Back Propagation Fails to Separate Where Perceptrons Succeed

MARTIN L. BRADY, RAGHU RAGHAVAN, MEMBER, IEEE, AND JOSEPH SLAWNY

Abstract—**It is widely believed that the back propagation algorithm in neural networks, for tasks such as pattern classification, overcomes the limitations of the perceptron. We construct several counterexamples to this belief. We also construct linearly separable examples which have a unique minimum which fails to separate two families of vectors, and a simple example with four two-dimensional vectors in a single layer network showing local minima with a large basin of attraction. Thus back propagation is guaranteed to fail in the first, and likely to in the second, example. We show that even multilayered (hidden layer) networks can also fail in this way to classify linearly separable problems. Since our examples are all linearly separable, the perceptron would correctly classify them. Our results disprove the presumption, made in recent years, that, barring local minima, back propagation will find the best set of weights for a given problem.**

I. Introduction

THE *Back Propagation* (*BP*) algorithm has been proposed as a training procedure for neural network models [7]. The importance of BP is that it can be applied to multilayered neural networks, whereas the earlier Perceptron Learning Procedure [6] applies only to one-layer networks. While one-layer networks are limited in the transformations they can perform, multilayered networks are much more general. BP is gradient descent on a function of *least squared errors* (*LSE*), where the errors will be defined shortly, and many problems have been reported for which BP converged to the LSE solution [7]. In fact, it has been suggested that BP is capable of training neural networks to perform arbitrary transformations, and that barring local minima, the BP algorithm finds optimal sets of synaptic interconnection strengths (*weights*) [7], [10].

In this paper, we point out some of the weaknesses of the BP technique. In particular, we consider one of the most rudimentary tasks which a neural network might be used to perform—classifying a set of distinguishable stimuli into two sets. We show that numerous simple classes of separable problems exist for which BP fails to find a separating solution. All of our examples would be correctly classified by the Perceptron Learning Procedure [5].

Since BP is a gradient descent technique, the learning algorithm can get stuck in nearby local minima. It has been observed that networks with hidden layers can have local suboptimal minima in the error function when a nonlinear activation function is used, causing the BP algorithm to fail [7], however this situation has been largely downplayed and ignored. We show that for nonlinear activation functions (e.g., the logistic function), even one-layer networks (no hidden layers) can contain suboptimal local minima. The example given in Section III is extremely simple, consisting of only four input vectors over a one-layer, one-output, two-input network. Furthermore, we show that the region of attraction to the local minima can greatly dominate the region for the global minimum, implying a high probability of failure of the BP algorithm for such instances.

In Section IV, we construct examples where optimal LSE solutions do not minimize the number of misclassifications. For these examples, BP is guaranteed to converge to the wrong set of weights. This is shown to occur even for very restricted problem domains, such as Boolean input vectors and equilength input vectors. Although it has been previously noted that LSE solutions do not necessarily minimize the number of misclassifications [8] the specific results obtained are new and stronger than the previous statements. Such deviations are shown to be widespread and occurring even in very simple problems.

Finally, in Section V we extend our results to hidden layer networks. We construct and analyze an example which has a non-separating local minimum. While hidden layer networks were known to have local minima, our example is linearly separable. In the concluding section we mention relevant previous work in this area.

II. Formulation of the Problem

We now formulate the problem more precisely. We first consider k-input feedforward networks with no hidden units, and with one output unit (see Fig. 1). A pattern p of inputs is characterized by k real numbers, $i_{p,1},\cdots,i_{p,k}$. An activation function f transforms the inputs into an output of the system. The function f is assumed to be differentiable and have a strictly positive first derivative (i.e., *strictly increasing*). Often, a *sigmoid* or *logistic* function, $f(x)=1/(1+e^{-\beta x})$, is used, as a family of differentiable functions approaching the step function as $\beta\to\infty$.

The network is characterized by the weights w_i leading from the ith input unit to the output unit, and by the bias θ, of the output unit. Writing the weights and inputs as $(k+1)$-dimensional vectors, $w=[w_1,\cdots,w_k,\theta]$, $i=[i_1,\cdots,i_k,1]$, the output due to input pattern p is $o_p\equiv$

Manuscript received July 1, 1988; revised November 4, 1988 and December 16, 1988. This work was supported by internal Lockheed funds under IR&D project RDD504 and under IR&D project RDD360. This paper was recommended by Guest Editors R. W. Newcomb and N. El-Leithy.

M. L. Brady and R. Raghavan are with the Lockheed R&DD, 0-9740/B202, Palo Alto, CA 94304.

J. Slawny is with the Center for Transport Theory and Mathematical Physics, Virginia Polytechnic Institute and State University, Blacksburg, VA 24061.

IEEE Log Number 8826717.

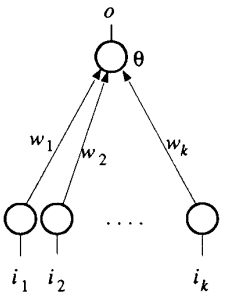


Fig. 1. A k-input one-layer network.

$f(\boldsymbol{w}\cdot\boldsymbol{i}_p)$ where the dot denotes the scalar product. We also introduce the notation $\tilde{\boldsymbol{i}}$ to represent the k-dimensional vector $[i_1,\cdots,i_k]$ corresponding to $\boldsymbol{i}=[i_1,\cdots,i_k,1]$. For each pattern p, the output is compared to its target value, or *teacher*, t_p. The error for the pattern p is defined to be

$$E_p=\frac{1}{2}(o_p-t_p)^2 \tag{1a}$$

and the total error function is

$$E(\boldsymbol{w})=\sum_p E_p. \tag{1b}$$

The BP algorithm implements gradient descent on the function of LSE's. Hence, the algorithm moves along the gradient

$$\nabla E(\boldsymbol{w})=\sum_p\big(f(\boldsymbol{w}\cdot\boldsymbol{i}_p)-t_p\big)f'(\boldsymbol{w}\cdot\boldsymbol{i}_p)\boldsymbol{i}_p \tag{2}$$

of E. We consider only the situation where the teacher values are *attainable*, i.e., for each t_p there exists a real number u_p (unique, by the strict monotonicity of f) such that $f(u_p)=t_p$.

We will denote by I the set of input vectors, i.e., $I=\{\boldsymbol{i}_n,\ n=1,\cdots,p\}$. When we want to emphasize the dependence of the error function on the set of input vectors, I, we write $E_I(\boldsymbol{w})$ instead of $E(\boldsymbol{w})$. Then (1) can be written as

$$E_I(\boldsymbol{w})=\frac{1}{2}\sum_{i\in I}\big(f(\boldsymbol{w}\cdot\boldsymbol{i})-t_i\big)^2. \tag{3}$$

We will continue to use this notation even when some of the inputs coincide.

In our problems, the set of patterns consist of two subsets, and we write $I=I^+\cup I^-$ for the corresponding decomposition of I. Then one says that a weight vector $\boldsymbol{w}$ *separates* I if, for any $\boldsymbol{i}^+$ in I^+ and $\boldsymbol{i}^-$ in I^-,

$$\boldsymbol{w}\cdot\boldsymbol{i}^+>\boldsymbol{w}\cdot\boldsymbol{i}^-.$$

I is *separable* if there exists a weight vector $\boldsymbol{w}$ which separates I. The chief results in this paper are the construction of examples which are separable, but for which the minima of E do not separate. In the examples, the teachers assume two values, $t^+>t^-$, where $t_i=t^+$ (resp. t^-) for all $\boldsymbol{i}$ in I^+ (resp. I^-).

