ABSTRACT

We consider in this paper the behavior of the least squares problem that arises when one attempts to train a feedforward net with no hidden neurons. It is assumed that the net has monotonic non-linear output units. Under the assumption that a training set is separable, that is that there is a set of achievable outputs for which the error is zero, we show that there are no non-global minima. More precisely, we assume that the error is a training monotonic non-linear output units. Under the assumption that gation techniques do not need to be employed, it does provide a examples cease to exist, and in fact that one has a convergence result in [lo] however applies only to linear response theorem that caly parallels that for perceptrons. The con-

2 Definitions and Statement of Main Result

Definition 2.1 A continuously differentiable function $h : \mathbb{R} \to \mathbb{R}$ is a penalty function if there is some nonempty interval $I \subseteq \mathbb{R}$ so that:

1. $a \in I \implies h(a) = 0$
2. $a \notin I \implies h(a) > 0$ and $h'(a) \neq 0$.

By "interval" we mean infinite or finite, or even just one point. Observe that the hypotheses imply that $I = \{ a \mid h(a) = 0 \} = \{ a \mid h'(a) = 0 \}$ and in particular that $I$ must be closed.

We shall use the standard inner product notation

$$(x, y) = \sum_{i=1}^{n} x_i y_i$$

and the norm

$$\|x\| = \sqrt{(x, x)} = \sqrt{\sum_{i=1}^{n} x_i^2}$$

throughout this note.

Definition 2.2 An $E : \mathbb{R}^n \to \mathbb{R}$ is a cost function if it has the form

$$E(x) = \sum_{i=1}^{m} h_i((v^i, x))$$

where $h_i$ is a penalty function and $v^i \in \mathbb{R}^n$, for each $i = 1, \ldots, m$. 

This work was partially supported by NSF grants DMS88-03390 and DMS83-01878-01, by US Air Force grant AFOSR-86-0225, and by the CAIP Center, Rutgers University, with funds provided by the New Jersey Commission on Science and Technology and by CAIP's industrial members.
Example 2.3 Threshold LMS problems for neural nets with no hidden neurons and linear or nonlinear monotone response characteristics give rise to cost functions in the above sense. Specifically, assume given a fixed function

$$\theta : \mathbb{R} \rightarrow \mathbb{R}$$

with the property that \(\theta(0) = 0\) and \(\theta'(a) > 0\) for all \(a\). Assume that we are also given two sets of \(n\)-vectors

$$\{v^1, \ldots, v^l\} \quad \{v^{l+1}, \ldots, v^m\}$$

as well as two real numbers

$$\alpha < 0 < \beta$$

in the range of \(\theta\) (the "target values" for the first and second set respectively). We say that the sets (2) are (linearly) separable in case that there exists a vector \(z^* \in \mathbb{R}^n\) so that

$$\langle z^*, v^i \rangle < 0 \quad \text{and} \quad \langle z^*, v^j \rangle > 0$$

for each \(i = 1, \ldots, l\) and each \(j = l+1, \ldots, m\). Such a vector will be called a separating vector. Equivalently, there exists in case a vector \(z^* \in \mathbb{R}^n\) so that

$$\theta(\langle z^*, v^i \rangle) < \alpha \quad \text{and} \quad \theta(\langle z^*, v^j \rangle) > \beta$$

for each \(i = 1, \ldots, l\) and each \(j = l+1, \ldots, m\). These are penalty functions; for instance in the first case we have that

$$I = \{a \mid a \leq \theta^{-1}(\alpha)\}$$

and therefore

$$h(a) = 2(\theta(a) - \alpha)\theta'(a) > 0$$

when \(a \notin I\). Note that the sets (2) are separable if and only if there exists some \(z^*\) for which \(E(z^*) = 0\). \(\square\)

Often in neural net research one uses \(\theta(a) = \tanh(a)\), and one picks \(\alpha \in (-1,0)\) and \(\beta \in (0,1)\). Equivalently under a simple coordinate rescaling one could use the logistic function

$$\frac{1}{1 + e^{-a}}$$

in which case one takes \(\alpha\) near 0 and \(\beta\) near 1.

