Journal of Mathematical Systems, Estimation, and Control
Vol. 6, No. 2, 1996, pp. 1–18

Local Minima and Attractors at Infinity for Gradient Descent Learning Algorithms*

Kim L. Blackmore† Robert C. Williamson†
Iven M.Y. Mareels†

Abstract

In the paper "Learning Nonlinearly Parametrized Decision Regions", an online scheme for learning a very general class of decision regions is given, together with conditions on both the parametrization and on the sequence of input examples under which good learning can be guaranteed to occur. In this paper, we discuss these conditions, in particular the requirement that there be no non-global local minima of the relevant error function, and the more specific problem of no attractor at infinity. Somewhat simpler sufficient conditions are given. A number of examples are discussed.

Key words: error function, local minimum, attractor at infinity, learning, decision region

AMS Subject Classifications: 68T05

1 Introduction

In [3] we presented a gradient descent based learning algorithm which was motivated by neural network learning algorithms. We gave a deterministic analysis of the convergence properties of the algorithm, using techniques of dynamical systems analysis. In the course of this analysis, we derived conditions under which the algorithm is guaranteed to learn effectively.

The conditions for convergence of the algorithm include restrictions on the topological properties of an associated error surface: specifically that it does not have non-global local minima, and does not produce an attractor

*Received November 5, 1993; received in final form June 30, 1994. Summary appeared in Volume 6, Number 6, 1996.

†This work was supported by the Australian Research Council.

at infinity. As the error surface is defined by an average over all of the training examples, it is difficult to test conditions on the error surface. The aim of this paper is to develop simpler sufficient conditions which are easier to test in applications. We have had some success in this aim, particularly with the question of attractors at infinity, although the question of existence of local minima remains intractable for some very simple parametrizations. This is not surprising when one considers the value to general optimization theory of identifying functions with no local minima. This question has long been investigated, with few helpful results [5]. The results given in this paper suggest considerable difficulty in analysing more complicated neural network parametrizations. However some potentially tractable areas for further analysis are identified.

We present a number of examples of applying these results to understanding the behaviour of the algorithm for different classes of decision regions. We look at three different parametrizations for the class of half spaces containing the origin, and show that two of these satisfy the conditions for convergence, but the third does not. The third parametrization is shown to produce an attractor at infinity. This is reminiscent of behaviour observed in neural network learning, where estimate parameters may drift off to infinity. This exercise shows that the large scale behaviour of the learning algorithm is very much dependent on the choice of parametrization. The fact that such different behaviour is observed strongly suggests (but of course does not prove) that for more complex decision regions an even wider range of behaviours may be apparent with different parametrizations of the decision boundary. We also look at a problem motivated by a radar problem, where the decision regions are stripes, and at decision regions which are an intersection of two half spaces. We show that there is no attractor at infinity for our parametrization of an (approximate) intersection of two half spaces.

In the following section we outline the algorithm presented in [3] and state the conditions for convergence that we will be discussing. In section 3 we assume that the example points are uniformly distributed. Section 3.1 contains conditions under which both local minima and attractors at infinity are excluded, whereas section 3.2 focuses on the exclusion of attractors at infinity when local minima may or may not be present. In section 4 we look at the application to learning half spaces, in section 5 we look at the problem of learning stripes, and in section 6 we discuss intersections of half spaces. In section 7 we raise the question of whether additional problems are occur when the examples are nonuniformly distributed, and we show that the results of section 3 must be modified. Section 8 concludes.

2 Problem Formulation and Previous Results

It is assumed that a sequence of data samples $((x_k, y_k))_{k \in \mathbb{Z}^+}$ are received, where $x_k \in X \subset \mathbb{R}^n$ and $y_k \in \{-1, 1\}$. X is called the *sample space* and y_k indicates the membership of x_k in some *decision region* $\Sigma \subset X$. If x_k is contained in Σ then $y_k = 1$; otherwise $y_k = -1$.

We assume that Σ belongs to a class C of subsets of X for which there exists some parameter space $A \subset \mathbb{R}^m$ and some epimorphism (onto mapping) $\Sigma : A \to C$. Moreover, it is assumed that there exists a parametrization f for C, defined below.

Definition 2.1 *Let C be a class of closed subsets of X with parameter space $A = \mathbb{R}^m$ and some epimorphism $\Sigma : A \to C$. A* parametrization *of C is a function $f : A \times X \to \mathbb{R}$, such that for all $a \in A$*

$$f(a,x) \begin{cases} > 0 & \textit{if} \quad x \in \textit{ interior of } \Sigma(a) \\ = 0 & \textit{if} \quad x \in \textit{ boundary of } \Sigma(a) \\ < 0 & \textit{if} \quad x \notin \Sigma(a) \end{cases} \tag{2.1}$$

In addition, $f(a,x)$ is required to be twice differentiable with respect to a, on $A \times X$; f, $\frac{\partial f}{\partial a}$, and $\frac{\partial^2 f}{\partial a^2}$ are to be bounded in a compact domain; and f must be Lipschitz continuous in x in a compact domain.

The algorithm proposed in [3] is as follows:

Algorithm 2.2
Step 0: *Choose the stepsize : $\mu \in (0, \infty)$.*
Choose a boundary sensitivity parameter: $\varepsilon \in (0, \infty)$.
Choose an initial parameter estimate: $a_0 \in A$.
Step 1: *Commencing at $k = 0$, iterate the recursion below:*

$$a_{k+1} = a_k - \mu \left. \frac{\partial f}{\partial a} \right|_{(a_k, x_k)} (g(a_k, x_k) - y_k), \tag{2.2}$$

where

$$g(a,x) := \frac{2}{\pi} \arctan \left(\frac{f(a,x)}{\varepsilon} \right). \tag{2.3}$$

In [3], algorithm 2.2 is be shown to be a perturbation of stepwise gradient descent of the cost function

$$J(a) = \lim_{K \to \infty} \frac{1}{K} \sum_{k=0}^{K-1} \left[f(a, x_k)(g(a, x_k) - g(a^*, x_k)) - \frac{\varepsilon}{\pi} \ln \left(1 + \frac{f(a, x_k)^2}{\varepsilon^2} \right) \right] \tag{2.4}$$

where $a^* \in A$ is the true parameter vector. Based on this fact, we derived in [3] two sets of conditions, either of which guarantees that the estimate parameters a_k asymptotically enter and remain in a neighbourhood of the parameters of the true decision region. This implies that after sufficiently many iterations, the misclassified region is very small.