In Section V we extend our results to hidden layer networks. In such a network, one or more intermediate

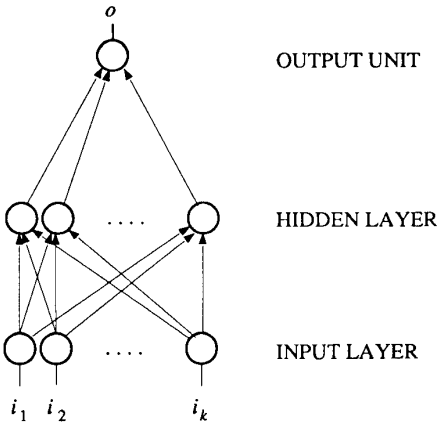


Fig. 2. A two-layer network.

layers of units compute a transformation of the inputs, and their outputs form the input to the next layer. For example, consider a two layer network with k inputs, l hidden units, and one output unit (see Fig. 2). Let $\boldsymbol{w}^j$ denote the $(k+1)$-dimensional vector of weights leading to the jth hidden unit (the last component of $\boldsymbol{w}^j$ is the corresponding bias, θ^j). Then the output of the hidden unit is

$$h_j=f(\boldsymbol{w}^j\cdot\boldsymbol{i}).$$

Defining $\boldsymbol{h}=[h_1,h_2,\cdots,h_l,1]$, we have

$$o_i=f\left(\left(\sum_j w_j^{\text{out}}f(\boldsymbol{w}^j\cdot\boldsymbol{i})\right)+\theta^{\text{out}}\right)=f(\boldsymbol{w}^{\text{out}}\cdot\boldsymbol{h})$$

where $\boldsymbol{w}^{\text{out}}=[w_1^{\text{out}},\cdots,w_l^{\text{out}},\theta^{\text{out}}]$ is the vector of weights leading to the output unit, and the bias, respectively. We also introduce the $((l+1)+l(k+1))$-dimensional vectors:

$$\boldsymbol{w}=[\boldsymbol{w}^{\text{out}},\boldsymbol{w}^1,\cdots,\boldsymbol{w}^l]$$

and

$$\boldsymbol{i}^{\#}=\big[\boldsymbol{h},w_1^{\text{out}}f'(\boldsymbol{w}^1\cdot\boldsymbol{i})\boldsymbol{i},\cdots,w_l^{\text{out}}f'(\boldsymbol{w}^l\cdot\boldsymbol{i})\boldsymbol{i}\big]$$

(note that $\boldsymbol{h}$ depends on $\boldsymbol{i}$). Again, the BP technique trains the network by performing gradient descent on the LSE function

$$E_I(\boldsymbol{w})=\frac{1}{2}\sum_{i\in I}(o_i-t_i)^2$$

and for the gradient of E we have a formula similar to (2):

$$\nabla E_I(\boldsymbol{w})=\sum_{i\in I}(o_i-t_i)f'(\boldsymbol{w}^{\text{out}}\cdot\boldsymbol{h}_i)\boldsymbol{i}^{\#}. \tag{4}$$

All of our examples are formulated according to the following scheme. We start with a separable problem, $I=I^+\cup I^-$, for which E has a unique, easily identifiable separating minimum, $\boldsymbol{w}_0$. We then modify I by adding to it one more input vector, $\boldsymbol{s}$, which we call the *spoiler*. The new set of inputs, $\hat{I}=I\cup\{\boldsymbol{s}\}$, is still separable. However, $E_{\hat{I}}$ will have a minimum close to $\boldsymbol{w}_0$ which does not separate $\hat{I}^-=I^-$ from $\hat{I}^+=I^+\cup\{\boldsymbol{s}\}$.

In the next two sections, we show nonseparating minima of two different types. In Section III, we demonstrate error

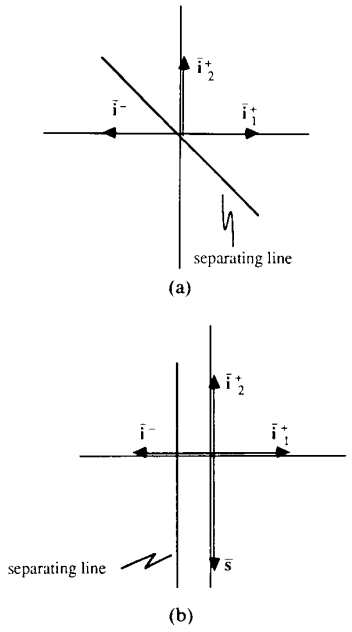


Fig. 3. Illustration of Example 1. (a) Three vector problem. (b) Augmented problem $\hat{I}$.

functions for nonlinear activation functions which have local nonseparating minima in addition to a global minimum. In this case, the effect is due to the nonlinearity of the activation function. We show that even one layer networks can have local minima. In this example the spoiler is a vector longer than those which comprise the zero-error solution.

Section IV shows examples for which the error surface has only one minimum, which is nonseparating. These examples hold for *any strictly increasing monotone transformation*. These examples do not employ a long spoiler vector. Instead, there are many vectors of the same length. All of our examples are linearly separable, and therefore are optimally solved by a one-layer perceptron.

III. Non-Separating Local Minimum

In this section, we present an example for which the error function contains a nonseparating local minimum. The example is extremely simple, containing only four input patterns. This example requires a nonlinear transformation, f, such as the logistic. The precise conditions on f are shown in Appendix I. Our analysis of Example 1 forms the basis of the remaining examples.

3.1. Example 1

The network has two inputs, i.e., the vectors w and i are three dimensional. The set I contains only three vectors, $I^- = \{i^-\}$, $I^+ = \{i_1^+, i_2^+\}$. The two-dimensional diminished versions of I, $\bar{i}^- = [-1, 0]$, $\bar{i}_1^+ = [1, 0]$, $\bar{i}_2^+ = [0, 1]$ are shown in Fig. 3(a). Since i^-, i_1^+, i_2^+ are linearly independent, the system of equations

$$w \cdot i^- = u^-; \quad w \cdot i_1^+ = u^+; \quad w \cdot i_2^+ = u^+$$

has a unique solution, namely,

$$w_0 = \frac{1}{2}[(u^+ - u^-), (u^+ - u^-), (u^+ + u^-)].$$

Here, $u^+, u^- < u^+$ are the arguments of f which yield the teachers: $f(u^+) = t^+, f(u^-) = t^-$.

