BACKPROPAGATION SEPARATES WHERE PERCEPTRONS DO[1]

Eduardo D. Sontag
Héctor J. Sussmann
Rutgers Center for Systems and Control
Department of Mathematics, Rutgers University, New Brunswick, NJ 08903

ABSTRACT

Feedforward nets with sigmoidal activation functions are often designed by minimizing a cost criterion. It has been pointed out before that this technique may be outperformed by the classical perceptron learning rule, at least on some problems. In this paper, we show that no such pathologies can arise if the error criterion is of a *threshold* LMS type, i.e., is zero for values "beyond" the desired target values. More precisely, we show that if the data are linearly separable, and one considers nets with no hidden neurons, then an error function as above cannot have any local minima that are not global. Simulations of networks with hidden units are consistent with these results, in that often data which can be classified when minimizing a threshold LMS criterion may fail to be classified when using instead a simple LMS cost. In addition, the proof gives the following stronger result, under the stated hypotheses: the continuous gradient adjustment procedure is such that from any initial weight configuration a separating set of weights is obtained in finite time. This is a precise analogue of the Perceptron Learning Theorem. The results are then compared with the more classical pattern recognition problem of threshold LMS with linear activations, where no spurious local minima exist even for nonseparable data: here it is shown that even if using the threshold criterion, such bad local minima may occur, if the data are not separable and sigmoids are used.

1 Introduction

This paper deals with the behavior of the so-called backpropagation technique (Hinton, 1987; Rumelhart & McClelland, 1986) used in training feedforward nets for pattern classification. This technique is based on (1) proposing an error function that penalizes missclassifications and then (2) attempting to minimize this function using a gradient descent method (the name "backpropagation" arises from the use of the chain rule to compute partial derivatives recursively through the network's layers). In (Wittner & Denker, 1987; Brady, Raghavan, & Slawny, 1989), examples are given illustrating the fact that even if the training data are linearly separable –a case already treated satisfactorily by linear programming techniques as well as the classical "perceptrons"–, a net performing gradient search may get stuck in a solution which fails to classify correctly. The first of these papers (see also Shrivastava & Dasgupta, 1987) pointed out that (a) it might be possible to overcome these difficulties by using a *threshold* LMS procedure, where one does not penalize numerical values which are already beyond the

[1]This work was partially supported by NSF grants DMS88-03396 and DMS89-02994, and by the CAIP Center, Rutgers University.
Keywords: Backpropagation, pattern classification, nonlinear least squares, neural networks
Phone: (908)932-3072; email: sontag@hilbert.rutgers.edu

targets, and (b) in the threshold case one indeed has, under certain assumptions on the activation functions, a convergence theorem that closely parallels that for perceptrons. The apparent contradiction with the title of (Brady, Raghavan, & Slawny, 1989) is explained by the fact that this latter reference did not use threshold but rather "exact" LMS.

In this paper, we extend the result in (Wittner & Denker, 1987) so as to include sigmoidal neurons, which were excluded by their assumptions. To be precise, we show that if the data are linearly separable and one considers nets with no hidden neurons and monotonic non-linear output units, then the threshold error function has no non-global local minima and in fact the continuous gradient adjustment procedure is such that from any initial weight configuration a separating set of weights is obtained in finite time. The main result is stated in terms of the convergence of gradient procedures for the minimization of a general class of cost functions that includes this and other examples of interest in neural networks. The technique of proof involves some elementary stability arguments. We also show how our result is an analogue of the Perceptron Learning Theorem.

The paper is organized as follows. First we give an intuitive discussion of the problem, including a comparison with perceptrons, and we see how a threshold LMS criterion is suggested by this comparison. After that, we describe the type of error function to be considered and state the main theorem, discussing its application to sigmoidal nets. Then we provide some examples, and in particular we show that if the training set is *not* separable, there may be nonglobal local minima *even if* a threshold LMS is used. We also compare the situation with the case of linear response units (Duda & Hart, 1973, pp.148-149), and remark that a basic difference with that case is due to the lack of convexity in the cost function: for linear activations, there are no spurious local minima even for nonseparable data, in contrast to sigmoidal nets. In the final section we prove the theorem.