One of the conditions in the first set (assumption *B2* in [3]) is that the cost function J has a unique critical point in A. If $a = a^*$ then each term in the sum $\frac{\partial J}{\partial a}$ is zero. Therefore the true parameter a^* is always a critical point of J, and *B2* says that a^* is the *only* critical point of J.

Assumption *B2* cannot hold when there is more than one a for which $f(a, \cdot) \equiv f(a^*, \cdot)$. This occurs in many important examples where there is symmetry in the parametrization. We say there is more than one true parameter value. In [3] we showed that *B2* is not necessary for effective learning, and it is sufficient to replace it with two conditions in the second set: *(C3)* local minima of J only occur at the points where $f(a, \cdot) \equiv f(a^*, \cdot)$ and *(C4)* for all $a_0 \in A$, the solution of the initial value problem (IVP)

$$\dot{a}(t) \quad = \quad -\mu \left.\frac{\partial J}{\partial a}\right|_{a(t)} \qquad ; \qquad a(0) = a_0 \tag{2.5}$$

does not cross the boundary of A, where A is assumed to be compact. That is, $a(t) \in A$ for all $t \geq 0$.

For many interesting examples, $A = \mathbb{R}^m$ for some $m > 0$, in which case *C4* says that, for all $a_0 \in A$ the solution of (2.5) is bounded for all $t \geq 0$. We say in that case that there is *no attractor at infinity* for the ordinary differential equation in (2.5). If *C4* is violated, the estimate parameters generated by algorithm 2.2 may drift off to infinity as the algorithm updates. This undesirable behaviour has been observed in neural network learning, and in practice it would be good to anticipate this divergence, and avoid it if possible.

For a given parametrization, it is not clear how to test conditions on the cost function J. This is due in part to the averaging involved in defining J, and partly due to the application of the arctan squashing function to the parametrization before taking the difference between $f(a, \cdot)$ and $f(a^*, \cdot)$.

3 Results

In this section we ignore the influence that different input sequences can play in the learning. This is achieved by assuming that the input examples $(x_k)_{k \in \mathbb{Z}^+}$ *cover* the sample space X. By this we mean that for any integrable function $f : X \to \mathbb{R}$,

$$\lim_{K \to \infty} \frac{1}{K} \sum_{k=0}^{K-1} f(x_k) \quad = \quad \frac{1}{\text{vol } X} \int_X f(x) dx. \tag{3.1}$$

Here vol $X = \int_X dx$. This is essentially a deterministic way of saying that the input examples are uniformly distributed. In particular, this means that

$$J(a) \quad = \quad \frac{1}{\text{vol } X} \int_x \left[f(a,x)(g(a,x) - g(a^*,x) - \frac{\varepsilon}{\pi} \ln \left(1 + \frac{f(a,x)^2}{\varepsilon^2} \right) \right] dx. \tag{3.2}$$

3.1 Local minima

First we deal with condition *B2*, that the cost function J has a unique critical point a^*. This is guaranteed if J is strictly increasing along rays originating at a^*. That is, the directional derivative at any point a in the direction $a - a^*$ should be positive.

Definition 3.1 *The directional derivative of $J : A \to I\!R$ at point $a \in A$ and in direction $b \in A$, is*

$$\mathcal{D}_b J(a) := \left(\left. \frac{dJ}{da} \right|_a \right)^\top b. \tag{3.3}$$

Similarly, the directional derivative of $f : A \times X \to I\!R$ with respect to $a \in A$ at point $(a,x) \in A \times X$, in direction $b \in A$, is

$$\mathcal{D}_b f(a,x) := \left(\left. \frac{\partial f}{\partial a} \right|_{(a,x)} \right)^\top b. \tag{3.4}$$

Note that $\mathcal{D}_0 J(a) = \mathcal{D}_0 f(a,x) = 0$ for any values of a and x. Theorem 3.2 and corollary 3.3 use the directional derivative to give sufficient conditions that *B2* is satisfied. For brevity, we use the notation

$$\mathcal{F}(a,x) \quad := \quad \mathcal{D}_{a-a^*} f(a,x)(f(a,x) - f(a^*,x)) \tag{3.5}$$

$$\mathcal{J}(a,x) \quad := \quad \mathcal{D}_{a-a^*} f(a,x)(g(a,x) - g(a^*,x)), \tag{3.6}$$

where g is defined by (2.3). Note that $\mathcal{F}(a^*,x) = \mathcal{J}(a^*,x) = 0$. For any $a \in \mathbb{R}^m$ we partition X into the two subsets $S_a := \{x \in X : \mathcal{F}(a,x) \geq 0\}$ and $T_a := X \backslash S_a$.

Theorem 3.2 *Assume $X \subset I\!R^n$ is compact and (x_k) covers X. Let $f : I\!R^m \times X \to I\!R$ be a parametrization of some class of decision regions, and assume $a^* \in I\!R^m$.*

If, for any $a \in I\!R^m$, there exists a set $R_a \subset S_a$ such that

$$\inf_{x \in R_a} \mathcal{F}(a,x) \text{ } vol \text{ } R_a > \left(1 + \sup_{x \in R_a} \left(\frac{f(a,x)}{\varepsilon} \right)^2 \right) \sup_{x \in T_a} |\mathcal{F}(a,x)| \text{ } vol \text{ } T_a, \tag{3.7}$$

then $\left. \frac{\partial J}{\partial a} \right|_a = 0$ if and only if $a = a^*$, where J is defined by (2.4).

The proof of this theorem is given in the appendix.