Consider now an augmented system $\hat{I}$ with $\hat{I}^- = I^-$, $\hat{I}^+ = I^+ \cup \{s\}$, where for the spoiler, s, we have $\bar{s} = [0, -y]$, $y > 0$. The system $\hat{I}$ is obviously separable for any y (see Fig. 3(b)). On the other hand, let us assume that $E_{\hat{I}}$ has a local minimum w_s which is close to w_0 for y large. Then, this local minimum w_s of $E_{\hat{I}}$ fails to separate the separable system $\hat{I}$. We prove the latter as follows. Since

$$2w_0 \cdot s = ((u^- - u^+)y + (u^+ + u^-)) \to -\infty \quad \text{as } y \to +\infty \tag{5}$$

$w_s \cdot s \to -\infty$ as $y \to +\infty$. On the other hand, since for large y, $f(w_s \cdot i^-)$ (resp. $f(w_s \cdot i_i^+)$) is close to t^- (resp. t^+), one has

$$f(w_s \cdot s) < f(w_s \cdot i^-) < f(w_s \cdot i_i^+)$$

for all y large enough. Thus the local minimum fails to classify s correctly.

It remains to show that our assumption on w_s is correct. The proof is contained in the following section, using the result of Lemma 1, contained in Appendix I.

3.2. Proof of the Existence of a Nearby Local Minimum

The existence of the local minimum w_s of $E_{\hat{I}}$ near w_0 can be proved either using a standard fixed point theorem, or more geometrically, which is the approach we will take. First, we show that for the example without the spoiler, we have a nondegenerate minimum. The Hessian of E, $\partial^2 E / \partial w^2$, at w_0 is the matrix M with entries

$$M_{jk} = \sum_{i \in I} (f'(w_0 \cdot i))^2 i_j i_k. \tag{6}$$

Hence, if I contains at least three linearly independent vectors, as is the case, the matrix M is positive definite, and therefore w_0 is an isolated critical point of E_I. (This is also intuitively obvious from (3) since under the above condition on I, the system of equations $f(w \cdot i) = t_i (i \in I)$ has a unique solution, if any at all.)

Let $B(w_0, r)$ be the (closed) ball of radius r centered at w_0; it is now easy to see that there exists $r > 0$ with the properties:

- $\partial^2 E_I / \partial w^2$ is positive definite at all points of $B(w_0, r)$. More precisely, there exists $\lambda > 0$ such that $\partial^2 E_I / \partial w^2 \geq \lambda \mathbf{1}$ for all $w \in B(w_0, r)$;
- $w \cdot s \to -\infty$ as $y \to +\infty$ uniformly in $w \in B(w_0, r)$. More precisely, there is a $\mu > 0$ such that $w \cdot s \leq -\mu y$ for all $w \in B(w_0, r)$ and all y large enough. (This follows from (5).)

Let a^- be the limit of $f(x)$ as $x \to -\infty$. It follows from

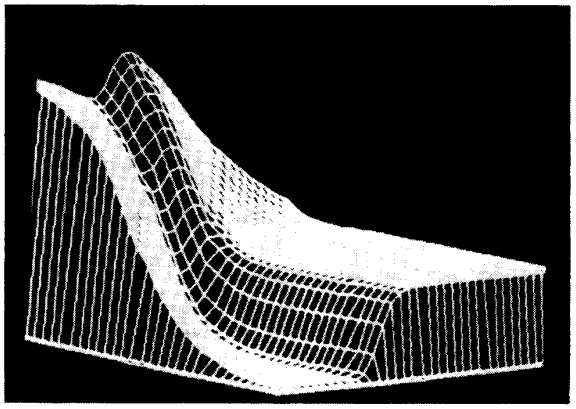


Fig. 4. Energy surface for Example 1.

(2) that as $y \to +\infty$,

$$E_{\hat{f}}(w) \to E_I(w) + \frac{1}{2}(a^- - t^+)^2 \qquad (7)$$

uniformly for $w \in B(w_0, r)$. Moreover, assuming that f' and f'' also have limits at $-\infty$, we obtain that

$$\nabla E_{\hat{f}}(w) \to \nabla E_I(w)$$
$$\partial^2 E_{\hat{f}}(w)/\partial w^2 \to \partial^2 E_I(w)/\partial w^2$$
$$\text{as } y \to +\infty \qquad (8)$$

uniformly for $w \in B(w_0, r)$.

When conditions (7) and (8) are satisfied, then for all y large enough, $E_{\hat{f}}$ has a unique critical point w_s in $B(w_0, r)$. Furthermore, $E_{\hat{f}}(w) > E_{\hat{f}}(w_s)$ for all $w \in B(w_0, r)$, $w \neq w_s$, and $w_s \to w_0$ as $y \to +\infty$, as required. The abstract result justifying these assertions is stated as Lemma 1 in Appendix I. The precise conditions required on f are:

- f is twice continuously differentiable with $f' > 0$;
- f, f', and f'' have finite limits at $+\infty$ and $-\infty$.

Thus the logistic satisfies our conditions, but linear functions do not.

3.3. Simulation Results for Example 1

In this section, we display the energy landscape for Example 1. Our display plots the error in the vertical dimension against the two weights w_1 and w_2. Since we cannot incorporate three parameters into our display, we have modified Example 1 slightly so that both the global minimum as well as the point corresponding to the correct separating solution have the same value for the third parameter, namely the bias. Thus both these points appear in our pictures, and varying the bias around the given value does not change the landscape significantly.

The error surface, $E_{\hat{f}}(w)$, for Example 1 is shown in Fig. 4. The local minimum occurs in the middle of a large flat region, and any initial set of weights in this region converges to a nonseparating solution. For teacher values $t^+ = 0.9$, $t^- = 0.1$, $y = 5$ yields a nonseparating local minimum. In this case, the local minimum is very near the zero error solution: $w_0 = [2.19723, 2.19723, 0]$, while $w_s = [2.19727, 2.18092, 0.00616]$.

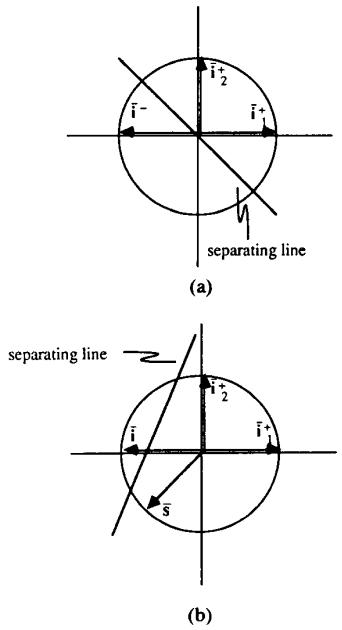


Fig. 5. Normalized input vectors: Example 2. (a) Three-vector problem I. (b) Augmented problem $\hat{I}$.

IV. Nonseparating Global Minimum

In the previous example, the spoiler vector was long. The situation of nonseparating minima cannot, however, be avoided by equilength vectors. In fact, in the next example, in which all vectors are of equal length, the local nonseparating minimum, w_s, is in fact the *global* minimum of $E_{\hat{f}}(w)$. Further, we show that the function $E_{\hat{f}}(w)$ has only one minimum (w_s) for these examples. The situation is thus more dramatic than in the previous, since proper implementation of BP is guaranteed to converge to this nonseparating weight vector. We then show how to extend these ideas to other types of problems. We note that, unlike in Example 1, this result holds for any valid activation function, f, including linear.

