BACKPROPAGATION SEPARATES WHEN PERCEPTRONS DO

Eduardo D. Sontag
Héctor J. Sussmann
Department of Mathematics, Rutgers University, New Brunswick, NJ 08903
(201)932-3072 (Sontag) - (201)932-5407 (Sussmann) - *E-mail:* sycon@fermat.rutgers.edu

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 of a *threshold* LMS type, in that the error function is zero for values "beyond" the target value.

Our proof gives in addition the following stronger result: the continuous gradient adjustment procedure is such that *from any initial weight configuration* a separating set of weights is obtained *in finite time*. Thus we have a precise analogue of the perceptron learning theorem.

We contrast our results with the more classical pattern recognition problem of threshold LMS with linear output units.

1 Introduction

There has been some interest in understanding the behavior of *backpropagation* (see e.g. [3], [6]) in feedforward nets with no hidden neurons. Although this case can also be approached from the point of view of perceptrons, in the sense that backpropagation techniques do not need to be employed, it does provide a testing ground for hypotheses about the local minima structure of the cost functions that appear in the general case.

In [10] and [1], the authors give examples illustrating the fact that while a training set may be separable, a net performing backpropagation (gradient) search may get stuck in a solution which fails to separate. The first of these papers pointed out that if one uses instead a *threshold* LMS procedure, where one does not penalize values "beyond" the targets, then such counterexamples cease to exist, and in fact that one has a convergence theorem that closely parallels that for perceptrons. The convergence result in [10] however applies only to linear response units, as we discuss later. In independent work in the control-theoretic literature, [7] had obtained related results, which we will also discuss.

In this note, we prove that even for arbitrary (monotonic) nonlinear responses the gradient descent procedure is such that, for separable data, *from any initial weight configuration*, a separating set of weights is obtained *in finite time*. In fact, we provide a result about the convergence of gradient procedures for a very general class of cost functions that includes this and other examples of interest in neural networks.

We also show how to modify the example given in [9] to conclude that there is a training set consisting of 125 binary vectors and a network configuration for which there are nonglobal local minima, even if threshold LMS is used. In this example, the training set is of course *not* separable.

We also compare our results to threshold LMS with linear output units ([2], pp.148-149), and remark that a basic difference with that case is due to the lack of convexity in the cost function in the nonlinear case.

This work was partially supported by NSF grants DMS88-03396 and DMS83-01678-01, by US Air Force grant AFOSR-88-0235, 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.

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. $a \in I \Rightarrow h(a) = 0$
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.

We shall use the standard inner product notation

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

and the norm

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

throughout this note.

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

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

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

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

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

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

$$\{v^1, \ldots, v^l\} \quad \{v^{l+1}, \ldots, v^m\} \tag{2}$$

as well as two real numbers

$$\alpha < 0 < \beta$$

in the range of θ (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 $x^* \in \mathbb{R}^n$ so that

$$\langle v^i, x^* \rangle < 0 \text{ and } \langle v^j, x^* \rangle > 0 \tag{3}$$

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 that case a vector $x^* \in \mathbb{R}^n$ so that

$$\theta(\langle v^i, x^* \rangle) < \alpha \text{ and } \theta(\langle v^j, x^* \rangle) > \beta \tag{4}$$

for each $i = 1, \ldots, l$ and each $j = l+1, \ldots, m$. To each two sets (2) and corresponding numbers α, β we associate the error function E for which

$$h_i(a) := \begin{cases} (\theta(a) - \alpha)^2 & \text{if } \theta(a) > \alpha \\ 0 & \text{if } \theta(a) \leq \alpha \end{cases} \tag{5}$$

if $i = 1, \ldots, l$ and instead

$$h_i(a) := \begin{cases} (\theta(a) - \beta)^2 & \text{if } \theta(a) < \beta \\ 0 & \text{if } \theta(a) \geq \beta \end{cases} \tag{6}$$

for $i = l+1, \ldots, m$. 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 x^* for which $E(x^*) = 0$. □

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

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

in which case one takes α near 0 and β 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

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

instead of the previous penalty functions. Note that separability is *not* in general equivalent to the existence of an x^* so that $E(x^*) = 0$, 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 x^* for which $E(x^*) = 0$. Then, for each x^0 the solution $x(\cdot)$ of the gradient differential equation*

$$\dot{x} = -\nabla E(x)' \tag{7}$$

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 ($E = 0$).

