Complex Systems **3** (1989) 91–106

Backpropagation Can Give Rise to Spurious Local Minima Even for Networks without Hidden Layers

Eduardo D. Sontag
Héctor J. Sussmann
Department of Mathematics, Rutgers University,
New Brunswick, NJ 08903, USA

Abstract. We give an example of a neural net without hidden layers and with a sigmoid transfer function, together with a training set of binary vectors, for which the sum of the squared errors, regarded as a function of the weights, has a local minimum which is not a global minimum. The example consists of a set of 125 training instances, with four weights and a threshold to be learned. We do not know if substantially smaller binary examples exist.

1. Introduction

Backpropagation (bp) is one of the most widely used techniques for neural net learning. (Cf. Rumelhart and Hinton [3], Hinton [2] for an introduction and references to current work.) The method attempts to obtain interconnection weights that minimize missclassifications in pattern recognition problems.

Though there are many variants of the basic scheme, they are all based on the minimization of a cost function through the use of gradient descent for a particular nonlinear least squares fitting problem. Thus bp is subject to the usual problems associated to local minima, and indeed many experimenters have found training instances in which bp gets stuck in such minima. It is often asserted, however, that even when these minima do occur their domain of attraction is small, or that even if not true minima, the obtained networks tend nonetheless to classify correctly. In addition, it seems to be "folk knowledge" that no spurious local minima can happen when there are no hidden neurons. (The argument made in this last case is roughly that the problem should be analogous to the standard quadratic least squares problem, in which neurons have a linear response map.)

We approach the problem from a purely mathematical point of view, asking what are the constraints that the local minima structure will ultimately impose on any bp-like method. In [4] we remarked that even for the case of no hidden neurons there may be "bad" solutions of the gradient descent algorithm, and this point was also raised by Brady, Raghavan, and Slawny in

the last section of [1]. The main point of the latter reference is to deal with the second of the above assertions. Through a careful and rigorous analysis they show that even if a training set is separable, that is, if it is recognizable by a perceptron, weights obtained through bp may *not* classify the data correctly. (A modification of the cost function, as described in [5] and discussed below, allows one to avoid this problem, however.) In addition, the domain of attraction of such "bad" weight configurations is very large. So neither of the above three assertions is, in fact, correct in general. Of course, it is entirely possible that "real" problems — as opposed to mathematically constructed ones — will not share these pathologies. In that case, it becomes even more urgent to characterize those features of such real problems that are not included in the present formulation.

The constructions of spurious local minima that existed until now did not use binary but rather real-valued inputs. Further, in one case ([1]) the fact that outputs are not allowed to take limiting values ($\{-1,1\}$, or $\{0,1\}$, depending on the conventions,) is critical. Our main result will show that there are indeed counterexamples with binary inputs and outputs, which a situation often encountered in practice.

Precisely, we consider a network with n input neurons and one output neuron. We let $x_1,\ldots,x_n$ be the connection weights, so that the output b computed by the net for an input vector $a=(a_1,\ldots,a_n)$ is given by

$$b=\theta(a_1x_1+a_2x_2+\ldots+a_nx_n). \tag{1.1}$$

Here $\theta:\mathbb{R}\to\mathbb{R}$ is a sigmoid function, i.e. a strictly increasing smooth function such that $\theta(u)$ goes to -1 as $u\to-\infty$ and to 1 as $u\to+\infty$.

Suppose we are given a finite sequence $\mathbf{c}=(c^1,c^2,\ldots,c^m)$ of *input-output pairs* $c^j=(a^j,b^j)$, where the a^j are vectors in $\mathbb{R}^n$, and the b^j are real numbers belonging to the closed interval $[-1,1]$. We then want to choose the weights x_j so as to minimize the error function

$$E(x_1,x_2,\ldots,x_n)=\sum_{j=1}^{m}\left[\theta(a_1^jx_1+a_2^jx_2+\ldots+a_n^jx_n)-b^j\right]^2. \tag{1.2}$$

We will give an example showing that the function E can have local minima that are not global minima. This will show, in particular, that any algorithm for minimizing E which is based on some version of gradient descent may get stuck in a local minimum of E which is not the desired solution. This situation is in marked contrast with the case of Boltzmann machines, where the "folk fact" that, if there are no hidden neurons, then there are no spurious local minima, actually is a true theorem. (Cf., e.g. [6].)

In our example, all the components of the a^j, and all the b^j, will be binary (i.e. equal to 1 or -1). If, as is sometimes the case, one wishes to consider outputs b^j that satisfy $|b^j|<1$ rather than $|b^j|=1$, then it is easy to construct an example for this situation as well, since the property that E has local minima that are not global minima is stable under small perturbations of the function E.

This latter remark is of interest because of the recent result obtained by the authors (see [5]) where it is proved that if (1) one modifies the cost function to be of a *threshold-LMS* type, i.e. one does not penalize "overclassification," and (2) the training data is separable in the sense of perceptrons, then it *is* true indeed that there are no local minima that are not global. Moreover, if (1) and (2) hold, then the gradient descent procedure converges globally, from any initial condition and in finitely many steps, to the minimum. But stability under small perturbations allows us to conclude that even if a threshold-LMS criterion is used, there still will in general exist badly behaved local minima (if (2) does not hold).