4.1. Example 2

All the input vectors for Example 2 will be of the same length. We again use a two input network, so w and i have three components. The set I contains $3n$ input vectors, of three different types, $I^+ = (i_1^+, \cdots, i_{2n}^+)$, $I^- = (i_1^-, \cdots, i_n^-)$, where

$$\bar{i}_1^+ = \bar{i}_3^+ = \cdots = \bar{i}_{2n-1}^+ = [1,0]$$
$$\bar{i}_2^+ = \bar{i}_4^+ = \cdots = \bar{i}_{2n}^+ = [0,1]$$
$$\bar{i}_1^- = \bar{i}_2^- = \cdots = \bar{i}_n^- = [-1,0].$$

The spoiler is

$$\bar{s} = \frac{1}{\sqrt{2}}[-1,-1].$$

This example is illustrated in Fig. 5. There are n copies of each of three vectors, and a single spoiler.

We again pick any $t^- < t^+$ in the range of f and define u^-, u^+ as before. We now note the following.

- As in Example 1, the system of equations $w \cdot i_p = u_p$, for all p, has a solution, namely $w_0 = 1/2[(u^+ - u^-), (u^+ - u^-), (u^+ + u^-)]$.
- E_I has a unique non-degenerate (global) minimum w_0. (In fact, it is not hard to see that w_0 is also a unique *local* minimum of E_I.)
- Weight vector w_0 does not separate $\hat{I}$, where $\hat{I}^+ = I^+ \cup \{s\}$ and $\hat{I}^- = I^-$, since

$$f(w_0 \cdot s) = f\left(\frac{1-\sqrt{2}}{2}u^+ + \frac{1+\sqrt{2}}{2}u^-\right) < f(u^-) = t^-. \tag{9}$$

- Any weight vector separating $\hat{I}$ has least square error of $O(n)$. More precisely, there exists $c > 0$ such that for all w separating $\hat{I}$,

$$E_{\hat{I}}(w) \geq cn. \tag{10}$$

To prove (10), it is enough to show that there exists $c > 0$ such that for any w separating $\hat{I}$,

$$(f(w \cdot i^-) - t^-)^2 + (f(w \cdot i_1^+) - t^+)^2 + (f(w \cdot i_2^+) - t^+)^2 \geq c. \tag{11}$$

Now, the set of w which are separating for $\hat{I}$ is contained in the set W of w for which

$$w \cdot i^- \leq w \cdot i_1^+; \qquad w \cdot i^- \leq w \cdot i_2^+; \qquad w \cdot i^- \leq w \cdot s. \tag{12}$$

By (9), W does not contain w_0. Since W is closed and convex, one can show (see Proposition, Section 4.3) that there exists $\delta > 0$ such that for each $w \in W$, one of the three numbers, $|w \cdot i^- - u^-|$, $|w \cdot i_1^+ - u^+|$, $|w \cdot i_2^+ - u^+|$ is $\geq \delta$. But then (11) follows, since f is strictly increasing.

Thus, there exists a separable example for which no solution of the LSE problem separates. While inequality (10) and a continuity argument are sufficient to show that the global minimum is nonseparating, it can be shown that our example has the following strong property:

For n large enough, $E_{\hat{I}}$ has a unique critical point. This critical point, w_s, tends to w_0 as $n \to \infty$, and $E_{\hat{I}}(w) > E_{\hat{I}}(w_s)$ for $w \neq w_s$ (i.e., w_s is a strict global minimum).

The proof of this is sketched in Appendix II.

4.2. Simulation Results for Example 2

For large enough n, Example 2 has been shown to have a unique, nonseparating minimum. Numerical results indicate that for teachers $t^+ = 0.75$, $t^- = 0.25$, $n = 6$ is sufficient for the result to hold. With teachers $t^+ = 0.9$, $t^- = 0.1$, $n = 15$ will suffice.

A picture of the energy landscape for Example 2 is given in Fig. 6. Our display plots the error in the vertical dimension against the two weights w_1 and w_2. As before, we have modified Example 2 slightly so that both the global minimum as well as the point corresponding to the correct separating solution have the same value for the bias. With n chosen large enough the surface is bowl shaped, with only a single minimum, and this minimum fails to separate the problem.

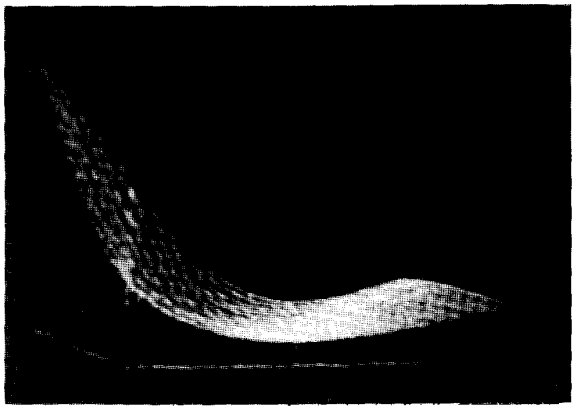
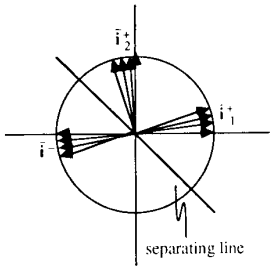
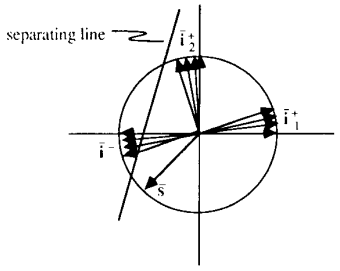


Fig. 6. Energy surface for Example 2.



(a)



(b)

Fig. 7. Normalized input vectors: Example 3.

4.3. Additional Examples with Non-Separating Global Minima

Example 3: (All vectors different, normalized)

This is really just a variation of Example 2, but with all the input vectors different. We "split" the three vectors of I into three groups, each consisting of n vectors which are close to either i^-, i_1^+, or i_2^+. The new situation is illustrated in Fig. 7.

For the sake of concreteness, let ϵ be a positive number and let $i^-_{k,\epsilon}$, $i^+_{2k-1,\epsilon}$ and $i^+_{2k,\epsilon}$, for $k=1,2,\cdots,n$ be the vectors obtained from i^-, i^+_1, and i^+_2, respectively, through a rotation by angle $(k-1)\epsilon$. In particular, $i^-_{1,\epsilon}=i^-$, $i^+_{1,\epsilon}=i^+_1$ and $i^+_{2,\epsilon}=i^+_2$.

We set $I^-(\epsilon)=\{i^-_{k,\epsilon},\ k=1,\cdots,n\}$, $I^+(\epsilon)=\{i^+_{k,\epsilon},\ k=1,\cdots,2n\}$, $\hat{I}^-(\epsilon)=I^-(\epsilon)$, $\hat{I}^+(\epsilon)=I^+(\epsilon)\cup\{s\}$ and $\hat{I}(\epsilon)=\hat{I}^+(\epsilon)\cup\hat{I}^-(\epsilon)$. We note that all the vectors of $\hat{I}(\epsilon)$ are of the same length. While inequality (10) and a continuity argument are sufficient to show that the minima are nonseparating, it can be shown that our example has the following strong property:

For n large enough and ϵ small enough, $E_{\hat{I}}$ has a unique critical point. This critical point, w_s, tends to w_0 as ϵ tends to 0, and $E_{\hat{I}}(w)>E_{\hat{I}}(w_s)$ for $w\neq w_s$.

