_d^n} \sup_{x \in X} |g(x) - f(x)| \leq c(p, n)\delta^n B \frac{n^{\frac{3}{2}(p+1)}}{(d+n)^{p+1}}.$$

Following the proof of Theorem 4.2, a value of δ is sought which will guarantee that the function approximation error is less than the absolute value of g for all $x \notin \partial D + \delta$. This involves finding δ_d such that

$$c(n, p)\, \frac{\delta_d^n B}{(d+n)^{p+1}} < c(n, p)\delta_d^{p+1+n} M.$$

This only holds for all large d if $\delta_d > \frac{c(n,p)}{d+n}$. But the decision region approximation error is a constant times δ_d, so if this method is used the lower bound on the degree of approximation must be less than 1.

Thus it has been demonstrated that if g is the convolution of the discriminant function with any function h of compact support, then an argument analogous to that of Theorem 4.2 cannot give a bound with better order of magnitude in d. (Of course this does not preclude better bounds existing.) Thus the degree of approximation result obtainable by the method of Theorem 4.2 does not depend on the smoothness of the intermediate function g, which is constructed solely for the sake of the proof.

5 Neural Network Decision Regions

Using the techniques of the last section, a result similar to Theorem 4.2 which applies when the approximating decision regions are defined by neural networks will be derived. In order to do this, function approximation result similar to Theorem 4.1, for functions defined by neural networks is used. First some notation:

1. A *sigmoidal function* is any continuous function $\sigma : \mathbb{R} \to \mathbb{R}$ which satisfies

$$\lim_{x \to -\infty} \sigma(x) = 0, \qquad \lim_{x \to \infty} \sigma(x) = 1.$$

2. $\mathcal{N}_d^n$ is the space of functions defined by two hidden layer feedforward neural networks with n inputs, $n(d^n + 1)$ nodes in the first layer, and $d^n + 1$ nodes in the second layer. That is, $\mathcal{N}_d^n$ is the space of all linear combinations

$$\sum_{j=0}^{d^n} \alpha_j \sigma \left(\sum_{k=1}^{n} \beta_{jk} \sigma \left(\gamma_{jk}^\top x + \delta_{jk} \right) + \varepsilon_j \right)$$

   where $x, \gamma_{jk} \in \mathbb{R}^n$, and $\alpha_j, \beta_{jk}, \delta_{jk}, \varepsilon_j \in \mathbb{R}$. In order to identify elements in $\mathcal{N}_d^n$, one must specify $(n^2 + 2n + 2)(d^n + 1)$ real numbers.

3. $\mathcal{CN}_d^n$ is the class of *neural network decision regions*. Each decision region in $\mathcal{CN}_d^n$ is the positive domain of a function in $\mathcal{N}_d^n$. Specifically,

$$\mathcal{CN}_d^n := \left\{ \Sigma \subset X : \exists f \in \mathcal{N}_d^n \text{ satisfying } \begin{array}{lll} f(x) \geq 0 & \text{if} & x \in \Sigma \\ f(x) < 0 & \text{if} & x \notin \Sigma \end{array} \right\}.$$

Theorem 5.1 is a special case of Theorem 3.4 of Mhaskar [20]. (In [20] the function is defined on $[0, 1]^n$ rather than $[-1, 1]^n$.)

Theorem 5.1 Mhaskar *Let $X = [-1,1]^n$ and $g : X \to \mathbb{R}$. If g is Lipschitz continuous with Lipschitz constant L then there exists a constant $c(\sigma, n)$ such that*

$$\inf_{f \in \mathcal{N}_d^n} \sup_{x \in X} |g(x) - f(x)| \leq \frac{c(\sigma, n)L}{d}$$

for all $d \geq 1$.

This result is very similar to Theorem 4.1 in the case $p = 0$. Using the technique in Section 4, the following theorem holds:

Theorem 5.2 *Let $D \subset X = [-1,1]^n$. If ∂D is a union of finitely many $n-1$ dimensional submanifolds of $\mathbb{R}^n$ then for any $r \in (0,1)$ there exist constants c, $c(r, D, n) > 1$ such that*

$$\inf_{\Sigma \in \mathcal{CN}_d^n} V(D, \Sigma) < \frac{c \ \text{area}(\partial D)}{d^r}$$

for all $d \geq c(r, D, n)$.

Comparing Theorem 5.2 with Theorem 4.2, it can be seen the two classes of approximating decision regions give the same upper bound on the degree of approximation for the two class $\mathcal{CP}_d^n$ and $\mathcal{CN}_d^n$. Moreover, the number of parameters necessary to specify elements in the approximating classes is of order d^n in both cases. Thus there is no apparent advantage of one class over the other. However lower bounds on the degree of approximation are needed in order to conclude that polynomial decision regions and neural network decision regions are equally capable approximators of decision regions.