Although the results in this paper are mathematically quite straightforward, they do serve to emphasize the need for care in the choice of error function. Obviously, it would be far better to have a positive result for more general nets than those with no hidden layers, since after all the need for extra layers justifies considering sigmoidal nets to begin with. An analogous result cannot hold in that generality, as spurious local minima can occur for multilayer-classifiable data, but one can expect at least the domains of attraction to be larger when using threshold errors, and this is confirmed easily by computer simulations. For more on the rigorous analysis of local minima in the multilayer case, the reader is referred to (Blum, 1989).

2 Perceptrons Versus Error Minimization

Neural nets are typically applied to solve binary classification problems. In these, labeled examples and counterexamples are presented during a training stage, and weights are adjusted so as to make the network's numerical output match in some sense the desired classification.

Assume that we are given a data sequence of labeled n-vectors

$$w^1, \epsilon_1;\ w^2, \epsilon_2;\ \ldots;\ w^m, \epsilon_m \tag{1}$$

where $\epsilon_i = +$ or $\epsilon_i = -$ for each i. We say that the data are *linearly separable* in case there exists a hyperplane H in $\mathbb{R}^n$ such that all the points w^i with $\epsilon_i = +$ are on an opposite side of H than the points with $\epsilon_i = -$. Adding a coordinate equal to one to each w^i, and writing $v^i := (w^i, 1)$, the condition is that there must exist a vector x^* in dimension $n+1$ such that

$$\langle v^i, x^* \rangle < 0 \text{ and } \langle v^j, x^* \rangle > 0 \quad \text{for each } v_i \in S_- \text{ and each } v_j \in S_+ \ , \tag{2}$$

where we use the standard inner product notation

$$\langle x, y \rangle = \sum_{i=1}^{n} x_i y_i$$

and we denote by S_+ the set of all vectors v^i such that $\epsilon_i = +$ and S_- the set of those with $\epsilon_i = -$. An x^* as in (2) is called a *separating vector*. The problem of determining if there exists such an x^* is a simple linear programming question, and there are very efficient methods for solving it as well as for finding explicit solutions x^*. More in connection with nets, the classical *perceptron learning procedure* (see e.g. Duda & Hart, 1973) provides a recursive rule for finding one such solution provided that any exist.

Although the perceptron rule is very simple, we will briefly recall it, so we can compare it with error minimization. First an arbitrary starting value is selected as an estimate for x^*. Then, the vectors v^i are presented one after the other in an infinite sequence, with the only restriction that every element must appear infinitely often. For each element of this sequence, the corresponding inequality is tested. If the sign of the inequality is wrong, the estimate for x^* is updated as follows:

$$x^* := x^* + v^i$$

if v^i is in S_+, and

$$x^* := x^* - v^i$$

if v^i is in S_-; if the sign was right, no change is made. It is well-known (see e.g. Duda & Hart, 1973) that this procedure converges in finitely many steps to a solution x^* if the original data are linearly separable.

To compare this with the error minimization technique, we consider a fixed differentiable function

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

with the property that $\theta(0) = 0$ and $\theta'(a) > 0$ for all a (and, for technical reasons, such that the derivative θ' is locally Lipschitz). Let $t_+ := \lim_{a\to+\infty} \theta(a)$ and $t_- := \lim_{a\to-\infty} \theta(a)$, where $-\infty \le t_- < t_+ \le \infty$. We pick two real numbers

$$t_- < \alpha < 0 < \beta < t_+$$

in the range of θ —the "target values" for the negative and positive examples, respectively; see for instance (Blum, 1989; Brady, Raghavan, & Slawny, 1989; Wittner & Denker, 1987). Often in neural net research one uses $\theta(u) = \tanh(u)$ or, equivalently under a simple coordinate rescaling, the logistic function $1/(1 + e^{-u})$. If x^* is as in (2) then there exists some $\gamma > 0$ so that $\gamma\langle v^i, x^*\rangle \le \theta^{-1}(\alpha)$ and $\gamma\langle v^j, x^*\rangle \ge \theta^{-1}(\beta)$ respectively. That is,

$$\theta(\langle v^i, x^*\rangle) \le \alpha \text{ and } \theta(\langle v^j, x^*\rangle) \ge \beta \text{ for each } v_i \in S_- \text{ and each } v_j \in S_+ \;, \tag{3}$$

after redefining x^* as γx^*. Conversely, if an x^* exists so that (3) holds, then the data must be linearly separable and this same x^* is a separating vector. Now consider the cost