The proof is sketched in Appendix II.

We will now use a perturbative argument to show that for small enough ϵ, any separating set of weights has error $O(n)$, thus proving that the situation of Example 2 holds for small but nonzero ϵ. We note that the set of separating weight vectors of $\hat{I}(\epsilon)$ is contained in the closed convex set W defined by (12). Weight vector w_0 does not separate $\hat{I}(\epsilon)$; we now show that for all small enough $\epsilon>0$,

$$\inf_{w\in W} E_{\hat{I}(\epsilon)}(w)>cn. \tag{13}$$

Since $E_{\hat{I}(\epsilon)}(w)\to E_{\hat{I}}(w)$ as $\epsilon\to 0$ for each w, our problem is in getting a good enough lower bound on the LHS of inequality (13), which is uniform in n. (While it is presumably true that the LHS of (13) is also a continuous function of ϵ, we have no proof of it, apart from the case when f is linear.) In view of our results about E_I, it is enough to establish the following.

Write i_1,i_2,i_3,t_1,t_2,t_3, for $i^+_1,i^+_2,i^-,t^+,t^+,t^-$, respectively, and for any $j_1,j_2,j_3\in\Re^3$, let

$$F(J)=F(j_1,j_2,j_3)=\inf_{w\in W}\left[\sum_i\left(f(w\cdot j_i)-t_i\right)^2\right].$$

We claim that *there exist $c'>0$, $\delta>0$ such that if* $\operatorname{dist}(j_i,i_i)<\delta$ *for* $i=1,2,3$, *then* $F(J)\geq c'$. That this is true follows from the following:

Proposition: Let $w,j_i\in\Re^k$, $u_i\in\Re$, $i=1,\cdots,n$, and suppose the family $\{j_i,\ i=1,\cdots,n\}$ generates $\Re^k$. Let W be an intersection of a finite number of closed subspaces of $\Re^k$. Then there exists a unique $w(J)$ minimizing

$$\sum_i\left(w\cdot j_i-u_i\right)^2$$

over $w\in W$, and $w(J)$ depends continuously on J (and u).

The uniqueness of w is obtained from standard results on the geometry of convex sets in Euclidean spaces familiar from treatments of least squares and pseudo-inverses. The continuity may be obtained by the use of Lemma 1 in Appendix I. Details of the proof are omitted.

Now, let u_i, for $i=1,2,3$, be such that $f(u_i)=t_i$ and apply the above proposition to (J,u). Since i_1,i_2,i_3 generate $\Re^3$, the assumptions on (J,u) are satisfied for j_i close enough to i_i. Moreover, for J close to I, $w(J)$ will be close to $w(I)$. However, since $w_0\notin W$,

$$\sum_i\left(w(I)\cdot i_i-u_i\right)^2\neq 0.$$

By the proposition, there exists a neighborhood, say U, of I and a constant $c''>0$ such that for all $J\in U$ and all $w\in W$

$$\sum_i\left(w\cdot j_i-u_i\right)^2\geq c''.$$

This implies the existence of the c' and δ of our statement about $F(I)\geq c'$.

We now apply the statement that $F(J)\geq c'$ to $j_1=i^+_{2k-1,\epsilon}$, $j_2=i^+_{2k,\epsilon}$, $j_3=i^-_{k,\epsilon}$ and obtain that inequality (13) holds for ϵ small enough. Thus our assertion is proved.

Example 4: (Boolean vectors)

We show that examples of nonseparating minima can be constructed even for the very restricted case in which all vectors are Boolean. We consider a one layer network having m inputs and one output. Define a problem I as follows. Sets I^+ and I^- each contain 2^{m-2} input vectors, of two different types:

$$\begin{aligned} I^+\colon\ I_{110}&=\{[1,1,0,i_4,\cdots,i_m,1]\}\\ I_{111}&=\{[1,1,1,i_4,\cdots,i_m,1]\}\\ I^-\colon\ I_{100}&=\{[1,0,0,i_4,\cdots,i_m,1]\}\\ I_{010}&=\{[0,1,0,i_4,\cdots,i_n,1]\}\end{aligned}$$

where $i_4,i_5,\cdots,i_m$ take on all $n=2^{m-3}$ $(0,1)$ values: $i_4,i_5,\cdots,i_m\in\{0,1\}$.

The spoiler s is the single vector $s=[0,0,1,0,0,\cdots,0,1]$. This example can be analyzed as a three-input problem because the inputs i_4 through i_m do not affect the character of the solution. They merely allow for 2^{m-3} "copies" of each input vector type to be in fact different. The three-input portions of this example are represented graphically in Fig. 8, along with their respective separating planes.

The separating plane in Fig. 8(a) represents a zero-error solution w_0 for I which does not separate the spoiler vector. Furthermore, it is easy to see that any weight vector w_s which separates $\hat{I}$ has error proportional to n. In order to capture s, w_3 must be greater than zero. Now simply pair the vectors of I^+ which differ only at position 3: $[1,1,0,i_4,\cdots,i_m,1]$ and $[1,1,1,i_4,\cdots,i_m,1]$. Since $w_3>0$, each of the n pairs incurs some minimum error at position 3, independent of n. We omit the detailed proof of these statements.

Example 5: (Higher dimensions)

The previous example has shown that high dimensional examples can be constructed quite easily. We can construct the analog of Example 2 in high dimensions by considering:

$$\begin{aligned} I^+\colon\ I_{1,0}&=\{[1,0,i_3,i_4,\cdots,i_{n+2},1]\}\\ I_{0,1}&=\{[0,1,i_3,i_4,\cdots,i_{n+2},1]\}\\ I^-\colon\ I_{-1,0}&=\{[-1,0,i_3,i_4,\cdots,i_{n+2},1]\}\end{aligned}$$

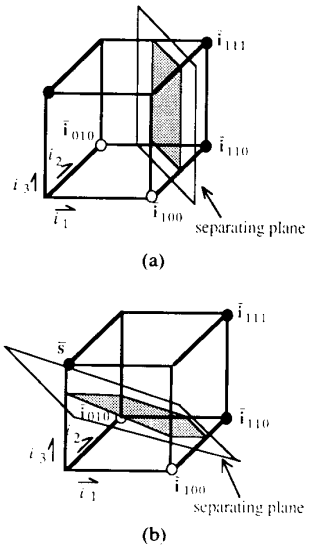


Fig. 8. Boolean vector example.

where $i_k = 1$ and $i_3 = i_4 = \cdots = i_{k-1} = i_{k+1} = \cdots = i_{n+2} = 0$ for $3 < k < n+2$. For example, set $I_{1,0}$ consists of the vectors $[1,0,1,0,0,\cdots,0,1]$, $[1,0,0,1,0,\cdots,0,1]$, $[1,0,0,0,1,\cdots,0,1]$, and so forth. The augmented instance $\hat{I}$ includes spoiler vector $s = [-1/\sqrt{2}, -1/\sqrt{2}, 1, 0, 0, \cdots, 0, 1]$.