Example 2.4 Instead of "margins" and a threshold LMS one could also use a different type of penalty function, leading to a different kind of error function. With the same notations as above, this would be the case when one employs

$$h(a) = \begin{cases} (\theta(a) - \alpha)^2, & i = 1, \ldots, l \\ (\theta(a) - \beta)^2, & i = l+1, \ldots, m \end{cases}$$

instead of the previous penalty functions. Note that separability is not in general equivalent to the existence of an \(z^*\) so that \(E(z^*) = 0\), in this case.

Our main result, to be proved in the next section, is as follows.

Theorem 1 Let \(E\) be a cost function, and assume that there exists at least one \(z^*\) for which \(E(z^*) = 0\). Then, for each \(z^0\) the solution \(z(t)\) of the gradient differential equation

$$\dot{z} = -\nabla E(z)$$

with \(z(0) = z^0\) is defined for all \(t \geq 0\),

$$\dot{z} = \lim_{t \to \infty} z(t)$$

exists, and \(E(z) = 0\). In particular, every local minimum of \(E\) is global (\(E = 0\)).

We next discuss the consequences of this result in the case of example 2.3. Assume that two sets (2) are given, and that we pose the problem of minimizing \(E\), for any arbitrary choice of \(\alpha, \beta\) (with \(\alpha < 0 < \beta\)). In general (see last section) \(E\) may have false (non-global) locally minima. However, if the sets happen to be linearly separable then we do know from the theorem that such bad minima do not exist. More importantly, solving the differential equation (7) with a random initial state will result in a solution which converges to a minimizing value. In particular, since \(E \rightarrow 0\) along trajectories, \(\alpha\) is strictly negative, and \(\beta\) is strictly positive, there will be some finite time \(t_0\) so that \(x(t_0)\) separates.

Note that the convergence result applies to a continuous gradient modification. One might ask about the recursive discrete version

$$x_{k+1} = x_k - \rho \nabla E(x_k), \quad x_0 = x^0$$

where \(\rho > 0\) is a "learning rate." The following says that, for the example of interest, this will also converge to a solution, provided that \(\rho\) be small enough.

Corollary 2.5 If \(E\) is an in example 2.3 then for each initial vector \(x^0\) there exists a real number \(\rho\) so that the solution of the iteration (8) is so that \(z_K\) separates, for some integer \(K \geq 0\).

Proof. Consider the solution of the differential equation (7). As discussed before, there will be some time \(t_0\) so that

$$\langle v^i, x(t_0) \rangle < 0 \quad \text{and} \quad \langle v^j, x(t_0) \rangle > 0$$

for each \(i = 1, \ldots, l\) and each \(j = l+1, \ldots, m\). The equation (8) is nothing more than the Euler algorithm for calculating the solution of (7) and one knows that, if \(z_K\) denotes the solution of the Euler iteration at time \(k\) using \(\rho := \rho/k\), then

$$\|x(t_k) - z_K\| = O(\frac{\rho}{k})$$

which goes to zero as \(k \to \infty\) ([4], chapter 8). Any point close enough to \(x(t_k)\) still separates, since the inequalities (9) still hold for such a point, so for \(\rho = t_0/k\) small it indeed holds that \(z_K\) separates.

Regarding example 2.4, the same conclusions hold provided that the target values \(\alpha, \beta\) are selected so that separability of the two sets is equivalent to the existence of an \(z^*\) so that \(E(z^*) = 0\). This is always true (for any \(\alpha, \beta\) in the case of the first example, because of the equivalence of separability and the possibility of solving (4), but is in general impossible in the second example.

In fact, the paper ([1]) shows many examples of separable vectors and values \(\alpha, \beta\) for which had local minima may appear.