In the example we use the sigmoid function $\tanh(u)$, which is up to a simple rescaling the logistic function

$$\frac{1}{1+e^{-u}}, \tag{1.3}$$

which is routinely used when the binary values are taken to be $\{0,1\}$ rather than $\{1,-1\}$. We prefer the latter convention, since the mathematics becomes much more symmetric.

Remark 1.1. As described, our setting does not involve thresholds. However, it is easy to transform our example with n input neurons and no thresholds into an example with $n-1$ input neurons and thresholds. Indeed, it is well known that the presence of a threshold is equivalent to having one neuron who activation is always equal to 1. Since the sigmoid function is odd, we can always change the sign of an input vector, so as to make sure that the first component is 1, and leave the error function unchanged, provided that we also change the sign of the output.

The example given is rather complicated, and it is very possible that a simpler one exists. However, we have been unable so far to simplify our construction.

Intuitively, the existence of local minima is due to the fact that the error function E is the superposition of functions that may have minima at different points. In the more classical case of linear response units, each of these terms is a *convex* function, so no difficulty arises, because a sum of convex functions is again convex. In contrast, sigmoidal units give rise to nonconvex functions, and so there is no guarantee that the sum will have a unique minimum. In order to actually exhibit such an example, it is necessary to obtain terms whose minima are far apart, and to control the second derivatives in such a way that the effect of such minima is not cancelled by the other terms (as happens in the case of convex functions). The calculations are rather involved, and we base them upon a change of variables which converts the function E into a *rational* function. Properties of local minima of rational functions are decidable (theory of real-closed fields), so in principle this change of variables is of some interest besides the role that it plays in the present paper.

2. The example

We let $\theta : \mathbb{R} \to \mathbb{R}$ be the function $\theta(u) = \tanh(u)$, i.e.

$$\theta(u) = \frac{e^u - e^{-u}}{e^u + e^{-u}}. \tag{2.1}$$

We will take $n = 5$. (As indicated above in Remark 1.1, an obvious modification of our example will then yield an example with *four* input neurons plus a threshold.) The inputs will be the following 11 vectors:

$$\begin{aligned}
\mathbf{v}_1 &= (1,1,1,-1,-1),\\
\mathbf{v}_2 &= (1,1,-1,1,-1),\\
\mathbf{v}_3 &= (1,-1,1,-1,1),\\
\mathbf{v}_4 &= (-1,1,1,-1,1),\\
\mathbf{v}_5 &= (-1,1,1,1,-1),\\
\mathbf{v}_6 &= (-1,-1,-1,1,1),\\
\mathbf{v}_7 &= (-1,-1,1,-1,1),\\
\mathbf{v}_8 &= (-1,1,-1,1,-1),\\
\mathbf{v}_9 &= (1,-1,-1,1,-1),\\
\mathbf{v}_{10} &= (1,-1,-1,-1,1),\\
\mathbf{v}_{11} &= (1,1,1,1,1).
\end{aligned}$$

The first five vectors will be repeated 15 times. The sixth to tenth vectors are repeated just once, and the eleventh vector is repeated 45 times. The outputs b^j are always equal to 1.

We will show:

Theorem 1. *With the above choice of inputs and outputs and of the sigmoid function θ, the error function E has local minima that are not global minima.*

3. Proof of Theorem 1

If $x = (x_1, \ldots, x_n)$, $y = (y_1, \ldots, y_n)$, are vectors in $\mathbb{R}^n$, we use $\langle x, y\rangle$ denote the *inner product* of x and y, i.e. $\langle x, y\rangle = \sum_{i=1}^n x_i y_i$. We use I to denote the set $\{1, -1\}$, so I^n is the set of all vectors of length n all whose components are equal to 1 or -1.

If $\mathbf{a} = (a^1, \ldots, a^m)$ is a finite sequence of vectors in I^n, we let $\varphi_{\mathbf{a}} : \mathbb{R}^n \to \mathbb{R}$ be the function given by

$$\varphi_{\mathbf{a}}(x) = \sum_{j=1}^{m} \Big(\theta(\langle a^j, x\rangle) - 1\Big)^2. \tag{3.1}$$

Our goal is to produce an example of a sequence $\mathbf{a}$ such that the function $\varphi_{\mathbf{a}}$ has a local minimum which is not a global minimum.

In order to simplify our calculations, let us rewrite the function $\varphi_{\mathbf{a}}$ in the form

$$\varphi_{\mathbf{a}}(x) = \sum_{a \in I^n} \alpha_a \Big(\theta(\langle a, x\rangle) - 1\Big)^2, \tag{3.2}$$

where the numbers α_a are nonnegative integers. (Precisely, each α_a is the number of times that the vector a occurs in the sequence $\mathbf{a}$.)

It is easy to verify that

$$\Big(\theta(u) - 1\Big)^2 = \frac{4}{(1 + \mathrm{e}^{2u})^2}. \tag{3.3}$$

Let us make the transformation $\mathcal{T}$ given by $\xi_i = \mathrm{e}^{2x_i}$ (so that ξ_i takes values in $\mathbb{R}_+$, the set of positive real numbers), and use ξ to denote a vector $(\xi_1, \ldots, \xi_n) \in \mathbb{R}_+^n$. Write

$$\xi^a = \xi_1^{a_1} \xi_2^{a_2} \ldots \xi_n^{a_n}. \tag{3.4}$$

Then we have $\varphi_{\mathbf{a}}(x) = 4\psi_{\mathbf{a}}(\xi)$, where

$$\psi_{\mathbf{a}} = \sum_{a \in I^n} \frac{\alpha_a}{(1 + \xi^a)^2}. \tag{3.5}$$

So it suffices find an $\mathbf{a}$ such that $\psi_{\mathbf{a}}$ has a local minimum that is not a global minimum.