The regions S_a and T_a can be interpreted as "good" and "bad" regions respectively. If the current estimate parameter is a_k, and x_k falls in S_{a_k}, then gradient descent on the instantaneous cost, $f(a_k,x_k)(g(a_k,x_k) - g(a^*,x_k)) - \frac{\varepsilon}{\pi}\ln\left(1+\frac{f(a_k,x_k)^2}{\varepsilon^2}\right)$, will cause an update which brings the estimate closer to the true parameter vector. Thus if $x_k \in S_{a_k}$ (and μ is sufficiently small), then $\|a_{k+1} - a^*\| \leq \|a_k - a^*\|$, where a_{k+1} is determined by (2.2). However, if x_k is chosen in T_{a_k}, then the estimate parameters will move away from the true parameters. Averaging relies on the idea that the effect of the erroneous updates is negligible. This requires that for any $a \in A = \mathbb{R}^m$, the volume of T_a is small compared to that of S_a, and updates made when x_k falls in T_a are not significantly larger than when x_k falls in S_a. The following corollary deals with the special case where the volume of T_a is zero.

Corollary 3.3 *Assume $X \subset \mathbb{R}^n$ is compact and (x_k) covers X. Let $f : \mathbb{R}^m \times X \to \mathbb{R}$ be a parametrization of some class of decision regions, and assume $a^* \in \mathbb{R}^m$.*

If, for any $a \in \mathbb{R}^m$, there exists a set $U_a \subset X$ such that the closure of U_a is X and $\mathcal{F}(a,x) > 0$ for all $x \in U_a$ then $\left.\frac{\partial J}{\partial a}\right|_a = 0$ if and only if $a = a^*$, where J is defined by (2.4).

Proof: From the definition of S_a, it is clear that $U_a \subset S_a \subset X$, but the closure of U_a is X, so $S_a = X$. Therefore $\operatorname{vol} S_a = \operatorname{vol} X$ and $\operatorname{vol} T_a = 0$ for all $a \in \mathbb{R}^m$. Because $\mathcal{F}(a,x) > 0$ for all $x \in U_a$, there exists a set $R_a \subset U_a$ such that $\operatorname{vol} R_a > 0$ and $\inf_{R_a} \mathcal{F}(a,x) > 0$. Thus Theorem 3.2 applies. ▮

The above results rely on showing that J is strictly increasing along rays in parameter space originating at a^*. If this is relaxed to J being nondecreasing along the rays, we may have critical points of J which are not at a^*, so *B2* is violated. However, provided J is not constant along the rays, the behaviour of the algorithm will not be significantly altered by this relaxation, because the critical points will not be local minima of J, and all solutions of the IVP (2.5) will be bounded for all $t \geq 0$. That is, assumptions *C3* and *C4* will be satisfied. As shown in [3], this is sufficient for convergence of the algorithm.

To this point, we have only considered situations where there is a unique true parameter vector. If this is not the case, *B2* cannot hold, and there must be regions where the directional derivative of J along rays originating at any of the true parameter vectors will be negative. In this case it is much more difficult to test for local minima. We may have some success if the true parameter vectors are isolated and countable, and $\mathbb{R}^m$ can be partitioned into convex sets Λ^i, each containing exactly one true parameter vector, a^{*i}. Then if the directional derivative along all rays originating at a^{*i} is positive at all points in Λ^i, for all possible i, then there cannot be any local minima of J, and there is no attractor at infinity. Thus *C3* and *C4* are satisfied. In practice, it is generally difficult to determine the sets Λ^i, so this is probably not a useful result.

3.2 Attractors at infinity

Even when local minima of the cost function exist, it is useful to know that there is no attractor at infinity.

Algorithm 2.2 uses a finite step size and gradient descent on the instantaneous cost, rather than the average cost J, so the estimate parameters will jump around rather than moving smoothly to the minima of J. Moreover, they will continue to jump around even when they are in a local minimum, or even at the global minimum of J. Large deviations theory suggests that estimate parameters will eventually escape from the local minima (under some additional assumptions) [1]. They may also escape from the global minimum, though this is harder because the size of the updates is smaller when the value of J is closer to zero.

If there is an attractor at infinity, large deviations theory suggests that estimate parameters will eventually appear in the basin of attraction of this attractor. They will then head off to infinity and may become so large that no amount of jumping around will cause them to return to the basin of attraction of the global minimum. So if there is an attractor at infinity then ultimately the estimate parameters will converge there, even though the value of J at this attractor may be large.

If there is no attractor at infinity, all local minima of J are constrained to lie within some compact set. Parameters will not wander off to infinity but will spend most of their time near a local or global minimum of J. It is much less likely that the estimate parameters will leave a global minimum than a local minimum (since the cost driving the algorithm will be less), so the estimate parameters are likely to spend most of their time near the global minimum.

A classic result dealing with attractors at infinity gives the following (see page 204 of [4]):

Lemma 3.4 *Let $J : \mathbb{R}^m \to [0,4)$ have continuous second derivatives. If $J^{-1}([0,c])$ is compact for all $c \in [0,4)$, and $\frac{\partial J}{\partial a}\big|_a \neq 0$ except at a finite number of points $a^{*1}, \ldots, a^{*r} \in \mathbb{R}^m$, then for all $a_0 \in A$, the solution of the IVP (2.5) is bounded for all $t \geq 0$.*

So, under the conditions of Lemma 3.4, if $A = \mathbb{R}^m$ then assumption *C4* is satisfied. Using Lemma 3.4 involves determining the lower level sets

$J^{-1}([0,c])$, which is not a simple problem. Essentially the idea here is that the cost function keeps increasing as the parameter a goes to infinity in any direction. We are not worried about the shape of the cost surface for finite parameter values, but only as the parameter becomes large. Therefore we use the following, more easily tested result.