1-640
3 Proof of Main Result

The following simple lemma will be useful in the proof.

Lemma 3.1 If $h$ is any penalty function and if $b \notin I$ then

$$(b - a)h'(b) > 0$$

for all $a \in I$.

Proof. We assume that $I$ is bounded above, that is

$I = [b_0, \infty)$ or $I = (-\infty, b_0]$ and $b > b_0$; if instead $b$ is to the left of $I$ the proof is entirely analogous. Since $b - a > 0$ for all $a \in I$, we must show that $h'(b) > 0$.

Since $h'$ is known to be nonzero outside $I$, it has constant sign on $(b_0, \infty)$. So if $h'(b) < 0$ then it would have to be always negative in that interval, from which it would follow that

$$0 \leq h(b) < h(b_0) = 0,$$

a contradiction.

To prove the theorem we first establish the following facts:

$$\forall z \in \mathbb{R}^n, E(z) \neq 0 \Rightarrow \nabla E(z). (z - z^*) > 0$$

(10)

and

$$\forall z \in \mathbb{R}^n, \nabla E(z). (z - z^*) \geq 0$$

(11)

where $z^*$ is any vector satisfying $E(z^*) = 0$. Pick any fixed $x \in \mathbb{R}^n$, and for these $x, z^*$ introduce the scalar function

$$g(r) := E(z^* + r(z - z^*))$$

and observe that

$$g'(1) = \nabla E(z)(z - z^*)$$

so that the desired conclusions are about $g'(1)$. On the other hand, because of the form (1) of $E$, this derivative is the same as

$$\sum_{i=1}^{m} (b_i - a_i) h'(b_i)$$

(12)

where

$$a_i = (v^i, z^*)$$

and

$$b_i = (v^i, x)$$

for each $i = 1, \ldots, m$. Since $E(z^*) = 0$, it follows that all $a_i \in I$. The terms for which $b_i \notin I$ all vanish, because $h'$ is zero on $I$, while the terms with $b_i \in I$ are positive by lemma 3.1. Thus (11) holds. If $E(x) \neq 0$ then not all $b_i$ can be in $I$, from which it follows that at least one term is positive; so (10) holds too.

With respect to any fixed $z^*$ for which $E(z^*) = 0$ we define the function

$$V(x) := \frac{1}{2}||x - z^*||^2$$

to play the role of a Lyapunov function for the gradient system (7). Along its trajectories, we have that

$$\frac{dV(x(t))}{dt} = V(x(t))$$

(13)

where we are denoting

$$\dot{V}(x) := -\nabla E(z). (z - z^*)$$

as is usually done in qualitative ODE theory. From (11) we know that

$$\dot{V}(x) \leq 0$$

for all $x$, so $V$ decreases along trajectories. Furthermore, from (10) we also know that

$$V(x) = 0 \Rightarrow E(z) = 0$$

for all $x$.

For any initial condition $x(0)$, the trajectory $x(t)$ remains in the compact set

$$\{x \mid V(x) \leq V(x(0))\}$$

so it is defined for all $t \geq 0$.

The LaSalle Invariance Principle (see for instance [5], theorem 6.4) guarantees then that there is some real number $\mu$ such that

$$x(t) \to E^{-1}(0) \cap V^{-1}(\mu)$$

(14)

for this trajectory. If $\mu = 0$ this reduces to one point, and the theorem is proved for that trajectory. This value may not be zero, but we next prove that by modifying $V$ (that is, choosing a $V$ corresponding to a different $z^*$) it can be made zero. If it is established, the theorem will be proved.

Suppose then that $x(t)$ is a trajectory for which $\mu > 0$, and pick any $\omega$-limit point $z$ of this trajectory, that is to say some point to which a subsequence $x(t_i), t_i \to \infty$, converges. By (14), $E(z) = 0$. So we can repeat the above argument using $z$ as the new $z^*$. Now necessarily $\mu = 0$, and we are done.