Now pick, once and for all, a subset $\mathcal{A}$ of I^n such that no vector $a \in I^n$ satisfies $a \in \mathcal{A}$ and $-a \in \mathcal{A}$. Suppose we are given a collection γ of nonnegative integers γ_a for $a \in \mathcal{A}$. Then we can consider the function

$$\psi^{\mathcal{A},\gamma}(\xi) = \sum_{a \in \mathcal{A}} \left(\frac{\gamma_a}{(1 + \xi^a)^2} + \frac{\gamma_{-a}}{(1 + \xi^{-a})^2} \right). \tag{3.6}$$

It is clear that every such function is of the form $\psi_{\mathbf{a}}$ for some appropriate choice of $\mathbf{a}$.

It is convenient to rewrite (3.6) in the simpler form

$$\psi^{\mathcal{A},\gamma}(\xi) = \sum_{a \in \mathcal{A}} \frac{\gamma_a + \gamma_{-a}\xi^{2a}}{(1 + \xi^a)^2}, \tag{3.7}$$

i.e.

$$\psi^{\mathcal{A},\gamma}(\xi) = \sum_{a \in \mathcal{A}} h(\gamma_a, \gamma_{-a}, \xi^a), \tag{3.8}$$

where

$$h(p, q, u) = \frac{p + qu^2}{(1 + u)^2}. \tag{3.9}$$

It will be useful to understand the behavior of the function $h(p, q, \cdot)$ for a particular pair p, q of nonnegative integers. Let us use $'$ to denote differentiation with respect to u. Then an easy computation shows that

$$h'(p,q,u) = \frac{2(qu-p)}{(1+u)^3}. \tag{3.10}$$

Therefore, we have

Lemma 3.1. If $p > 0$ and $q > 0$, then the function $h(p,q,\cdot)$ is globally minimized at $u = \frac{p}{q}$, and the minimum value is $\frac{pq}{p+q}$. Moreover, $h(p,q,\cdot)$ is strictly decreasing for $u < \frac{p}{q}$, and strictly increasing for $u > \frac{p}{q}$. In particular, $u = \frac{p}{q}$ is the only critical point of $h(p,q,\cdot)$.

On the other hand, it is clear that, if p or q (but not both) vanishes, then the function $h(p,q,\cdot)$ does not have a minimum. Lemma 3.1 shows that one can place this minimum at any rational point $\bar{u}$ of $\mathbb{R}_+$ by suitably choosing p and q.

If we now choose an $a \in I^n$ and positive integers p, q, and consider the function $\xi \to h(p,q,\xi^a)$, we see that this function is *globally minimized* at all points ξ in the set $S(a,p,q)$ of those $\xi \in \mathbb{R}^n_+$ such that $\xi^a = \frac{pq}{p+q}$.

Now suppose we choose n linearly independent vectors $a^1,\ldots,a^n$ in I^n, and positive numbers p_i, q_i, $i = 1,\ldots,n$. Then the sets $S(a^i,p_i,q_i)$ intersect at exactly one point. (To see this, just notice that, under the transformation $\mathcal{T}$, the set $S(a,p,q)$ corresponds to the hyperplane $\langle a,x\rangle = \frac{1}{2}\log\left(\frac{pq}{p+q}\right)$.) This point is then the global minimum of the function. So we have established:

Lemma 3.2. Let $a^1,\ldots,a^n$ be linearly independent members of I^n, and let $p_1,\ldots,p_n$, $q_1,\ldots,q_n$ be positive numbers. Then the function Ψ given by

$$\Psi(\xi) = \sum_{i=1}^{n} \frac{p_i + q_i\xi^{2a^i}}{(1+\xi^{a^i})^2} \tag{3.11}$$

has a unique global minimum at the point $\bar{\xi}_\Psi$ characterized by $(\bar{\xi}_\Psi)^{a^i} = \frac{p_i}{q_i}$ for $i = 1,2,\ldots,n$, and the value of $\Psi(\bar{\xi}_\Psi)$ is equal to v_Ψ, where

$$v_\Psi = \frac{p_1q_1}{p_1+q_1} + \frac{p_2q_2}{p_2+q_2} + \ldots + \frac{p_nq_n}{p_n+q_n}. \tag{3.12}$$

We now specialize even further, and choose $n = 5$. Moreover, we choose the five vectors a^i to be such that three of their components are equal to 1, and the other two equal to -1. For instance, we can choose a^1, a^2, a^3, a^4, a^5 to be, respectively, the vectors $(1,1,1,-1,-1)$, $(1,1,-1,1,-1)$, $(1,-1,1,-1,1)$, $(-1,1,1,-1,1)$ and $(-1,1,1,1,-1)$. *From now on it will be assumed that $n = 5$ and the a^i are the five vectors listed above.* Finally, we choose all the p_i to be equal to a number p, and all the q_i equal to 1. Then it is clear that the point $\tilde{\xi}_p = (p,p,p,p,p)$ is none other than $\bar{\xi}_\Psi$. So we have:

Lemma 3.3. The function Ψ has a unique global minimum at $\tilde{\xi}_p$, and its value there is equal to $\frac{5p}{p+1}$.