$$E(x) := \sum_{i=1}^{m} \Big(\delta_i - \theta(\langle v^i, x\rangle)\Big)^2 \tag{4}$$

where δ_i equals β if $v^i \in S_+$ and equals α otherwise. One now minimizes E as a function of x. If a small value results, it follows that each $\theta(\langle v^i, x\rangle)$ is approximately equal to δ_i, and the classification problem is solved. Unfortunately, there are local minima problems associated to this procedure; see for instance, (Wittner & Denker, 1987; Brady, Raghavan, & Slawny, 1989; Sontag, 1988). Spurious (i.e., non-global) local minima can occur *even if the data are separable.* Moreover, even when there happens to be only one local (and thus necessarily global) minimum, the resulting solution may fail to separate (hence the title of Brady, Raghavan, & Slawny, 1989). As this general error-minimization procedure is in principle exactly the same that is used when dealing with the far more interesting case of networks with hidden layers and data that is classifiable to higher order than linear, it is of interest to study the problem in this simpler case, where a comparison with the perceptron is possible. One may expect that the same pathologies will arise in the more general case; indeed, this has recently been shown by (Blum, 1989).

There are two very different reasons for the above local minima to appear. One of the reasons has to do with the highly nonconvex nature of the problem, when using nonlinear functions θ. We discuss this in Section 4. Another reason, and the one that interests us more in this note, is as follows. When minimizing E, one is trying to fit the values α and β *exactly*, whereas it would obviously be sufficient for classification that $\theta(\langle v^i, x\rangle)$ be less than α for v^i in S_-, and bigger than β for v^i in S_+. The precise fitting may force the parameters x to be chosen so as to make most terms small at the cost of leaving one term large. This can be illustrated with a simple example. Assume that $n = 1$ and the data consists of five points w^i so that w^1 and w^2 are very close to -1, w^3 and w^4 are very close to 1, and $w^5 = -0.9$. One desires for $\{w^1, w^2\}$ to be classified as "$-$" and $\{w^3, w^4, w^5\}$ as "$+$". Obviously this data are linearly separable (pick $x^* = (1, 0.95)$). We now pose the minimization problem, using $\theta = \tanh$ and the target values $\alpha := -0.6$, $\beta := 0.6$. That is, we must minimize the error

$$(0.6 - \theta(-0.9x_1 + x_2))^2 + 2\,(0.6 - \theta(x_1 + x_2))^2 + 2\,(-0.6 - \theta(-x_1 + x_2))^2$$

as a function of $x = (x_1, x_2)$. There is a unique local (and global) minimum, attained at the unique value of (approximately) $x = (0.4855, 0.258)$. It turns out that this vector does not separate: for $v = (-0.9, 1)$, $\langle v, x\rangle = (-0.9)0.4855 + 0.258 < 0$. Therefore, in minimizing E, the classification of w^5 has been traded-off for a better fit of the rest of the data. Note that the cutoff between classes happens approximately at $w = -1/2$. On the other hand, with $x^* = (13.863, 13.17)$, equation (3) does hold, as the values of for $(-1, 1)$, $(-0.9, 1)$ and $(1, 1)$ are approximately at -0.6, 0.6, and 1, respectively; the cutoff happens at $w = -0.95$. This x^* does not arise from the minimization of E, but it can be obtained by not adding to the error if $\theta(\langle v^i, x\rangle)$ is already less than α and $x \in S_-$, and similarly for $x \in S_+$ if the value exceeds β. In other words, the corresponding term in Equation (4) is taken to be zero. So one must minimize instead

$$E^*(x) \;:=\; \sum_{i=1}^{m} h_i(\langle v^i, x\rangle) \tag{5}$$

where

$$h_i(a) :=\; (\theta(a) - \alpha)_+^2 \;\text{ if }\; \epsilon_i = -$$

and

$$h_i(a) :=\; (\beta - \theta(a))_+^2 \;\text{ if }\; \epsilon_i = +$$

and we are denoting

$$(u)_+^2 :=\; \frac{1}{2}(u + |u|) = \begin{cases} 0 & \text{if } u \le 0 \\ u & \text{if } u > 0 \end{cases}$$

for all numbers u. The new error function E^* is differentiable, but in general is not second-order differentiable. (However, its partial derivatives are locally Lipschitz, which is all that will be needed in order to have a well-posed gradient differential equation.) Note that for separable data the global minimum of E^* is zero and it is achieved at any x^* that satisfies (3); conversely, if the data are not separable then the global minimum is strictly positive.