Lemma 3.5 *Let $J : \mathbb{R}^m \to [0,4)$ have continuous second derivatives and $J(a^*) = 0$ for some $a^* \in \mathbb{R}^m$. If there exists a constant $C > 0$ such that $\mathcal{D}_{a-a^*} J(a) > 0$ for all $a \in \mathbb{R}^m$, $\|a\| \geq C$ then for all $a_0 \in A$, the solution of the IVP (2.5) is bounded for all $t \geq 0$.*

Proof: Let $a(t) \in \mathbb{R}^m$. If $\|a(t)\| \geq C$ then $\mathcal{D}_{a-a^*} J(a) > 0$ so the solution of the IVP (2.5) moves towards a^*, in the sense that $\|a(t+\varepsilon) - a^*\| \leq \|a(t) - a^*\|$ for all sufficiently small $\varepsilon > 0$. Thus all solutions of the IVP (2.5) enter and remain in the closed ball with centre a^* and radius C. ▮

Lemmas 3.4 and 3.5 are true for any functions J which satisfy the stated assumptions. We can now use the ideas of the last section to determine assumptions which are specific to the cost function J defined in (2.4).

Theorem 3.6 *Assume $X \subset \mathbb{R}^n$ is compact and (x_k) covers X. Let $f : \mathbb{R}^m \times X \to \mathbb{R}$ be a parametrization of some class of decision regions, and assume $a^* \in \mathbb{R}^m$.*

If there exists a constant $C > 0$ such that, for all $a \in \mathbb{R}^m$, $\|a\| \geq C$, there exists a set $R_a \subset S_a$ such that

$$\inf_{x \in R_a} \mathcal{F}(a,x) \ \mathit{vol}\ R_a > \left(1 + \sup_{x \in R_a} \left(\frac{f(a,x)}{\varepsilon}\right)^2\right) \sup_{x \in T_a} |\mathcal{F}(a,x)| \ \mathit{vol}\ T_a, \tag{3.8}$$

then for all $a_0 \in A$, the solution of IVP (2.5) is bounded for all $t \geq 0$, where J is defined in (2.4).

Proof: The proof of this result follows similar lines to the proof of Theorem 3.2. ▮

In order to show that there is no attractor at infinity, we are only ever interested in what happens for large $\|a\|$. For some cases it may be sufficient to only calculate the limiting behaviour of $\mathcal{F}(ca,x)$ as $c \to \infty$, which is often much simpler than calculating $\mathcal{F}(a,x)$ for finite values of a. In particular, if $\lim_{c\to\infty} \mathcal{F}(ca,x)$ exists for all a and x, is never negative, and for each a $\lim_{c\to\infty} \mathcal{F}(ca,x)$ is positive for some values of x then Theorem 3.6 will hold. Or if, for each a, $\mathcal{F}(ca,x) \to \infty$ for some values of x, and $\lim_{c\to\infty} \mathcal{F}(a,x)$ exists for all other values of x then Theorem 3.6 will hold. Thus we have the following results.

Theorem 3.7 *Assume $X \subset \mathbb{R}^n$ is compact and (x_k) covers X. Let $f : \mathbb{R}^m \times X \to \mathbb{R}$ be a parametrization of some class of decision regions, and assume $a^* \in \mathbb{R}^m$. If*

1. $\lim_{c\to\infty} \mathcal{F}(ca, x) \geq 0$ *for all* $a \in A,\ a \neq 0,\ x \in X$.
2. *For all $a \in \mathbb{R}^m$, $a \neq 0$, there exists a set $V_a \subset X$ and a constant $r > 0$ such that vol $V_a \neq 0$ and $\lim_{c\to\infty} \inf_{x\in V_a} \mathcal{F}(ca, x) \geq r$.*

then for all $a_0 \in A$, the solution of IVP (2.5) is bounded for all $t \geq 0$, where J is defined in (2.4).

Proof: Choose $a \in \mathbb{R}^m$ such that $\|a\| = 1$. Define

$$\alpha_a := \frac{r \text{ vol } V_a}{2 \text{ vol } X \left(1 + \sup_{c>0} \sup_{x\in V_a} \frac{f(ca,x)^2}{\varepsilon^2}\right)}. \tag{3.9}$$

The first assumption implies that $\sup_{x\in T_{ca}} |\mathcal{F}(ca, x)| \to 0$ as $c \to \infty$, so there exists a constant C_{α_a} such that

$$\sup_{x\in T_{ca}} |\mathcal{F}(ca, x)| < \alpha_a \qquad \forall c \geq C_{\alpha_a}. \tag{3.10}$$

Assumption 2 implies that there exists a constant C_a such that

$$\inf_{x\in V_a} \mathcal{F}(ca, x) \geq \frac{r}{2} \qquad \forall c \geq C_a. \tag{3.11}$$

Let $C = \sup_{\|a\|=1}\{C_{\alpha_a}, C_a\}$.

For any $a \in \mathbb{R}^n$, we can write $a = c_n a_n$, where $c_n = \|a\|$ and $a_n = \frac{a}{\|a\|}$. By equation 3.10, if $\|a\| > C$ then

$$\sup_{x\in T_{c_n a_n}} |\mathcal{F}(c_n a_n, x)| \quad < \frac{r \text{ vol } R_a}{2 \text{ vol } X \left(1 + \sup\limits_{c>0} \sup\limits_{x\in R_a} \dfrac{f(ca_n, x)^2}{\varepsilon^2}\right)} \tag{3.12}$$

$$\tag{3.13}$$

where $R_a = V_{a_n}$. By definition, vol $T_a \leq$ vol X, so equation 3.11 implies that

$$\sup_{x\in T_a} |\mathcal{F}(a, x)| \quad < \frac{\inf_{x\in R_a} \mathcal{F}(a, x) \text{ vol } R_a}{\text{vol } T_a \left(1 + \sup_{x\in R_a} \frac{f(a,x)^2}{\varepsilon^2}\right)}. \tag{3.14}$$

Thus Theorem 3.6 holds. ∎

Theorem 3.8 *Assume $X \subset \mathbb{R}^n$ is compact and (x_k) covers X. Let $f : \mathbb{R}^m \times X \to \mathbb{R}$ be a parametrization of some class of decision regions, and assume $a^* \in \mathbb{R}^m$. If, for all $a \in \mathbb{R}^m,\ a \neq 0$,*

1. *There exists $r_a > 0$ such that $\lim_{c\to\infty} \mathcal{F}(ca, x) \geq -r_a$ for all $x \in X$.*
2. *There exists a set $V_a \subset X$ such that vol $V_a \neq 0$ and $\lim_{c\to\infty} \inf_{x\in V_a} \mathcal{F}(ca, x) = \infty$.*

then for all $a_0 \in A$, the solution of IVP (2.5) is bounded for all $t \geq 0$, where J is defined in (2.4).