Notice that this value is very close to 5 if p is very large. On the other hand, if $\tilde{\xi}_1$ denotes the point $(1,1,1,1,1)$ (i.e. the point of $\mathbb{R}^5_+$ that corresponds to the origin under the transformation $\mathcal{T}$), then $\Psi(\tilde{\xi}_1) = \frac{5}{4}(p+1)$. Since $\frac{p}{p+1} \leq \frac{p+1}{4}$, with equality holding only if $p = 1$, we see that $\Psi(\tilde{\xi}_p) < \Psi(\tilde{\xi}_1)$ unless $p = 1$, in which case the points $\tilde{\xi}_p$ and $\tilde{\xi}_1$ coincide. Moreover, when p is very large the value v_Ψ is approximately equal to 5, whereas $\Psi(\tilde{\xi}_1) = \frac{5}{4}(p+1)$.

We now let $\hat{\Psi}$ be the function

$$\hat{\Psi}(\xi) = h(\hat{p}, \hat{q}, \xi^{\hat{a}}), \tag{3.13}$$

where the vector $\hat{a} \in I^n$ and the numbers $\hat{p}$, $\hat{q}$ are chosen so that the minima of $\hat{\Psi}$ (i.e. the points in the set $S(\hat{a}, \hat{p}, \hat{q})$) are very far from $\tilde{\xi}_p$. This can be achieved by taking $\hat{a} = (1,1,1,1,1)$ to begin with, so that the value of $\hat{\Psi}$ at $\tilde{\xi}_p$ is $\frac{\hat{p}+\hat{q}p^{10}}{(1+p^5)^2}$, whereas the value at $\tilde{\xi}_1$ is $\frac{\hat{p}+\hat{q}}{4}$. Notice that the condition on $\hat{p}$, $\hat{q}$ that would make $\tilde{\xi}_p$ belong to $S(\hat{a}, \hat{p}, \hat{q})$ would be $\hat{p} = p^5\hat{q}$, so in particular $\hat{p}$ would have to be much larger than $\hat{q}$, since we are going to choose p large. We will choose $\hat{p}$, $\hat{q}$ so that we are very far from this situation, by taking $\hat{q}$ much larger than $\hat{p}$. Actually, to make matters even easier, we will take $\hat{p}$ to be 0, and choose $\hat{q}$ "sufficiently large." (Precisely how large will be seen below.)

We now let $\Psi^* = \Psi + \hat{\Psi}$. In particular,

$$\Psi^*(\tilde{\xi}_p) = \frac{5p}{1+p} + \frac{\hat{q}p^{10}}{(1+p^5)^2}, \tag{3.14}$$

and

$$\Psi^*(\tilde{\xi}_1) = \frac{5p}{4} + \frac{\hat{q}}{4} + \frac{5}{4}. \tag{3.15}$$

It is then clear that $\Psi^*(\tilde{\xi}_p) > \Psi^*(\tilde{\xi}_1)$ if p and $\hat{q}$ are large enough. For instance, one can easily verify that

Lemma 3.4. If $p \geq 2$ and $\hat{q} \geq 2p$ then $\Psi^*(\tilde{\xi}_p) > \Psi^*(\tilde{\xi}_1)$.

Now let us use $B(\rho, p)$ to denote the closed ball with center $\tilde{\xi}_p$ and radius ρ, provided that $0 < \rho < \sqrt{5}p$, so that $B(\rho, p) \subseteq \mathbb{R}^5_+$. Then it is clear that the inequality

$$\Psi^*(\xi) > \Psi^*(\tilde{\xi}_1) \tag{3.16}$$

will hold for $\xi \in B(\rho, p)$ if ρ is sufficiently small, In particular, this will imply that, if the function has a local minimum in the interior of $B(\rho, p)$, then this is not a global minimum.

We need to know how ρ can be chosen so that 3.16 will hold on $B(\rho, p)$. We have

$$\begin{aligned}
\Psi^*(\xi) &= \Psi(\xi) + \hat{\Psi}(\xi) && (3.17)\\
&\geq \Psi(\tilde{\xi}_p) + \hat{\Psi}(\xi) && (3.18)\\
&= \frac{5p}{p+1} + \hat{\Psi}(\xi) && (3.19)\\
&= \frac{5p}{p+1} + \hat{q}\frac{\xi^{2\hat{a}}}{(1+\xi^{\hat{a}})^2} && (3.20)\\
&= \frac{5p}{p+1} + \hat{q}\eta(\xi^{\hat{a}})^2, && (3.21)
\end{aligned}$$

where

$$\eta(u) = \frac{u}{1+u} = 1 - \frac{1}{1+u}. \tag{3.22}$$

Clearly, η is an increasing function of u for $u \in \mathbb{R}_+$. On the other hand, the function $\xi^{\hat{a}} = \xi_1\xi_2\xi_3\xi_4\xi_5$ is bounded below on $B(\rho, p)$ by $(p-\rho)^5$, so the lower bound

$$\Psi^*(\xi) \geq \frac{5p}{p+1} + \hat{q}\left[1 - \frac{1}{1+(p-\rho)^5}\right]^2 \tag{3.23}$$

holds throughout $B(\rho, p)$. Suppose we choose ρ, p such that $p \geq 2$ and $4\rho < p$. Then $\frac{5p}{p+1} > \frac{1}{4}$. Also, $(p-\rho)^5 > 7$, and so $\Psi^*(\xi) \geq \frac{1}{4} + \frac{49\hat{q}}{64}$ throughout $B(\rho, p)$. If we choose $\hat{q} = kp$, then 3.16 will hold for $\xi \in B(\rho, p)$ as long as $k \geq 3$. So we have shown

Lemma 3.5. If $p \geq 2$, $\hat{q} \geq 3p$, $0 < \rho < \frac{p}{4}$, then Inequality 3.16 holds throughout the ball $B(\rho, p)$.