We now compare minimization of E^* to the perceptron procedure. A discrete gradient descent step (with stepsize ρ) for minimizing E^* takes the form of updating x by:

$$x \;:=\; x + \tilde{\rho} v^i \;,$$

where

$$\tilde{\rho} \;=\; (\delta_i - \theta(\langle v^i, x\rangle))\,\theta'(\langle v^i, x\rangle)\,\rho \tag{6}$$

if the current classification of v^i is incorrect and $\tilde{\rho} = 0$ otherwise. This is the precise analogue of the perceptron rule (for which $\tilde{\rho}$ is always either zero or ± 1). When using E instead of E^* one would use Equation (6) always, *even if the classification was correct.* The parameters $x^* = (13.863, 13.17)$ were obtained numerically by minimizing E^*; alternatively, they can be found in closed form by first fitting exactly the values at -1 and -0.9 to obtain $x_1 = 10 \ln 4$ and $x_2 = 9.5 \ln 4$; the value at 1 turns out to be about 1, which is greater than 0.6 but still contributes zero error in E^*. The use of E^* was first suggested in (Wittner & Denker, 1987), who also proved a convergence result under somewhat restrictive hypotheses which do not allow for sigmoids. Below, we prove that there are no spurious local minima for E^* if the data are separable, and that the gradient descent differential equation converges in finite time to a separating solution, from any initial state. Moreover, we give a version that applies to a wider class of optimization problems. The only difficulty in the proof has to do with the fact that $\tilde{\rho}$ will tend to zero; one has to use an argument involving LaSalle invariance, but this is standard in dynamical systems theory. Thus we conclude that the use of E^* provides the correct generalization of the perceptron learning rule, and this provides strong evidence that one should always use such threshold-LMS procedures.

3 Penalty Functions and Examples

We give a general definition that captures what is needed in order to prove the result for threshold LMS cost functions.

Definition 3.1 A differentiable function $h : \mathbb{R} \to \mathbb{R}$, with locally Lipschitz derivative h', is a *penalty function* if there is some nonempty interval $I \subseteq \mathbb{R}$ so that:

1. $a \in I \Rightarrow h(a) = 0$
2. $a \notin I \Rightarrow h(a) > 0$ 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.

Definition 3.2 An $E : \mathbb{R}^n \to \mathbb{R}$ is a *cost function* if it has the form

$$E(x) = \sum_{i=1}^{m} h_i(\langle v^i, x \rangle) \tag{7}$$

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

Our main result, to be proved in Section 5, is as follows.

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

$$\dot{x} = -\nabla E(x)^{\mathrm{t}} \tag{8}$$

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

$$\tilde{x} = \lim_{t\to\infty} x(t)$$

exists, and $E(\tilde{x}) = 0$. In particular, every local minimum of E is global ($E = 0$).

3.1 The Example of Threshold LMS

Threshold LMS problems for neural nets with no hidden neurons and linear or nonlinear monotone response characteristics give rise to the cost functions $E = E^*$ introduced in Section 2 (we state everything in terms of the vectors $(w^i, 1)$, and for notational simplicity we write n instead of $n+1$). There is given a sequence of n-vectors $v^1, \ldots, v^m$ together with a sequence of desired signs $\{\epsilon_i = \pm\}$, as well as a map $\theta : \mathbb{R} \to \mathbb{R}$ as there, and any two values $\alpha < \beta$ in the range of θ. Associated to these is the error function E^* given in (5). Observe that the functions h_i are penalty functions; for instance for $\epsilon = -$ we have that

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

and therefore

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

when $a \notin I$.

In general (see Section 4), E^* may have spurious (non-global) locally minima. However, if the data happen to be linearly separable then we do know from the Theorem that such local minima do not exist. This is because, as discussed earlier, the data are separable if and only if there exists some x^* for which $E^*(x^*) = 0$. Note that this holds for any choice of the target values α, β, independently of the actual training data. Furthermore, one has the following observation:

Corollary 3.3 If E^* is as above, then, solving the differential equation (8) with an arbitrary initial state x^0, there is some t_0 so that $x(t)$ is a separating vector, for each $t \geq t_0$.