Proof: The proof follows along similar lines to Theorem 3.7, the main difference being that here we can choose $\inf_{x\in R_a} \mathcal{F}(a, x)$ as large as we wish, whereas in Theorem 3.7 we can choose $\sup_{x\in T_a} |\mathcal{F}(a, x)|$ as small as we wish. ■

Converse results, giving conditions under which solutions of (2.5) are unbounded (J has an attractor at infinity), can also be given. Theorem 3.8 will be used in sections 5 and 6 to show that for particular parametrizations of a stripe and the (approximate) intersection of two half spaces there is no attractor at infinity.

4 Application — Learning a Half Space

We consider decision regions which are half spaces in $\mathbb{R}^n$ containing the origin. Three different parametrizations will be discussed. The first is the natural choice of parametrization, and there appears to be no good reason why either of the other two would be chosen. However, for more complicated classes of decision regions, it can be difficult to find a suitable parametrization, and given two candidate parametrizations it is not immediately apparent which one is preferable. By looking at different parametrizations for a half space, we illustrate some of the issues that must be considered in choosing a parametrization.

The natural choice for a parametrization of the halfspace $a^\top x + 1 > 0$ is

$$f_1(a, x) \quad = \quad a^\top x + 1, \tag{4.1}$$

where the parameter space $A = \mathbb{R}^n$. This is the parametrization used in single node artificial neural networks. In this case, the directional derivative satisfies

$$\mathcal{F}_1(a, x) \quad = \quad ((a - a^*)^\top x)^2, \tag{4.2}$$

which is never negative, so Corollary 3.3 holds. It turns out that the cost function J_1 induced by f_1 is convex.

Another suitable parametrization is

$$f_2(a,x) \quad = \quad 1 - e^{-p(a^\top x+1)}. \tag{4.3}$$

Note that in this case the magnitude of $f_2(a,x)$ will be much larger for points x outside the decision region than for those an equal distance inside the decision region. Nonetheless, the decision regions defined by $f_1(a,\cdot)$ and $f_2(a,\cdot)$ are identical. Taking the directional derivative, we see that

$$\mathcal{F}_2(a,x) \quad = \quad p(a-a^*)^\top x(e^{p(a-a^*)^\top x} - 1)e^{-2p(a^\top x+1)}, \tag{4.4}$$

which is again never negative, since $e^z > 1 \iff z > 0$. Again, Corollary 3.3 shows that the cost function induced by f_2 has a unique critical point.

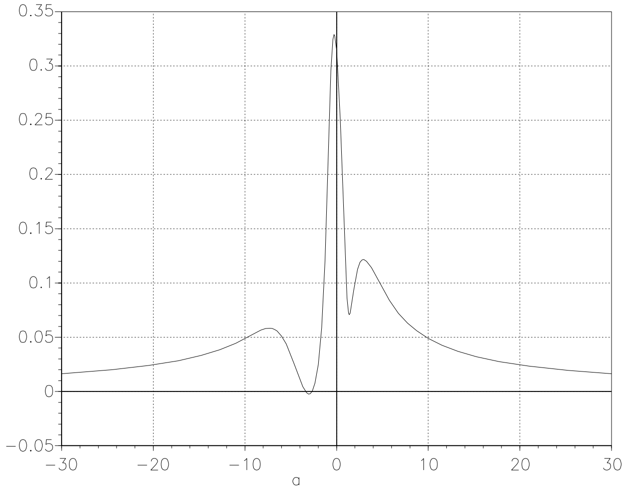

Figure 1: The cost function J_3 induced by f_3 when $a^* = -3$, $X = [-1,1]$, $\varepsilon = 0.001$ and example points (x_k) cover X. The average was determined by simple numerical integration.

Now consider the parametrization

$$f_3(a,x) \quad = \quad (a^\top x+1)e^{-(a^\top x+1)^2}. \tag{4.5}$$

For a particular a, the decision region identified by f_3 is identical to that identified by f_1. However in most simulations using this parametrization, estimate parameters drift off towards infinity. This suggests that there is an attractor at infinity generated by the parametrization (4.5). The directional derivative satisfies

$$\begin{aligned}\mathcal{F}_3(a,x) &= (a-a^*)^\top x(1-2(a^\top x+1)^2)e^{-(a^\top x+1)^2} \\ &\quad \left((a^\top x+1)e^{-(a^\top x+1)^2} - (a^{*\top} x+1)e^{-(a^{*\top} x+1)^2}\right). \quad (4.6)\end{aligned}$$

So for any $a \in \mathbb{R}^n$, in the limit $c \to \infty$,

$$\mathcal{F}_3(ca,x) \quad \to \quad 2c^3(a^{*\top}x+1)(a^\top x)^3 e^{-c^2(a^\top x)^2-(a^{*\top}x+1)^2} \to 0 \quad (4.7)$$

for all $x \in X$ such that $a^\top x \neq 0$. Thus in the limit $c \to \infty$, $S_{ca} = \{x \in X : (a^{*\top}x+1)a^\top x \geq 0\}$ and $T_{ca} = \{x \in X : (a^{*\top}x+1)a^\top x < 0\}$. We could try to prove there is an attractor at infinity by finding opposite bounds to those in Theorem 3.6. That is, we would try to show that for any $a \in \mathbb{R}^m$ there exists a set $W_a \subset X$ and a constant $r > 0$ such that vol $W_a \neq 0$ and $\lim_{c\to\infty} \mathcal{F}_3(ca,x) \leq -r$ for all $x \in W_a$. But (4.7) shows that no such r can be found in this case. Thus we have not been able to prove that there is *always* an attractor at infinity. However, for any one dimensional or two dimensional example, the cost function J_3 induced by f_3 can be plotted. Figure 1 shows the plot of J_3 for a particular one dimensional case. It can be seen from the figure that J_3 has one non global local minimum and furthermore is decreasing (albeit slowly) as $\|a\|$ goes to infinity.