4 Closing Remarks

If the hypothesis that $E(z^*) = 0$ for some $z^*$ is dropped, there may exist local minima of $E$ which fail to be global, even in the situation in example 2.3. For instance, in [9] a set of $m = 125$ vectors is given,

$$\{v^1, \ldots, v^m\}$$

all whose entries are equal to 1 or -1, for which

$$F(z) = \sum_{i=1}^{m} (\theta((v^i, z)) - 1)^2$$

has a local minimum which is not global, and $\theta = \tanh$. (There $n = 5$, which can be interpreted as giving 4 input neurons plus a bias weight, to be learned in a neural net with no hidden layers.) Let $z$ be so that $F$ has a local minimum at $z$ but there exists some $y$ so that

$$F(y) < F(z).$$

(15)

We will pick some number $\beta \approx 1$ which is larger than all the values

$$\theta((v^i, z))$$

and

$$\theta((v^i, y))$$

and consider the cost function $E$ given in example 2.3, where $l = 0$ and $\alpha$ is irrelevant. By continuity, for large enough $\beta$ it will hold that $x$ is a local minimum of $E$ and that (15) holds for the approximation $E$, that is, $E(y) < E(z)$. Thus we have an example where $E$ has a local nonglobal minimum. (If binary examples are not required, it is easy to construct examples with smaller $m$; see [8].)
This discussion serves also to illustrate the substantial difference that exists between the case of interest in neural nets, where a nonlinear function \( \theta \) is used, and the more standard case in pattern recognition, where one may consider a cost function as in example [2.3] but with \( \theta(a) = a \). (The "relaxation case" in [2], pp.147ff.) In that case, there are no nonglobal local minima even if the data is not separable. This is proved as follows. Each term 

\[ h_i(v^i, x) \]

in equation (1) is a convex function of \( x \), since along each line \( x + ry, r \in [0,1] \) the second derivative 

\[ \frac{d^2}{dx^2} h_i(v^i, x + r(v^i, y)) \]

is nonnegative: it equals 

\[ 2(v^i, y)^T h_i''(v^i, x) + r(v^i, y) \]

and the second derivative of \( h_i \) is always nonnegative, because \( h_i \) is quadratic in one interval and constant in another, as per equations (5) and (6). It follows that the cost function \( E \) is also convex, since it is the sum of convex functions, and therefore \( E \) has no bad local minima.

There is yet another important difference with the case \( \theta = \text{identity} \). In the above reference a result is proved which is somewhat analogous to corollary 2.5, but which establishes instead (with a different proof, for the "online" version where each term in the cost function is used one at a time, and with a small modification if the \( v_i \)'s are not unit vectors) that the discrete scheme (8) monotonically diminishes the distance to any fixed separating vector, for every fixed choice of \( \rho \in (0,2) \). This will not happen in general in the nonlinear case.

As we pointed out, the convergence result for the threshold-LMS problem is the one that has more interest. For the non-threshold case (example 2.4), the authors of the paper [7] already had established a related convergence result for nonlinear units. They dealt with discrete stochastic approximation rather than the gradient descent differential equation itself, which makes the techniques quite different. In addition certain hypotheses are made in that paper (binary inputs and a linear independence assumption on the data) that make their result somewhat more restricted, but a general proof based on their ideas (for the difference equation case) is very probably also possible.

Finally, we compare with the results in [10]. The authors there define a class of functions \( h \) called well-formed functions, which play the same role in the total cost as our penalty functions, and a result (not convergence of weights, but decrease of the error function to zero) is proved for the gradient differential equation. However, the definition of well-formed function does not include sigmoidal nonlinearities, since it requires that \( h \) have a derivative bounded away from zero while there are misclassifications. But the authors did emphasize the fact that threshold LMS is essential in order to avoid the examples where perceptrons classify but backprop doesn't.

### 4.1 Acknowledgment

The authors wish to thank Geoff Hinton for many useful comments and for asking the questions that led to this note.

### 5 References