Proof. Let $\tilde{x}$ be as in the Theorem, so that (3) holds at $\tilde{x}$. By continuity, $\theta(\langle v^i, x \rangle) < \alpha/2$ for each $v_i \in S_-$ and $\theta(\langle v^j, x \rangle) > \beta/2$ and each $v_j \in S_+$, for any x near $\tilde{x}$. Thus, for t large enough, $x^* = x(t)$ separates. $\blacksquare$

As stated, the convergence result applies only to continuous gradient descent. One might ask about the recursive discrete version

$$x_{k+1} := x_k - \rho \nabla E(x_k)^{\mathrm{t}}, \;\; x_0 = x^0 \tag{9}$$

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 ρ be small enough.

Corollary 3.4 If E^* is as above, then for each initial vector x^0 there exists a real number ρ so that the solution of the iteration (9) is so that x_K separates, for some integer $K \geq 0$.

Proof. Consider the solution of the differential equation (8). By Corollary 3.3, there is some t_0 so that $x(t_0)$ satisfies (3). The difference equation (9) is nothing more than the Euler algorithm for calculating the solution of (8) and one knows that, if x_k^ρ denotes the solution of the Euler iteration at time k using $\rho := t_0/k$, then

$$\|x(t_0) - x_k^\rho\| = O(\frac{t_0}{k})$$

which goes to zero as $k \to \infty$ (Isaacson & Keller, 1966, chapter 8). As before, any point close enough to $x(t_0)$ still separates, so for $\rho = t_0/k$ small it indeed holds that x_k^ρ separates. ∎

3.2 Exact LMS

Instead of a threshold LMS one could also use an "exact" LMS criterion, leading to a different kind of error function. With the same notations as above, this would be the case when one employs $h_i(a) := (\delta_i - \theta(a))^2$, where δ_i equals β if $v^i \in S_+$ and equals α otherwise. While Theorem 1 still guarantees global convergence to a solution of $E(x) = 0$ provided that one such solution exists, and in that case the corresponding vector x will indeed separate, in this example separability is *not* in general equivalent to the existence of an x^* so that $E(x^*) = 0$ (when the targets α, β have been chosen a priori, independently of the actual training data), and the (even global) minima of E need not separate, as discussed in Section 2.

4 Some Remarks

If the hypothesis that $E(x^*) = 0$ for some x^* is dropped, there may exist local minima of E which fail to be global, even in the situation of threshold LMS, and even if the vectors v^i are restricted to be binary, as in many applications. For instance, in (Sontag & Sussmann, 1989), the following labeled sequence of $m = 125$ vectors is given:

$$\begin{array}{rcl}
w^1, \ldots, w^{15} & = & (-1,-1,1,-1) \rightsquigarrow - \\
w^{16}, \ldots, w^{30} & = & (-1,-1,-1,1) \rightsquigarrow - \\
w^{31} & = & (1,1,-1,-1) \rightsquigarrow - \\
w^{32} & = & (1,-1,1,-1) \rightsquigarrow - \\
w^{33} & = & (-1,1,-1,1) \rightsquigarrow - \\
w^{34}, \ldots, w^{48} & = & (1,1,-1,-1) \rightsquigarrow + \\
w^{49}, \ldots, w^{63} & = & (1,-1,1,-1) \rightsquigarrow +
\end{array}$$

$$
\begin{aligned}
w^{64},\dots,w^{78} &= (-1,1,-1,1) \rightsquigarrow + \\
w^{79} &= (-1,-1,1,-1) \rightsquigarrow + \\
w^{80} &= (-1,-1,-1,1) \rightsquigarrow + \\
w^{81},\dots,w^{125} &= (1,1,1,1) \rightsquigarrow +
\end{aligned}
$$

for which, with $\varepsilon_i = -1$ for $i \leq 33$ and $\varepsilon_i = 1$ otherwise,

$$E(x) = \sum_{i=1}^{m} (\theta(\langle v^i, x\rangle) - \varepsilon_i)^2$$

has a local minimum which is not global, when $\theta = \tanh$. (Note that one must use the examples from (Sontag & Sussmann, 1989) rather than those from (Brady, Raghavan, & Slawny, 1989) or (Sontag, 1988), not just because of the interest in binary examples, but also because in the latter references outputs are not allowed to take limiting values $\{-1, 1\}$, which will be critical below.)