5 Application — Learning a Stripe

Next consider the parametrization

$$f_4(a,x) \quad = \quad \eta - (a^\top x+1)^2, \quad (5.1)$$

where $a, x \in \mathbb{R}^n$ and $n > 0$ is some fixed constant. The decision regions identified by f_4 are "stripes" in $\mathbb{R}^n$, where $\Sigma(a)$ is of width $\frac{2\sqrt{\eta}}{\|a\|}$ and is normal to a. Such decision regions arise in a radar problem. At a, the directional derivative is

$$\mathcal{F}_4(a,x) \quad = \quad 4((a-a^*)^\top x)^2(a^\top x+1)\left(\frac{1}{2}(a+a^*)^\top x+1\right). \quad (5.2)$$

This is nonnegative if $(a^\top x+1)(\frac{1}{2}(a+a^*)^\top x+1) \geq 0$, and negative otherwise. Figure 2 depicts the regions S_a and T_a for a particular choice of estimate parameter a and true parameter a^*, when $X \subset \mathbb{R}^2$ and $A = \mathbb{R}^2$. In the limit $C \to \infty$,

$$\mathcal{F}_4(ca,x) \quad \to \quad 2c^4(a^\top x)^4 \to 0 \quad (5.3)$$

for all $x \in \{x \in X : a^\top x \neq 0\}$ and if $a^\top x = 0$ then

$$\mathcal{F}_4(a,x) \quad = \quad -4(a^{*\top}x)^2\left(\frac{1}{2}a^{*\top}x+1\right). \tag{5.4}$$

Compactness of X implies that $s^* = \sup_{x\in X} a^{*\top}x$ exists, so assumption 1 of Theorem 3.8 is satisfied, with $r_a = 4s^{*2}(\frac{1}{2}s^*+1)$. Setting $V_a = \{x \in X : |a^\top x| \geq \varepsilon\}$ for some $\varepsilon > 0$, it is clear that assumption 2 of Theorem 3.8 is also satisfied, so there is no attractor at infinity for this parametrization.

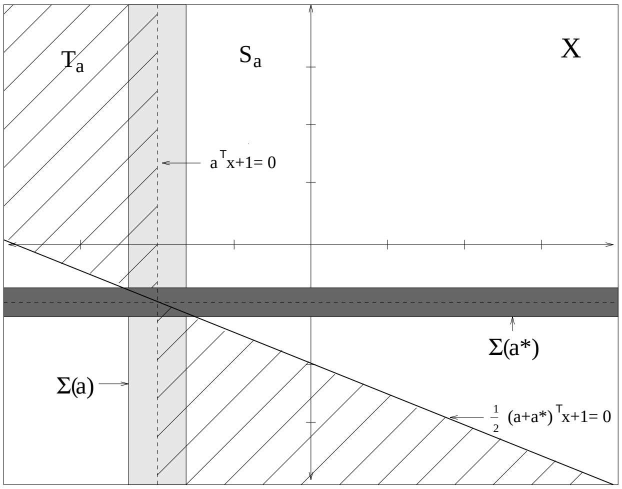

Figure 2: The "good", and "bad" regions S_a and T_a for a particular choice of a and a^*. Here the sample space is $X = [-1,1]^2$, the true parameter vector is $a^* = (0,4)$, the estimate is $a = (2,0)$, and the width is determined by $\mu = 0.04$. The dark shaded region is the true decision region, and the light shaded region is the estimate decision region. T_a is the diagonally hashed area, and S_a is the rest of X.

As an aside, note that if the input sequence (x_k) consists entirely of positive examples $(y_k = +1)$, then $X = \Sigma(a^*)$. Then S_a is much larger than T_a for any choice of a, while the size of the integrand is of the same order in both S_a and T_a. This is interesting because it explains why in simulations the estimate parameters converge much more smoothly if only positive examples are used in training. This property is a special feature of this particular class of decision regions, and certainly does not apply in general. Generally one would probably not get convergence at all with solely positive examples.

6 Application — Learning an Intersection of Two Half Spaces

Now consider decision regions which are intersections of half spaces containing the origin. If $X \subset \mathbb{R}^n$, the natural choice of parameter space is $\mathbb{R}^{2n}$. As in [3], we use the parametrization

$$f_p(a,x) \quad = \quad 1 - e^{-p(n_1^\top x+1)} - e^{-p(n_2^\top x+1)}, \tag{6.1}$$

where $a = vec(n_1, n_2)$, $n_1, n_2 \in \mathbb{R}^n$. There are two true parameter vectors in this case, $vec(n_1^*, n_2^*)$ and $vec(n_2^*, n_1^*)$, so the assumptions of Theorem 3.2 do not hold. The directional derivative satisfies

$$\begin{aligned}\mathcal{F}_p(a,x) \quad = \quad & p(n_1 - n_1^*)^\top x(e^{p(n_1-n_1^*)^\top x} - 1)e^{-2p(n_1^\top x+1)} \\ + \quad & p(n_1 - n_1^*)^\top x(e^{p(n_2-n_2^*)^\top x} - 1)e^{-2p(\frac{n_1+n_2}{2}^\top x+1)} \\ + \quad & p(n_2 - n_2^*)^\top x(e^{p(n_1-n_1^*)^\top x} - 1)e^{-2p(\frac{n_1+n_2}{2}^\top x+1)} \\ + \quad & p(n_2 - n_2^*)^\top x(e^{p(n_2-n_2^*)^\top x} - 1)e^{-2p(n_2^\top x+1)}. \end{aligned} \tag{6.2}$$

Due to the nonlinear, coupled nature of this equation, there appears to be no simple way of describing the regions S_a and T_a. The first and last terms are always positive, but the middle terms can be negative. All terms will be positive if $\text{sgn}\ (n_1 - n_1^*)^\top x = \text{sgn}\ (n_2 - n_2^*)^\top x$. So if $(n_1 - n_1^*) = c(n_2 - n_2^*)$ for some positive constant c, then $\mathcal{F}_p(a,x) \geq 0$ for all $x \in X$. That is, $S_a = X$ if $a \in \left\{(n_1, \frac{n_1 - n_1^*}{c} + n_2^*) : c \in \mathbb{R}\right\}$. As in the previous example, we can not show there are no local minima, but we can show that there is no attractor at infinity.