The problem I has a set of zero-error solutions w_0 for which $w_1 = w_2 = 1/2(u^+ - u^-) > 0$, and w_0 does not separate $\hat{I}$. To show that $\hat{I}$ separating solutions, w_s, have error proportional to n, consider the n pairs of vectors $[1,0,i_2,i_3,\cdots,i_{n+2},1]$, $[0,1,i_2,i_3,\cdots,i_{n+2},1]$. It is easy to show that for all w_s, $w_1 < w_2$. Once again, each pair incurs some amount of error, independent of n.

V. Example 6: Hidden Layer Networks

In this section, we extend our results to networks with hidden layers. It has been observed that the energy surface for multilayered networks can have local suboptimal minima, even for simple problems (e.g. two-input XOR). However, our example is more dramatic, since we construct a *linearly separable* problem for which a hidden layer network has a local non-separating minimum.

The strategy is as in Example 2 (or 3). First, we construct a family $I = I^+ \cup I^-$ of input vectors so that E_I has a local non-degenerate minimum, w_0, which is separating for I. Next, we define a spoiler, s, in such a way that w_0 does not separate the (separable) family $I \cup \{s\}$. Define the set I_n to consist of each vector of I repeated n times, and $\hat{I} = I_n \cup \{s\}$. By the argument of Example 2, $E_{\hat{I}}$ has a local non-degenerate minimum, w_s, which is close to w_0 for n large enough, and which therefore does not separate the (separable) family $\hat{I}$. *We thus obtain an example with a hidden layer for which the Perceptron convergence procedure would find a separating vector whatever the starting weights, while Back Propagation will fail to do so for some initial conditions.*

The construction will be worked out for a multilayered network with two inputs, two hidden units and one output unit. Adjacent layers of the network are fully interconnected, and the activation function is the logistic function. In the notation of Section II (see (4)) the nine-dimensional weight vector is written as

$$w = [w^{\text{out}}, w^1, w^2]$$

where w^{out} (resp. w^1, w^2) are the three-dimensional weight vectors formed by weights connecting the output unit (resp. the two hidden units) with the units of the level below, and the corresponding biases.

First, we must find a problem which has a non-degenerate zero-error solution. Let F denote the map which to every input i associates its image in the (h_1, h_2)-plane—the plane of outputs of the hidden layer:

$$F(i) = \big(f(w^1 \cdot i), f(w^2 \cdot i)\big).$$

Note that if we choose the vectors $\bar{w}^1$ and $\bar{w}^2$ to be linearly independent, as will be the case below, then for each $\bar{h}$ such that $0 < h_j < 1$, $j = 1,2$, there is exactly one input vector i such that $F(i) = \bar{h}$. Furthermore, for E_I to have a zero-error minimum, the points of $F(I)$ must lie on two parallel lines, one containing $F(I^+)$ and the other $F(I^-)$. If two such parallel lines are chosen, then w^{out} is uniquely defined by t^+ and t^-. These remarks suggest the following scheme for the construction:

- Choose two parallel lines in the (h_1, h_2)-plane. Choose the values $t^{\pm}$ of the teachers and compute w^{out}.
- Pick a family H^+ of points on one of the lines and a family H^- on the other, both families contained in the range of F.
- Choose the vectors $\bar{w}^1$ and $\bar{w}^2$ to be linearly independent, and define $I^+ = F^{-1}(H^+)$ and $I^- = F^{-1}(H^-)$.
- Adjust the values of w^j, $H^{\pm}$ to obtain:
  (a) families $I^{\pm}$ which are linearly separable, and
  (b) a non-degenerate minimum $w_0 = [w^{\text{out}}, w^1, w^2]$ of E_I, i.e., the Hessian of E_I at w_0 is positive definite.
- Choose a spoiler, s, so that $\hat{I}$ is again linearly separable, but w_0 does not separate $I^+ \cup \{s\}$ from I^-.

As this description suggests, one should be able to construct many examples of this kind, if any. We give one example; only the first few significant digits are given.

The separating line in the (h_1, h_2)-plane, the families $H^{\pm}$, and the point corresponding to the spoiler are shown in Fig. 9(b). The teacher values used are $t^+ = 0.75$, $t^- = 0.25$. The families $H^{\pm}$ and the teacher values yield

$$w^{\text{out}} \cong [-20.141, 16.186, 1.976]$$

with the weights leading to the hidden units

$$w^1 = [1.0, 0.1, 0] \quad \text{and} \quad w^2 = [2.0, -0.2, 0].$$

We obtain the zero-error solution w_0 for the family of

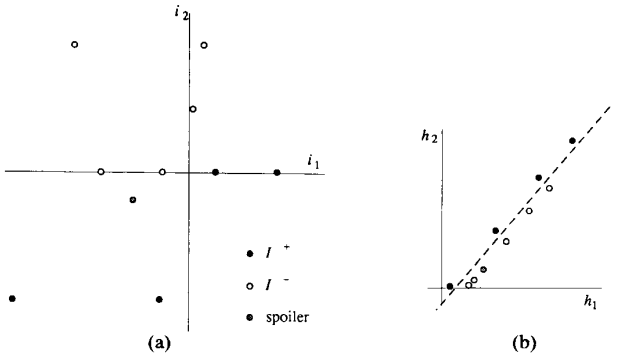


Fig. 9. Nonseparating minimum for the hidden layer network. (a) Input vectors for hidden layer example. (b) Transformed hidden layer space.

inputs $I^{\pm} = F^{-1}(H^{\pm})$:

$$I^+ \cong \{[1.386,0,1],[0.405,0,1],[-0.477,-2,1]\ [-2.783,-2,1]\}$$

$$I^- \cong \{[-1.386,0,1],[-0.405,0,1],[0.477,2,1]\ [-1.807,2,1],[0.069,1,1]\}.$$

These input vectors (and the spoiler) are shown in Fig. 9(a).

It remains to show that the zero-error solution is nondegenerate and to find a spoiler. With the error function

$$E_I(\boldsymbol{w}) = \frac{1}{2}\sum_{i\in I}\left(f(\boldsymbol{w}^{\text{out}}\cdot\boldsymbol{h}(i)) - t_i\right)^2$$

and with (4) for the gradient, we obtain for the Hessian $\partial^2 E_I/\partial \boldsymbol{w}^2$ at $\boldsymbol{w}_0$ the matrix M with entries

$$M_{jk} = \sum_{i\in I}\left(f'(\boldsymbol{w}^{\text{out}}\cdot\boldsymbol{h}(i))\right)^2 i_j^{\#} i_k^{\#}, \qquad j,k = 1,\cdots,9.$$

This is obtained in the same way as (6), taking into account that $o_i - t_i = 0$ at $\boldsymbol{w}_0$. The generalization of this formula to more layers is straightforward.

As in the case of no hidden layers, the matrix M is always positive semi-definite. It is positive definite, and therefore the minimum $\boldsymbol{w}_0$ is non-degenerate, if and only if its determinant, $\det(M)$, is nonzero, or, equivalently, if there are at least nine independent vectors in the family $\{\boldsymbol{i}^{\#}: \boldsymbol{i}\in I\}$. A computer calculation yields a nonzero value for $\det(M)$ with the above nine inputs. Though the numerical results leave no doubt that $\det(M)\neq 0$, to make our computations into a complete proof we would have to follow the roundups made during the computations, and thus turn it into a "computer-assisted proof." We did not do this, however, we stress that given the fact that $\det(M) \neq 0$, the arguments of Example 2 (and 3) do complete the proof.