We now have to show that, by suitably choosing p, ρ and $\hat{q}$, we can satisfy the hypotheses of the previous lemma and also guarantee that Ψ^* will have a local minimum in the interior of $B(\rho, p)$. The crucial point here is that Ψ has a minimum at $\tilde{\xi}_p$. The addition of $\hat{\Psi}$ should not disturb this fact too much, because $\hat{\Psi}$ is nearly constant in the neighborhood of $\tilde{\xi}_p$, and the Hessian matrix of Ψ at $\tilde{\xi}_p$ is nondegenerate. To make this precise, we need to have an upper bound on the gradient of $\hat{\Psi}$ and a lower bound on the second derivative of Ψ. Indeed, once we have established those bounds, the following lemma gives us the desired result. In the statement, $\|\ldots\|$ denotes the usual Euclidean norm, and $D_v f$, $D_v^2 f$ denote, respectively, the first and second directional derivatives of the function f in the direction of the unit vector $v \in \mathbb{R}^n$.

Lemma 3.6. Let f, g be C^2 functions on a closed ball $B \subseteq \mathbb{R}^m$ of radius r centered at a point $\bar{x} \in \mathbb{R}^m$. Assume that A, C are constants such that $\|\nabla g(x)\| \leq A$ for all $x \in B$, and $D_v^2 f(x) \geq C$ for all $x \in B$ and all unit vectors $v \in \mathbb{R}^m$. Then, if $\nabla f(\bar{x}) = 0$, and $Cr > 2A$, it follows that the function $f + g$ has a local minimum in the interior of B.

Proof. Let $F = f + g$. Let S be the boundary of B. We show that $F(x) > F(\bar{x})$ for all $x \in S$. The bound on ∇g implies that $|g(x) - g(\bar{x})| \leq Ar$ for $x \in S$, so $g(x) \geq g(0) - Ar$ for all such S. If $x \in S$, write $x = \bar{x} + rv$ with v a unit vector. Let $\tilde{f}(t) = f(\bar{x} + tv)$ for $0 \leq t \leq r$. Then the derivative $\tilde{f}'(t)$ vanishes at $t = 0$ and has a derivative bounded below by C. So $\tilde{f}'(t) \geq Ct$ for $0 \leq t \leq r$. But then $\tilde{f}(r) \geq \tilde{f}(0) + \frac{Cr^2}{2}$, i.e. $f(x) \geq f(\bar{x}) + \frac{Cr^2}{2}$. We then get $F(x) \geq F(\bar{x}) + \frac{Cr^2}{2} - Ar$. Since $Cr > 2A$, we have shown that $F(x) > F(0)$ for all $x \in S$. ∎

Lemma 3.6 enables us to get an *a priori* idea on how large one has to take r. Suppose we compute $D_v^2 f(\bar{x})$ for all v, and $\nabla g(\bar{x})$, and we find that $\bar{C} = \inf\{D_v^2 f(\bar{x}) : \|v\| = 1\}$, $\bar{A} = \|\nabla g(\bar{x})\|$. Then it is clear that, no matter how we choose r, the constants A and C are going to satisfy $C \leq \bar{C}$, $A \geq \bar{A}$, so the *smallest* r can possibly be is $r = \frac{2\bar{A}}{\bar{C}}$. In the case of interest to us, the functions f, g and the point $\bar{x}$ depend on the parameter p, and the numbers $\bar{A}$, $\bar{C}$ behave like p^{-5} and p^{-3}, respectively. So r should be chosen so that $r \sim p^{-2}$.

We want to apply Lemma 3.6 with $n = 5$, $\bar{x} = \tilde{\xi}_p$, $f = \Psi$, $g = \hat{\Psi}$ and $r = \rho$. The hypothesis that $\nabla f(\bar{x}) = 0$ holds since f has a minimum at $\bar{x}$. Naturally, the bounds on ∇g and D_v^2 hold for *some* choice of the constants A, C, with C not necessarily positive. We have to show that, by suitably choosing p, $\hat{q}$ and ρ, we can satisfy the condition $Cr > 2A$. (This will imply in particular that $C > 0$.) Obviously, the condition that $C > 0$ is related to the fact that the vectors a^i are linearly independent so that, in each direction, at least one of the five functions whose sum is Ψ is strictly convex near $\tilde{\xi}_p$. To make this precise, we must study the quadratic form Q given by

$$Q(v) = \sum_{i=1}^{5} \langle a^i, v \rangle^2 \text{ for } v \in \mathbb{R}^5. \tag{3.24}$$