6 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 4.1 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. As discussed in Section 4, using a smoother convolution kernel h to construct a smoother intermediate function f will not give a better degree of approximation using our methods. It is also clear that 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 4.2.

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.

7 Acknowledgements

The authors would like to thank Peter Bartlett for helpful discussions motivating this research, and the reviewers and associate editor of this journal for their comments.

This work was supported by the Australian Research Council.

References

[1] A.R. Barron. Universal approximation bounds for superposition of a sigmoidal function. *IEEE Transactions on Information Theory*, 39:930–945, 1993.

[2] A. Bengtsson and J.-O. Eklundh. Shape representation by multiscale contour approximation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:85–93, 1991.

[3] M. Berger and B. Gostiaux. *Differential Geometry: Manifolds, Curves, and Surfaces.* Springer-Verlag, New York, 1988.

[4] K.L. Blackmore, R.C. Williamson, and I.M.Y. Mareels. Learning nonlinearly parametrized decision regions. *Journal of Mathematical Systems, Estimation, and Control*, 1995. To appear.

[5] F. Broglia and A. Tognoli. Approximation of C^∞ functions without changing their zero-set. *Ann. Inst. Fourier, Grenoble*, 39(3):611–632, 1989.

[6] E.W. Cheney. *Introduction to Approximation Theory.* Chelsea, New York, 2nd edition, 1982.

[7] C. Darken, M. Donahue, L. Gurvits, and E. Sontag. Rate of approximation results motivated by robust neural network learning. In *Proceedings of the Sixth ACM Conference on Computational Learning Theory*, pages 103–109, 1993.

[8] R.M. Dudley. Metric entropy of some classes of sets with differentiable boundaries. *Journal of Approximation Theory*, 10:227–236, 1974.

[9] R.P. Feinerman and D.J. Newman. *Polynomial Approximation.* The Williams and Wilkins Company, Baltimore, 1974.

[10] G.H. Hardy, J.E. Littlewood, and G. Pólya. *Inequalities.* Cambridge University Press, Cambridge, 1988.

[11] K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

[12] K. Hornik, M. Stinchcombe, and H. White. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward neural networks. *Neural Networks*, 3:551–560, 1990.

[13] K. Hornik, M. Stinchcombe, H. White, and P. Auer. Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. Technical Report NC-TR-95-004, NeuroCOLT Technical Report Series, January 1995.

[14] N.V. Ivanov. Approximation of smooth manifolds by real algebraic sets. *Russian Math. Surveys*, 37(1):1–59, 1982.

[15] P.S. Kenderov and N.K. Kirov. A dynamical systems approach to the polygonal approximation of plane convex compacts. *Journal of Approximation Theory*, 74:1–15, 1993.

[16] N.P. Korneichuk. Approximation and optimal coding of plane curves. *Ukranian Mathematical Journal*, 41(4):429–435, 1989.

[17] A.P. Korostelev and A.B. Tsybakov. Estimation of the density support and its functionals. *Problems of Information Transmission*, 29(1):1–15, 1993.

[18] A.P. Korostelev and A.B. Tsybakov. *Minimax Theory of Image Reconstruction*. Lecture Notes in Statistics v. 82. Springer, New York, 1993.

[19] G.G. Lorentz. *Approximation of Functions*. Holt, Rinehart and Winston, New York, 1966.

[20] H.N. Mhaskar. Approximation properties of a multilayered feedforward artificial neural network. *Advances in Computational Mathematics*, 1:61–80, 1993.

[21] H.N. Mhaskar and C.A. Micchelli. Approximation by superposition of sigmoid and radial basis functions. *Advances in Applied Mathematics*, 13:350–373, 1992.

[22] P.P. Petrushev and V.A. Popov. *Rational Approximation of Real Functions*. Cambridge University Press, Cambridge, 1987.

[23] N.V. Shchebrina. Entropy of the space of twice smooth curves in $\mathbb{R}^{n+1}$. *Math. Acad. Sci. USSR*, 47:515–521, 1990.

[24] J. Sklansky and G.N. Wassel. *Pattern Classifiers and Trainable Machines*. Springer-Verlag, New York, 1981.

[25] H. Weyl. On the volume of tubes. *American Journal of Mathematics*, 61:461–472, 1939.

[26] R.C. Williamson and U. Helmke. Existence and uniqueness results for neural network approximations. *IEEE Transactions on Neural Networks*, 6(1):2–13, 1995.

[27] J.-S. Wu and J.-J. Leou. New polygonal approximation schemes for object shape representation. *Pattern Recognition*, 26:471–484, 1993.

[28] A.C. Zaanen. *Continuity, Integration and Fourier Theory*. Springer-Verlag, Berlin, 1989.