Now one needs only a spoiler to complete the construction:

$$\boldsymbol{s} = [-1,-0.4,1]$$

does the job (see Fig. 9(a)).

We emphasize that our example is *linearly separable*; a simple one layer, two input Perceptron would be guaranteed to correctly solve this problem. It is generally believed that such a simple problem would easily be solved by a multilayered BP trained network. We thus see that hidden layers are no guarantee for successful classification.

VI. Comments

We have constructed simple examples, where, both for one-layer and hidden layer networks, gradient descent will fail to separate two families of linearly separable vectors.

This work was motivated by the claims made for back propagation in [7], and other publications of the PDP group, that back propagation is suitable for learning to classify arbitrary sets of inputs. After completing this work, we became aware that it has been noted in the past that solutions with LSE are not the same as having a minimum number of misclassifications [8]. It is not stated that LSE solutions can have errors even when fully separating solutions exist. It has also been noted that the linear Widrow–Hoff rule can misclassify cases for which the perceptron would succeed [3]. However, in the latter case, it would seem to have been ascribed to the fact that the rule is applied to one sample pattern at a time. Our results show that such misclassifications are not the result of a particular rule, nor are they confined to cases with complicated energy landscapes.

As mentioned, an earlier version of some of our examples has appeared [2]. The Examples 1 and 2 there were first constructed by Joseph Slawny. (This was inadvertently omitted from the published proceedings.) After he announced the results at the Rutgers meeting on Statistical Mechanics in May 1988, he was informed by S. Solla that his group at the AT&T Bell Laboratories have obtained results similar to Example 2 for *linear f*. In this case, the cost function E is quadratic, and the results can be obtained by purely algebraic methods. Unfortunately, their results were not available to us at the time of writing.

After this paper had been accepted for publication, we learned from Amir Dembo about the work by Wittner and Denker [9]. The example of their Fig. 2 is similar in spirit to our earlier version of Example 2 [2], in which the input vectors were not of equal length. Their ν^1 is playing the role of spoiler. It is, however, suggested in this paper that hidden layer counterexamples will not work.

Appendix I

We formulate a lemma needed in the proof for Example 1 (Section III).

Lemma 1: Let K be an open subset of $\Re^n$ with a compact closure, $\bar{K}$. Suppose that f_n is a sequence of real valued continuous functions on $\bar{K}$ which converge uniformly to a (continuous) function f.

(a) If f attains its minimum at exactly one point of K, say x, and if f_n attains its minimum at x_n, then $x_n \to x$ as $n\to\infty$,

(b) If (i) both f and f_n, $n=1,2,\ldots$ are twice continuously differentiable in K, (ii) x is the only critical point of f in K, (iii) $\partial^2 f/\partial x^2 \geqslant \lambda\ \mathbf{1}$ for some $\lambda > 0$ and all $x\in K$

and (iv) $\partial^2 f_n/\partial x^2 \to \partial^2 f/\partial x^2$ uniformly on K, then for all n large enough, f_n has a unique critical point $x_n \in K$, and $f_n(x) > f_n(x_n)$ for all $x \in K$.

For a proof of (a) we note that due to the compactness of $\bar{K}$, it is enough to consider the case of a sequence x_n, converging to a point $y \in \bar{K}$. Then since $f_n(x_n) < f_n(z)$ for all $z \in K$, passing here to the limit with n and using the fact that $f_n(x_n) \to f(y)$ (by uniform convergence of f_n to f) we obtain that $f(y) \leqslant f(z)$ for all $z \in K$, which by the uniqueness of the minimum of f implies that $y = x$. (b) is rather obvious, given (a); we omit the proof.

Appendix II
Unique Minimum in Example 2

The proof has a perturbative component, of the type of Lemma 1, and a geometric one, which handles the situation "at infinity". The difficulty here is that in any neighborhood of infinity, there are points where the gradient of the perturbation $E_{\hat{I}}$ is larger than the gradient of E_I, and therefore a simple perturbative argument, which is sufficient in a finite region, must be supplemented by a more precise analysis. To make the exposition clearer we start with a geometric interpretation of the problem. Let I and E be as in (3) and assume that all the teachers, t_i are attainable. Let P^i be the plane in the w-space defined by

$$f(w \cdot i) = t_i.$$

Note that i is a vector perpendicular to P^i. Thus as is seen from formula (2) for the gradient, ∇E_i is a vector perpendicular to P^i and always pointing in a direction opposite to P^i. Moreover, $\nabla E_i(w) = 0$ if and only if $w \in P^i$ (also, the length of $\nabla E_i(w)$ tends to zero as it moves away from P^i towards infinity, but this is of no importance here).

Consider now the family of planes P^i, $i \in I$. It partitions the space into a finite number of convex regions, which we shall call *cells*, and their boundaries. Each cell is an intersection of a finite number of half-spaces. Some of the cells are bounded (*finite cells*), while others extend to infinity (*infinite cells*). Fig. 10 presents a two-dimensional version (somewhat misleadingly simple) of the problem. There are three finite and eight infinite cells in this example.

We claim that ∇E *is not zero at the points of the closure of the union of infinite cells*. Moreover, our analysis provides important information on the direction of ∇E in this region. To see this, let C be one of the infinite cells, let $P(C)$ be the family of planes bounding C and let $I(C)$ be the set of corresponding input vectors. Hence for each i one can define $\epsilon_i = \pm 1$ so that

(a) C is defined by the inequalities $C = \{w: w \cdot \epsilon_i i > \epsilon_i u_i$ for all $i \in I(C)\}$;

(b) $w \cdot \epsilon_i i > \epsilon_i u_i$ for any w in C and $i \notin I(C)$.

It then follows from the classical result of Farkas and Minkowski on linear inequalities [1] that for $i \notin I(C)$, $\epsilon_i i$ is a linear combination with *non-negative* coefficients of the vectors $\{\epsilon_i i: i \in I(C)\}$. This has the following geometric interpretation: For each cell C consider the cone $\hat{C}$ generated by the vectors perpendicular to P^i, $i \in I(C)$, and *pointing towards* C. Then for any $i \notin I(C)$ the vector $\pm i$ (the one pointing towards C) is in this cone $\hat{C}$. (One can strengthen this somewhat by considering only those planes which bound C "at infinity".) It can be deduced from this that for any w_0 there exists $R > 0$ such that $\nabla E(w) \neq 0$ for any w with $r(w) = \text{dist}(w, w_0) \geqslant R$. Moreover, for such w let $n(w)$ denote the outer unit vector perpendicular at w to the sphere $S(w_0, r)$ centered at w_0 and with radius $r = r(w)$, and let $e(w)$ be the unit vector $\nabla E(w)/\|\nabla E(w)\|$. Then R can be chosen so large that there exists $\eta > 0$ such that

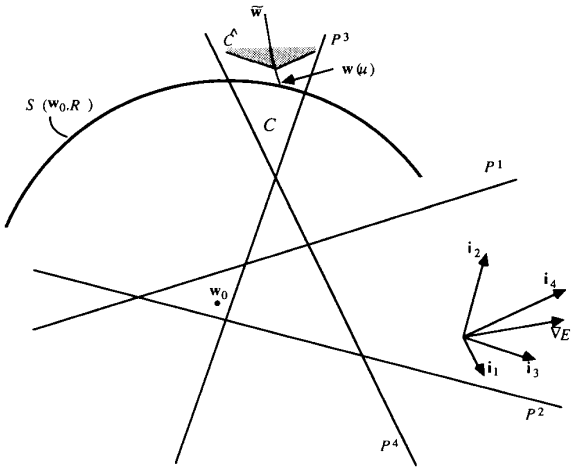


Fig. 10. A two-dimensional version of the geometrical analysis of ∇E. $S(w_0, R)$ is the sphere of radius R centered at w_0.

$$n(w) \cdot e(w) \geqslant \eta \tag{14}$$

for all w with $r(w) \geqslant R$. (If the point w is on one of the planes then consider cells defined by the remaining planes.)