Then Q is clearly nonnegative. Since the a^i form a basis of $\mathbb{R}^5$, $Q(v)$ can never vanish unless $v = 0$. So there exists a constant $c > 0$ such that $Q(v) \geq c\|v\|^2$ for all v. It will be useful to know an explicit value of c. A crude estimate, which will be sufficient for our purposes, is the bound (cf. Appendix A):

$$Q(v) \geq \frac{1}{3}\|v\|^2 \text{ for } v \in \mathbb{R}^5. \tag{3.25}$$

We now get a lower bound for the second derivative of Ψ on some neighborhood of $\tilde{\xi}_p$. An elementary calculation gives the formula (cf. Appendix B):

$$D_v^2 \Psi(\xi) = U(p, v, \xi) - V(p, v, \xi), \tag{3.26}$$

where

$$U(p, v, \xi) = 2 \sum_{j=1}^{5} \left\langle \frac{v}{\xi}, a^j \right\rangle^2 \cdot \left[\frac{2(p+1)\xi^{2a^j} - \xi^{3a^j} - p\xi^{a^j}}{(1 + \xi^{a^j})^4} \right], \tag{3.27}$$

and

$$V(p,v,\xi) = 2\sum_{j=1}^{5}\Big\langle \frac{v^2}{\xi^2}, a^j\Big\rangle \cdot \left[\frac{\xi^{2a^j} - p\xi^{a^j}}{(1+\xi^{a^j})^3}\right], \tag{3.28}$$

where $\frac{v}{\xi}$ denotes the vector

$$\frac{v}{\xi} = \Big(\frac{v_1}{\xi_1}, \frac{v_2}{\xi_2}, \ldots, \frac{v_5}{\xi_5}\Big), \tag{3.29}$$

and $\frac{v^2}{\xi^2}$ is defined similarly.

Now assume that v is a unit vector. We will get a lower bound for $D_v^2\Psi(\xi)$ on a ball $B(\rho,p)$ by getting a lower bound for U and an upper bound for the absolute value of V. Notice that, near $\tilde{\xi}_p$, ξ^{a^j} is approximately equal to p, and all the components of ξ are also approximately equal to p. So, if p is large enough so that we can ignore the "1" in $1+p$, then $U(p,v,\xi)$ is approximately equal to $2Q(v)p^{-3}$. So we have an approximate bound $U(p,v,\xi) \geq \frac{2}{3}p^{-3}$. On the other hand, $V(p,v,\xi)$ is a difference of two expressions, each of which is bounded in absolute value by a constant times p^{-3}. However, these two expressions are equal at $\tilde{\xi}_p$, which means in particular that the leading powers of p cancel, and $V(p,v,\xi)$ is actually $O(p^{-4})$ for ξ near $\tilde{\xi}_p$. To make all this precise, notice that on $B(\rho,p)$ we have the bounds

$$\frac{(p-\rho)^3}{(p+\rho)^2} \leq \xi^{a^j} \leq \frac{(p+\rho)^3}{(p-\rho)^2}. \tag{3.30}$$

Using this, we easily get:

$$2(p+1)\xi^{2a^j} - \xi^{3a^j} - p\xi^{a^j} \geq 2\frac{(p+1)(p-\rho)^6}{(p+\rho)^4} - \frac{(p+\rho)^9}{(p-\rho)^6} - p\frac{(p+\rho)^3}{(p-\rho)^2} \tag{3.31}$$

for $\xi \in B(\rho,p)$. So, if we write $u = \frac{\rho}{p}$, we have

$$2(p+1)\xi^{2a^j} - \xi^{3a^j} - p\xi^{a^j} \geq p^3\lambda_1(u) \tag{3.32}$$

where

$$\lambda_1(u) = 2\frac{(1+\frac{u}{\rho})(1-u)^6}{(1+u)^4} - \frac{(1+u)^9}{(1-u)^6} - \frac{u(1+u)^3}{\rho(1-u)^2}. \tag{3.33}$$

Also,

$$(1+\xi^{a^j})^4 \leq \left(1 + \frac{(p+\rho)^3}{(p-\rho)^2}\right)^4, \tag{3.34}$$

so that

$$(1+\xi^{a^j})^4 \leq p^4\lambda_2(u), \tag{3.35}$$

where

$$\lambda_2(u) = \left[\frac{u}{\rho} + \frac{(1+u)^3}{(1-u)^2}\right]^4. \tag{3.36}$$

We therefore have the bound

$$\frac{2(p+1)\xi^{2a^j} - \xi^{3a^j} - p\xi^{a^j}}{(1+\xi^{a^j})^4} \geq \frac{1}{p}\lambda\Big(\frac{\rho}{p}\Big), \tag{3.37}$$

where the function λ is given by

$$\lambda(u) = \frac{\lambda_1(u)}{\lambda_2(u)}. \tag{3.38}$$

Notice that $\lambda(0) = 1$, so the bound (3.37) says that the left-hand side of (3.37) is approximately bounded below by p^{-1} if $\frac{\rho}{p}$ is small. Then

$$U(p, v, \xi) \geq \frac{2}{p}\lambda\Big(\frac{1}{p}\Big)Q\Big(\frac{v}{\xi}\Big) \tag{3.39}$$

if $\xi \in B(\rho, p)$. In view of the lower bound for Q, we have

$$Q\Big(\frac{v}{\xi}\Big) \geq \frac{1}{3}\Big\|\frac{v}{\xi}\Big\|^2, \tag{3.40}$$