Let $a = vec(n_1, n_2) \in \mathbb{R}^{2n}$, where either n_1 may equal 0 or n_2 may equal 0, but not both. Define $\alpha := pn_1^\top x$ and $\beta := pn_2^\top x$. In the limit $c \to \infty$, $\mathcal{F}(ca, x)$ takes on the following values:
Case 1. $\alpha > 0$ and $\beta > 0$

$$\mathcal{F}_p(ca,x) \to 2(c\alpha e^{-c\alpha} + c\beta e^{-c\beta}) \to +0 \tag{6.3}$$

Case 2. $\alpha < 0$ and $\beta > 0$

$$\mathcal{F}_p(ca,x) \to -c\alpha e^{-2c\alpha} \to +\infty \tag{6.4}$$

Case 3. $\alpha < 0$ and $\beta < 0$

$$\mathcal{F}_p(ca,x) \to (-c\alpha e^{-c\alpha} - c\beta e^{c\beta})(e^{-c\alpha} + e^{-c\beta}) \to +\infty \tag{6.5}$$

Case 4. $\alpha = 0$ and $\beta > 0$

$$\mathcal{F}_p(ca,x) \to -pn_1^{*\top} x(e^{-pn_1^{*\top} x} - 1)e^{-2p} \geq 0 \tag{6.6}$$

Case 5. $\alpha = 0$ and $\beta < 0$

$$\mathcal{F}_p(ca, x) \to -c\beta e^{-2c\beta} \to +\infty \tag{6.7}$$

Case 6. $\alpha = 0$ and $\beta = 0$

$$\mathcal{F}_p(ca, x) = -p(n_1^* + n_2^*)^\top x(e^{-pn_1^{*\top} x} + e^{-pn_2^{*\top} x} - 2)e^{-2p} \tag{6.8}$$

Let $s_1 = \sup_{x\in X} n_1^{*\top} x$ and $s_2 = \sup_{x\in X} n_2^{*\top} x$. Setting $r_a = -p(s_1 + s_2)(e^{-ps_1} + e^{-ps_2} - 2)e^{-2p}$ and $V_a = \{x \in X : n_1^\top x \leq \varepsilon \text{ or } n_2^\top x \leq \varepsilon\}$, for some $\varepsilon > 0$, the assumptions of Theorem 3.8 are satisfied.

7 Nonuniformly Distributed Examples

The results and examples given so far in this paper have all assumed that the examples are such that (x_k) covers X. We have given conditions which guarantee that there is a unique critical point of the cost function, or that no attractors at infinity exist. But these conditions may not be sufficient if the examples do not cover X. From our definitions of the sets S_a and T_a in section 3, we see that it is often possible to construct malicious input sequences which will force the estimate parameters generated by algorithm 2.2 to diverge to infinity. If the parametrization is such that $T_a \neq \emptyset$ for all $a \in A$, a suitable malicious sequence is achieved by choosing $x_k \in T_{a_k}$ for each $k \geq 0$. In a similar vein, Sontag and Sussman [7] have given an example of a particular sequence of input examples which will cause the error surface for a simple neural network learning problem to have non global local minima.

These examples require special choices of the examples which vary with k, so this is quite a large departure from our original assumption that (x_k) covers X. A smaller relaxation of the covering assumption is that the examples are i.i.d. (independent and identically distributed) with some nonuniform distribution. In this case the results of section 3 no longer apply.

Assume that x_k are i.i.d. random variables with some known distribution $P(x) : X \to [0, 1]$. The cost function J becomes

$$J(a) = \frac{1}{\text{vol } X} \int_X \left[f(a, x)(g_\varepsilon(a, x) - g_\varepsilon(a^*, x)) - \frac{\varepsilon}{\pi} \ln\left(1 + \frac{f(a, x)^2}{\varepsilon^2}\right)\right] dP(x). \tag{7.1}$$

Results similar to those in section 3 can be derived, using $\inf P(x)$ and $\sup P(x)$ on the sets S_a and T_a. Whether or not the nature of the cost surface can be changed significantly by the choice of $P(\cdot)$ is unclear. Do there exist parametrizations for which there is no attractor at infinity when examples are uniformly distributed, but there is an attractor at infinity if examples are chosen according to some other distribution? If this is the case then satisfying the conditions for uniformly distributed points does not always guarantee good convergence if the points are nonuniformly distributed.

8 Conclusions

In this paper we have investigated conditions guaranteeing convergence of an algorithm for learning nonlinearly parametrized decision regions, discussed in [3]. The conditions given in [3] involve a cost function, where the average is taken over the input examples. The conditions are hard to test directly, so our aim here has been to develop simpler sufficient conditions which can be tested for a particular parametrization. The approach taken has been to relate the directional derivative of the cost function to the directional derivative of the parametrization, and integrate over the entire sample space.

A number of examples of the application of the new conditions to particular parametrizations have been given, and some interesting insights have been gained. We have shown that even half spaces can be parametrized in a way which will cause the learning algorithm to fail, though for the obvious linear parametrization, and for some nonlinear parametrizations of a half space, the algorithm will work. We have been able to explain why training with only positive examples gives smooth convergence for the parametrization of a stripe. The parametrization of an intersection of half spaces which was given in [3] has been investigated further, and it was shown that parameters will not drift off to infinity if points are uniformly distributed.