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 α, β (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 $E \to 0$ along trajectories, α is strictly negative, and β 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

$$x_{k+1} := x_k - \rho \nabla E(x_k), \ \ x_0 = x^0 \tag{8}$$

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 2.5 If E is an in example 2.3 then for each initial vector x^0 there exists a real number ρ so that the solution of the iteration (8) is so that x_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

$$\langle v^i, x(t_0) \rangle < 0 \text{ and } \langle v^j, x(t_0) \rangle > 0 \tag{9}$$

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 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 $k \to \infty$ ([4], chapter 8). Any point close enough to $x(t_0)$ still separates, since the inequalities (9) still hold for such a point, so for $\rho = t_0/k$ small it indeed holds that x_k^ρ separates. ∎

Regarding example 2.4, the same conclusions hold *provided* that the target values α, β are selected so that separability of the two sets is equivalent to the existence of an x^* so that $E(x^*) = 0$. This is always true (for any α, β) 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 α, β for which bad local minima may appear.

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

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

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

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

and observe that

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

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

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

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 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 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 (7). 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

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

for all x.

For any initial condition $x(0)$, the trajectory $x(\cdot)$ 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 μ such that

$$x(t) \to E^{-1}(0) \bigcap V^{-1}(\mu) \tag{14}$$

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, chosing a V corresponding to a different x^*,) it can be made zero. If 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 $x(t_i), t_i \to \infty$, converges. By (14), $E(\tilde{x}) = 0$. So we can repeat the above argument *using $\tilde{x}$ as the new "x^*"*. Now necessarily $\mu = 0$, and we are done. ∎

4 Closing 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 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

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

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 x be so that F has a local minimum at x but there exists some y so that

$$F(y) < F(x)\ . \tag{15}$$

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

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

and

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

and consider the cost function E given in example 2.3, where $l = 0$ and α is irrelevant. By continuity, for large enough β it will hold that x is a local minimum of E and that (15) holds for the approximation E, that is, $E(y) < E(x)$. 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, when a nonlinear function θ 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

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

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

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

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, 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 θ=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 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. But the authors did emphasize the fact that threshold LMS is essential in order to avoid the examples where perceptrons classsify 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

1. Brady, M., R. Raghavan and J. Slawny, "Backpropagation fails to separate where perceptrons succeed," submitted for publication. Summarized version in "Gradient descent fails to separate," in *Proc. IEEE International Conference on Neural Networks*, San Diego, California, July 1988, Vol. I, pp.649-656.

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

3. Hinton, G.E., "Connectionist learning procedures," Technical Report CMU-CS-87-115, Comp.Sci. Dept., Carnegie-Mellon University, June 1987.

4. Isaacson, E., and H.B. Keller, *Analysis of Numerical methods*, Wiley, 1966.

5. LaSalle, J.P., *The Stability of Dynamical Systems*, SIAM Publications, Philadelphia, 1976.

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

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

8. Sontag, E.D., "Some remarks on the backpropagation algorithm for neural net learning," Report SYCON-88-02, *Rutgers Center for Systems and Control*, June 1988.

9. Sontag, E.D. and H.J. Sussmann, "Backpropagation can give rise to spurious local minima even for networks without hidden layers," submitted.

10. Wittner, B.S., and J.S. Denker, "Strategies for teaching layered networks classification tasks," in *Proc. Conf. Neural Info. Proc. Systems,* Denver, 1987, Dana Anderson (Ed.), AIP Press.