The main result from (Sontag & Sussmann, 1989) is then: *The above error function E has at least one local minimum which is not a global minimum.*

The proof gives a somewhat stronger conclusion than stated, in that the local minimum in question is strict in the following sense: there exists a vector y_0 and a closed ball B not containing y_0 so that $E(y_0) < E(x)$ for all x in B, and so that the minimum of E on B is attained only in the interior of B.

We will next show how to obtain particular values α and β so that, for this same example and the threshold LMS cost, E^* has spurious local minima. Consider

$$F(x,\alpha,\beta) := \sum_{i=1}^{33} (\theta(\langle v^i, x\rangle) - \alpha)_+^2 + \sum_{i=34}^{125} (\beta - \theta(\langle v^i, x\rangle))_+^2$$

as a function of x and the real numbers α, β. Since $(u)_+$ is continuous in u, F is too. From the previous discussion, it follows that there exists some number $\varepsilon > 0$ and some x_0 in the interior of B, so that $F(x,-1,1) > f_0 + \varepsilon$ for all $x \in S$, where S is the boundary of the ball B and $f_0 := F(x,-1,1)$. Moreover, this ε can be chosen so that also $F(y_0,-1,1) < f_0 - \varepsilon$. Note that $F(x,-1,1) \geq f_0$ for all x in B.

By uniform continuity of F on compacts, for all $\alpha > -1$ and $\beta < 1$ sufficiently close to $-1, 1$ it will hold that

$$F(x,\alpha,\beta) > f_0 + \varepsilon$$

for all x in S as well as

$$F(y_0,\alpha,\beta) < f_0 - \varepsilon$$

and

$$F(x,\alpha,\beta) > f_0 - \varepsilon$$

for all x in B. Furthermore, we may assume that

$$F(x_0,\alpha,\beta) < f_0 + \varepsilon$$

so that the minimum of $F(x,\alpha,\beta)$ on B must be achieved only in the interior of B. For any such $(\alpha,\beta) \neq (-1,1)$ it then holds for $E^*(x) := F(x,\alpha,\beta)$ that E^* has a local minimum on B which cannot be global.

4.1 Comparison With Other Results

Th above discussion serves also to illustrate the substantial difference that exists between the case of interest in neural nets, when a nonlinear function θ is used, and a standard case in pattern recognition, that of threshold cost functions as before but with $\theta(a) = a$. (The "relaxation case" in (Duda & Hart, 1973, pp.147ff).) In that case, there are no nonglobal local minima *even if the data are not separable.* This is proved as follows. Each term

$$h_i(\langle v^i, x\rangle)$$

in equation (7) is a convex function of x, since along each line $x + ry, r \in [0, 1]$ the second derivative

$$\frac{d^2}{dr^2} h_i(\langle v^i, x\rangle + r\langle v^i, y\rangle)$$

is nonnegative: it equals

$$2\langle v^i, y\rangle^2 \; h_i''(\langle v^i, x\rangle + r\langle v^i, y\rangle)$$

and the second derivative of h_i is always nonnegative, because h_i is quadratic in one interval and constant in another. 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 θ=identity. In the above reference a result is proved which is somewhat analogous to Corollary 3.4, 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 (9) 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, the authors of the paper (Shrivastava & Dasgupta, 1987) 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) may be possible also.

Finally, we compare with the results in (Wittner & Denker, 1987). The authors here 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 missclassifications.

5 Proof of Main Result

The following simple lemma will be useful in the proof.

Lemma 5.1 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

$$I = [a_0, b_0] \text{ or } I = (-\infty, b_0]$$

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:

$$\boxed{\forall x \in \mathbb{R}^n, E(x) \neq 0 \Rightarrow \nabla E(x).(x - x^*) > 0} \tag{10}$$

and

$$\boxed{\forall x \in \mathbb{R}^n, \nabla E(x).(x - x^*) \geq 0} \tag{11}$$

where x^* is any vector satisfying $E(x^*) = 0$. Note that

$$\nabla E(x)(x - x^*) = \frac{\mathrm{d}}{\mathrm{d}\,r}|_{r=1}\, E(x^* + r(x - x^*))$$

so this expression equals

$$\sum_{i=1}^{m} (b_i - a_i) h'(b_i) \tag{12}$$

where

$$a_i = \langle v^i, x^* \rangle$$

and

$$b_i = \langle v^i, x \rangle$$

for each $i = 1, \ldots, m$. Since $E(x^*) = 0$, it follows that all $a_i \in I$. The terms for which $b_i \in I$ all vanish, because h' is zero on I, while the terms with $b_i \notin I$ are positive by Lemma 5.1. 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 x^* for which $E(x^*) = 0$ we define the function

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

to play the role of a Lyapunov function for the gradient system (8). Along its trajectories, we have that

$$\frac{dV(x(t))}{dt} = \dot{V}(x(t)) \tag{13}$$

where we are denoting

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

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

$$\dot{V}(x) = 0 \Rightarrow E(x) = 0 \tag{14}$$

for all x.