Appendix—Proof of Theorem 3.2

Let $a \in \mathbb{R}^m$. By the definition of a parametrization, both f and $\frac{\partial f}{\partial a}$ are bounded on a compact domain, so $\inf_{R_a} \mathcal{F}(a,x)$ and $\sup_{T_a} |\mathcal{F}(a,x)|$ are both finite. In this section we write, for instance, $\inf_{R_a} \mathcal{F}(a,x)$, rather than $\inf_{x \in R_a} \mathcal{F}(a,x)$ for the infimum over a subset of X, and similarly for supremum.

By (3.2), the directional derivative of J in the direction $a - a^*$ is

$$\mathcal{D}_{a-a^*} J(a) \quad = \quad \frac{1}{\text{vol } X} \int_{S_a} \mathcal{J}(a,x)dx + \frac{1}{\text{vol } X} \int_{T_a} \mathcal{J}(a,x)dx. \quad \text{(A.1)}$$

Note that the smoothness property of f, and hence g, has been used to interchange the order of differentiation and integration.

It is clear from the definitions of $\mathcal{F}$ and $\mathcal{J}$ that for any $a \in \mathrm{I\!R}^m$, $x \in X$,

$$\mathcal{J}(a,x) = \frac{g(a,x) - g(a^*,x)}{f(a,x) - f(a^*,x)} \mathcal{F}(a,x). \tag{A.2}$$

Because arctan is a strictly monotonic function, sgn $(g(a,x) - g(a^*,x)) =$ sgn $(f(a,x) - f(a^*,x))$. Thus for any $a \in \mathrm{I\!R}^m$, $x \in X$, sgn $\mathcal{J}(a,x) =$ sgn $\mathcal{F}(a,x)$. In particular, for any $x \in S_a$, $\mathcal{J}(a,x)$ is nonnegative, and for any $x \in T_a$, $\mathcal{J}(a,x)$ is negative.

From the intermediate value theorem, we know that for any $U \subset X$,

$$\inf_U \frac{\partial g}{\partial f} \leq \frac{g(a,x) - g(a^*,x)}{f(a,x) - f(a^*,x)} \leq \sup_U \frac{\partial g}{\partial f}. \tag{A.3}$$

But

$$\left.\frac{\partial g}{\partial f}\right|_{(a,x)} = \frac{2}{\pi} \frac{\varepsilon}{\varepsilon^2 + f(a,x)^2} < \frac{2}{\pi\varepsilon} \tag{A.4}$$

always, so $|\mathcal{J}(a,x)| < \frac{2}{\pi\varepsilon}|\mathcal{F}(a,x)|$ for any values of a and x. Similarly, for any $x \in U \subset X$, $|\mathcal{J}(a,x)| \geq \frac{2\varepsilon}{\pi(\varepsilon^2 + \sup_R f^2)}|\mathcal{F}(a,x)|$.

Now choose $a \in A$, $a \neq a^*$. From (A.1),

$$\begin{aligned}
\mathcal{D}_{a-a^*} J(a) \text{ vol } X \quad &\geq \quad \int_{R_a} \mathcal{J}(a,x)dx + \int_{T_a} \mathcal{J}(a,x)dx \\
&\geq \quad \inf_{R_a} \; \mathcal{J}(a,x) \int_{R_a} dx - \sup_{T_a} \; |\mathcal{J}(a,x)| \int_{T_a} dx \\
&\geq \quad \inf_{R_a} \frac{2\varepsilon}{\pi(\varepsilon^2 + \sup_{R_a} f^2)} \mathcal{F}(a,x) \text{ vol } R_a \\
&\qquad\quad - \sup_{T_a} \frac{2}{\pi\varepsilon} |\mathcal{F}(a,x)| \text{ vol } T_a \\
&\geq \quad \frac{2}{\pi\varepsilon} \inf_{R_a} \mathcal{F}(a,x) \text{ vol } R_a \frac{1}{1 + \sup_{R_a}(f/\varepsilon)^2} \\
&\qquad\quad - \frac{2}{\pi\varepsilon} \sup_{T_a} |\mathcal{F}(a,x)| \text{ vol } T_a \\
&\quad\; > 0
\end{aligned}$$

by assumption. Therefore $\left.\frac{\partial J}{\partial a}\right|_a \neq 0$ if $a \neq a^*$ as required.

References

[1] M.R. Frater, R.R. Bitmead, and C.R. Johnson, Jr. Escape from stable equilibria in blind adaptive equalization, *Conf. on Decision and Control*, Tuscon, Arizona, December 1992, 1756–1761.

[2] G.C. Goodwin and K.S. Sin. *Adaptive Filtering Prediction and Control*. New York: Prentice-Hall, 1984.

[3] K.L. Blackmore, R.C. Williamson, and I.M.Y. Mareels. SUMMARY: Learning nonlinearly parametrized decision regions, *J. Math. Systems, Estimation, and Control* **6(1)** (1996), 129-132. Full length paper available electronically.

[4] M.W. Hirsch and S. Smale. *Differential Equations, Dynamical Systems, and Linear Algebra.* New York: Academic Press, 1974.

[5] R. Horst and P.T. Thach. A topological property of limes-arcwise strictly quasiconvex functions, *Jour. Math. Analysis and Applications* **134** (1988), 426–430.

[6] M. Minsky and S. Papert. *Perceptrons: An Introduction to Computational Geometry.* Cambridge: MIT Press, 1969.

[7] E.D. Sontag and H.J. Sussmann. Back propagation can give rise to spurious local minima even for network without hidden layers, *Complex Systems* **3** (1989), 91–106.

DEPARTMENT OF ENGINEERING, AUSTRALIAN NATIONAL UNIVERSITY, CANBERRA ACT 0200, AUSTRALIA

DEPARTMENT OF ENGINEERING, AUSTRALIAN NATIONAL UNIVERSITY, CANBERRA ACT 0200, AUSTRALIA

DEPARTMENT OF ENGINEERING, AUSTRALIAN NATIONAL UNIVERSITY, CANBERRA ACT 0200, AUSTRALIA

Communicated by David Hill