so that

$$Q\Big(\frac{v}{\xi}\Big) \geq \frac{1}{3(p-\rho)^2}. \tag{3.41}$$

Therefore

$$U(p, v, \xi) \geq \frac{\Lambda(\frac{\rho}{p})}{p^3}, \tag{3.42}$$

for $\xi \in B(\rho, p)$, where

$$\Lambda(u) = \frac{2\lambda(u)}{3(1-u)^2}. \tag{3.43}$$

We now get an upper bound for $|V(p, v, \xi)|$ on $B(\rho, p)$. Clearly,

$$\left|\Big\langle \frac{v^2}{\xi^2}, a^j \Big\rangle\right| \leq \frac{\|v\|^2}{(p-\rho)^2}. \tag{3.44}$$

Also, we can write

$$\begin{aligned}
\left|\frac{\xi^{2a^j} - p\xi^{a^j}}{(1+\xi^{a^j})^3}\right| &= \xi^{a^j}\left|\frac{\xi^{a^j} - p}{(1+\xi^{a^j})^3}\right| && (3.45)\\
&\leq \frac{|\xi^{a^j} - p|}{(1+\xi^{a^j})^2} && (3.46)\\
&\leq \frac{|\xi^{a^j} - p|}{\xi^{2a^j}} && (3.47)\\
&\leq \frac{(p+\rho)^4}{(p-\rho)^6}|\xi^{a^j} - p| && (3.48)\\
&= \frac{1}{p^2}\nu_1\Big(\frac{\rho}{p}\Big)|\xi^{a^j} - p|, && (3.49)
\end{aligned}$$

where

$$\nu_1(u) = \frac{(1+u)^4}{(1-u)^6}. \tag{3.50}$$

On the other hand, the inequality

$$\frac{(p-\rho)^3}{(p+\rho)^2} \leq \xi^{a^j} \leq \frac{(p+\rho)^3}{(p-\rho)^2} \tag{3.51}$$

yields

$$\frac{(p-\rho)^3}{(p+\rho)^2} - p \leq \xi^{a^j} - p \leq \frac{(p+\rho)^3}{(p-\rho)^2} - p \tag{3.52}$$

so that

$$|\xi^{a^j} - p| \leq p\nu_2\Big(\frac{\rho}{p}\Big), \tag{3.53}$$

where

$$\nu_2(u) = \max\left(\frac{(1+u)^3}{(1-u)^2} - 1, 1 - \frac{(1-u)^3}{(1+u)^2}\right). \tag{3.54}$$

Combining all these bounds we get

$$|V(p,v,\xi)| \leq \frac{\nu(\frac{\rho}{p})}{p^3}, \tag{3.55}$$

where

$$\nu(u) = \frac{2\nu_1(u)\nu_2(u)}{(1-u)^2}. \tag{3.56}$$

Finally, we get

$$D_v^2\Psi(\xi) \geq \frac{\Lambda(\frac{\rho}{p}) - \nu(\frac{\rho}{p})}{p^3}, \tag{3.57}$$

as long as $\xi \in B(\rho, p)$.

Since $\Lambda(0) = \frac{2}{3} > 0$ and $\nu(0) = 0$, the lower bound for $D_v^2\Psi(\xi)$ given by the preceding formula is positive if $\frac{\rho}{p}$ is small enough.

Next we need an upper bound for $\nabla\hat{\Psi}$. We have (cf. Appendix B):

$$\frac{\partial\hat{\Psi}}{\partial\xi_i} = \frac{2\hat{q}\xi^{2\hat{a}}}{\xi_i(1+\xi^{\hat{a}})^3}, \tag{3.58}$$

so that

$$\left|\frac{\partial\hat{\Psi}}{\partial\xi_i}\right| \leq \frac{2\hat{q}}{\xi_i}\xi^{\hat{a}}. \tag{3.59}$$

In particular, if $\xi \in B(\rho, p)$, and $\hat{q} = kp$, we have the bound

$$\|\nabla\hat{\Psi}(\xi)\| \leq \frac{2k}{p^5}\mu\Big(\frac{\rho}{p}\Big), \tag{3.60}$$

where

$$\mu(u) = \frac{1}{(1-u)^6}. \tag{3.61}$$

If we let

$$C = \frac{\Lambda(\frac{\rho}{p}) - \nu(\frac{\rho}{p})}{p^3}, \tag{3.62}$$

$$A = \frac{2k}{p^5}\mu\Big(\frac{\rho}{p}\Big), \tag{3.63}$$

and

$$K = \frac{C\rho}{2A}, \tag{3.64}$$

then the hypothesis of Lemma 3.6 will be satisfied if $K > 1$. On the other hand,

$$K = \frac{p^2\rho\Big[\Lambda(\frac{\rho}{p}) - \nu(\frac{\rho}{p})\Big]}{2k\mu\Big(\frac{\rho}{p}\Big)}. \tag{3.65}$$

Since $\Lambda(0) = \frac{2}{3}$, $\nu(0) = 0$, $\mu(0) = 1$, it is clear that, for any fixed ρ and k, the inequality $K > 1$ will hold if p is sufficiently large. Moreover, if one chooses $k = 3$, then the conditions of 3.5 will also hold if p is sufficiently large.