Lemma 2: Assume that there exists w_0 and $R > 0$ such that

(a) All of the finite cells are contained in the (closed) ball $B(w_0, R)$ of radius R centered at w_0.

(b) Inequality (14) holds for $r(w) \geqslant R$.

(c) For any $w \in B(w_0, R)$,

$$E(w) > E(w_0). \tag{15}$$

Then (15) holds for any w, i.e., w_0 is a (unique) global minimum of E.

To prove this, consider any weight vector $\tilde{w}$ with $r(\tilde{w}) > R$. We can assume that there exists $\epsilon > 0$ such that inequality (14) holds for all w with

$$r(w) > R - \epsilon \tag{16}$$

otherwise we would decrease η slightly to achieve this. Let $s \to w(s)$ be the integral curve starting at $\tilde{w}$ of the vector field $w \to -e(w)$ defined in the region (16). Assuming f to be twice continuously differentiable, this integral curve is uniquely defined (until $w(s)$ reaches the boundary of

(16)) and is continuously differentiable. We have, by definition,

$$dw(s)/ds = -e(w(s)), \qquad w(0) = \tilde{w}.$$

Because of (14), the curve reaches $B(w_0, R)$ after a finite "time". For,

$$\frac{d}{ds} r(w(s)) = n(w(s))(-e(w(s))) \leqslant -\eta$$

and therefore $s \geqslant \eta^{-1}(r(\tilde{w}) - r)$, $w(s) \in B(w_0, R)$. Let u be the time of hitting of $B(w_0, R)$, i.e., $w(u) \in B(w_0, R)$ while $w(s) \notin B(w_0, R)$ for $s < u$. Then

$$\begin{aligned} E(\tilde{w}) &= E(w(u)) - \int_0^1 \frac{dE(w(s))}{ds}\, ds \\ &= E(w(u)) + \int_0^1 |\nabla E(w(s))|\, ds \\ &> E(w(u)) > E(w_0). \end{aligned}$$

This proves Lemma 2.

Lemma 2 is directly applicable to our example with multiple vectors, for

$$E_{\hat{I}} = n\left(E_{i^-} + E_{i_1^+} + E_{i_2^+} + \frac{1}{n} E_s\right).$$

Moreover, as is easy to see, E_I has exactly one critical point, w_0, which is also a strict global minimum: $E_I(w) > E_I(w_0)$ for $w \neq w_0$. Consider now the above construction applied to the planes P^{i^-}, $P^{i_1^+}$, $P^{i_2^+}$, P^s. We first choose r large enough so that all the finite cells of this system are contained in $B(w_0, r)$. Then we choose n so large that $E_{\hat{I}}$ has exactly one minimum, w_n, in $B(w_0, 3r)$, a minimum which is close to w_0, and $B(w_n, 2r)$ satisfies the assumptions of Lemma 2. That such n exists, and that for n large enough w_n is the only critical point in $B(w_0, 3r)$ follows directly from Lemma 1, since

$$\frac{1}{n} E_{\hat{I}} \to E_{i^-} + E_{i_1^+} + E_{i_2^+} \left(\equiv \frac{1}{n} E_I\right) \text{ as } n \to \infty.$$

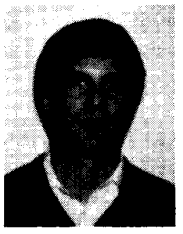
Thus application of Lemma 2 ends the proof in this case. When the vectors are split, as in Example 3, one needs an additional continuity argument, since the number of planes P^i is now changing with n. This is elementary though tedious, and omitted here. Thus for examples 2 and 3, we have, for large n, a unique global non-separating minimum.

Acknowledgment

The authors thank the referees as well as David Smith of the University of Maryland for suggestions that improved the presentation.

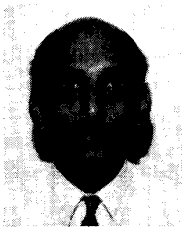
References

[1] E. F. Beckenbach and R. Bellman, *Inequalities*. Berlin, Germany: Springer-Verlag, 1961.
[2] M. Brady, R. Raghavan, and J. Slawny, *Proc. IEEE 2nd Int. Conf. on Neural Networks*, vol. 1, pp. 649–656, 1988.
[3] F. Fogelman Soulie, Y. Robert and M. Tchuente, (Eds.), *Automata Networks in Computer Science*, Princeton, NJ: Princeton Univ. Press, 1987.
[4] G. E. Hinton and T. J. Sejnowski, "Optimal perceptual inference," in *Proc. IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, pp. 448–453, 1983.
[5] M. Minsky and S. Papert, *Perceptrons*, (expanded edition) Cambridge, MA: MIT Press, 1988.
[6] F. Rosenblatt, *Principles of Neurodynamics*. New York: Spartan, 1962.
[7] D. E. Rummelhart and J. L. McClelland (Eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*. Cambridge, MA: MIT Press, 1986, (see esp. vol. 1, ch. 8).
[8] J. Sklansky and G. N. Wassel, *Pattern Classifiers and Trainable Machines*. New York: Springer-Verlag, 1981.
[9] B. Wittner and J. Denker, "Strategies for teaching layered network classification tasks," in *Proc. Amer. Inst. Physics Conf. on Neural Networks*, Denver, CO, pp. 850–859, Nov. 1987.
[10] D. Zipser and R. A. Andersen, "A black-propagation programmed network that simulates response properties of a subset of posterior parietal neurons," *Nature*, vol. 331, no. 25, pp. 679–684, 1988.



Martin L. Brady received the B.S., M.S., and Ph.D. degrees in electrical engineering from the University of Illinois at Urbana-Champaign in 1982, 1984, and 1987, respectively.

He is currently a research scientist at Lockheed's R&D Division in Palo Alto. His research interests include VLSI design algorithms, parallel and distributed computing, computational complexity and neural network computation and learning algorithms.



Raghu Raghavan (M'87) is presently Senior Staff Scientist at Lockheed's R&D Division. His research interests have included condensed matter physics, and more recently, massively parallel architectures, algorithms and pattern recognition.

Joseph Slawny is at the Center for Transport Theory and Mathematical Physics, Virginia Polytechnic and State University. His research interests are in statistical mechanics, dynamical systems and pattern recognition.