For any initial condition $x(0)$, the trajectory $x(\cdot)$ exists at least locally, and is unique by the Lipschitz hypotheses made. Furthermore, it remains in the compact set

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

so it is defined for all $t \geq 0$. (See for instance (Sontag, 1990), Proposition C.3.9.)

The *LaSalle Invariance Principle* (see for instance (LaSalle, 1976), Theorem 6.4, or the particular case in (Sontag, 1990), Lemma 4.6.6) says that if the trajectory is bounded and V is nonincreasing along this trajectory, then there is some real number μ such that the solution $x(t)$ converges to the set where $\dot{V}(x) = 0$ and $V(x) = \mu$. Because of (14), we conclude that

$$x(t) \ \rightarrow \ E^{-1}(0) \bigcap V^{-1}(\mu) \tag{15}$$

for this trajectory. Observe that this does not yet say that the solution is converging. If $\mu = 0$ this set does reduce to one point, and the theorem is proved for that trajectory. But this value may not be zero. However, we next prove that by modifying V (that is, chosing a V corresponding to a different x^*,) it can be made zero. Once that this is established, the theorem will be proved.

Suppose then that $x(\cdot)$ is a trajectory for which $\mu > 0$, and pick any ω-limit point $\tilde{x}$ of this trajectory, that is to say some point to which a subsequence $x(t_i), t_i \rightarrow \infty$, converges. By (15), $E(\tilde{x}) = 0$. So we can repeat the above argument *using* $\tilde{x}$ *as the new* "x^*". Now necessarily $\mu = 0$, and we are done. ∎

6 Acknowledgments

The authors wish to thank Geoff Hinton for many useful comments and for asking the questions that led to this note, and Ed Blum as well as an anonymous referee, for many useful suggestions.

7 References

Blum, E.K. (1989) Approximation of Boolean functions by sigmoidal networks: Part I: XOR and other two-variable functions. *Neural Computation*, **1**, 532-540.

Brady, M., Raghavan, R., and Slawny, J. (1989) Backpropagation fails to separate where perceptrons succeed. *IEEE Trans. Circuits and Systems*, **36**, 665-674.

Duda, R.O., and Hart, P.E. (1973) *Pattern Classification and Scene Analysis*, New York: Wiley.

Hinton, G.E. (1987) *Connectionist learning procedures* (Technical Report CMU-CS-87-115, Comp.Sci. Dept.), Pittsburg: Carnegie-Mellon University.

Isaacson, E., and Keller, H.B. (1966) *Analysis of Numerical methods*, New York: Wiley.

LaSalle, J.P. (1976) *The Stability of Dynamical Systems*, Philadelphia: SIAM Publications.

Rumelhart, D.E., and McClelland, J.L. (1986) *Parallel Distributed Processing: Explorations in the Microstructure of Cognition, volume* **1**. Cambridge: MIT Press.

Shrivastava, Y., and S. Dasgupta (1987) Convergence issues in perceptron based adaptive neural network models. In *Proc.25th. Allerton Conf. Comm. Contr. and Comp.* (pp. 1133-1141), Urbana: U.of Illinois.

Sontag, E.D. (1988) *Some remarks on the backpropagation algorithm for neural net learning*, (Report SYCON-88-02, Rutgers Center for Systems and Control), New Brunswick: Rutgers University.

Sontag, E.D. (1990) *Mathematical Control Theory: Deterministic Finite Dimensional Systems*, New York: Springer-Verlag.

Sontag, E.D. and H.J. Sussmann (1989) Backpropagation can give rise to spurious local minima even for networks without hidden layers. *Complex Systems*, **3**, 91-106.

Wittner, B.S., and J.S. Denker (1987) Strategies for teaching layered networks classification tasks. In Dana Anderson (Ed.), *Proc. Conf. Neural Info. Proc. Systems* New York: American Institute of Physics.