To prove Theorem 1, all we need to do is to verify that the conditions of 3.5 as well as the inequality $K > 1$ hold for $k = 3$ for some choice of ρ, not just for p sufficiently large, but for $p = 15$. So all we need is to find a value of ρ such that $4\rho < 15$, with the property that, if we plug in $p = 15$ and $k = 3$ in Equation 3.65, then $K > 1$. For each ρ, it is clear that there is a smallest p such that $K > 1$. Let this p be denoted by $p(\rho)$. Then a direct computation shows that, for ρ in the range between 0.08 and 0.14, the value of $p(\rho)$ is equal to 15. (As a function of ρ, $p(\rho)$ decreases for $\rho < 0.08$ and increases again for $\rho > 0.14$, so $p = 15$ is the best value that can be obtained from our estimates.) This completes the proof of Theorem 1. ∎

Acknowledgments

Sussmann's work was partially supported by NSF grant DMS83-01678-01, 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.

Sontag's work was partially supported by NSF grant DMS88-03396, by U.S. Air Force grants AFOSR-85-0247 and 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.

Appendix A. Derivation of Formula (3.25)

The bound we seek is the smallest singular value of A (where A is the matrix whose rows are the vectors $a^1, \ldots, a^5$), i.e. the smallest eigenvalue of $A^\dagger A$. It can be computed numerically, and turns out to be approximately 1/2.52. The following simple argument gives the bound used in the text.

Set $u_j = \langle a^j, v \rangle$. Using the explicit formulas for the a^j we find that

$$u_1 = v_1 + v_2 + v_3 - v_4 - v_5, \tag{A.1}$$

$$u_2 = v_1 + v_2 + v_4 - v_3 - v_5, \tag{A.2}$$

$$u_3 = v_1 + v_3 + v_5 - v_2 - v_4, \tag{A.3}$$

$$u_4 = v_2 + v_3 + v_5 - v_1 - v_4, \tag{A.4}$$

and

$$u_5 = v_2 + v_3 + v_4 - v_1 - v_5. \tag{A.5}$$

This system can easily be inverted, resulting in the following expressions for the v's in terms of the u's:

$$2v_1 = u_2 + u_3, \tag{A.6}$$

$$2v_2 = u_2 + u_4, \tag{A.7}$$

$$2v_3 = u_3 + u_5, \tag{A.8}$$

$$2v_4 = u_2 + u_3 + u_5 - u_1, \tag{A.9}$$

$$2v_5 = u_2 + u_3 + u_4 - u_1. \tag{A.10}$$

If we now square each equation and sum, we get

$$4\|v\|^2 = 2u_1^2 + 4u_2^2 + 4u_3^2 + 2u_4^2 + 2u_5^2 + S, \tag{A.11}$$

where S is the sum of the cross terms, which turns out to be given by

$$\begin{aligned} S \;=\; & -4u_1u_2 - 4u_1u_3 - 2u_1u_4 - 2u_1u_5 + \\ & +6u_2u_3 + 4u_2u_4 + 2u_2u_5 + 2u_3u_4 + 4u_3u_5. \end{aligned} \tag{A.12}$$

Using the bound $ab \leq \frac{a^2+b^2}{2}$ for each of the terms in the above sum, we get

$$4\|v\|^2 \leq 8u_1^2 + 12u_2^2 + 12u_3^2 + 6u_4^2 + 6u_5^2 \tag{A.13}$$

so $4\|v\|^2 \leq 12\|u\|^2$, i.e. $\|v\|^2 \leq 3Q(v)$. ∎

Appendix B. Appendix: Derivation of Formulas (3.26) and (3.58)

We use the identity

$$\frac{\partial \xi^a}{\partial \xi_i} = \frac{a_i}{\xi_i}\xi^a, \tag{B.1}$$

valid if $a = (a_1, \ldots, a_n) \in \mathbb{R}^n$.

If f is a function

$$f(\xi) = \frac{p + q\xi^{2a}}{(1+\xi^a)^2} \tag{B.2}$$

then

$$\begin{aligned} \frac{\partial f}{\partial \xi_i} &= \frac{\frac{2a_i}{\xi_i} q\xi^{2a}(1+\xi^a)^2 - \frac{2a_i}{\xi_i}\xi^a(1+\xi^a)(p+q\xi^{2a})}{(1+\xi^a)^4} \\ &= \frac{2a_i}{\xi_i}\frac{q\xi^{2a} - p\xi^a}{(1+\xi^a)^3} \ . \end{aligned} \tag{B.3}$$

Setting $p = 0$, $q = \hat{q}$, $a = \hat{a}$, we get (3.58).

If we now multiply (3) by v_i and sum over i, we get

$$D_v f = 2\Big\langle \frac{v}{\xi}, a \Big\rangle \frac{2a_i}{\xi_i}\frac{q\xi^{2a} - p\xi^a}{(1+\xi^a)^3}. \tag{B.4}$$

Differentiation with respect to ξ_i yields

$$\begin{aligned} \frac{\partial}{\partial \xi_i}(D_v f) &= 2\Big\langle \frac{v}{\xi}, a \Big\rangle \\ &\quad \left[\frac{\left(\frac{2a_i}{\xi_i} q\xi^{2a} - \frac{a_i}{\xi_i} p\xi^a\right)(1+\xi^a)^3 - 3\left(q\xi^{2a} - p\xi^a\right)(1+\xi^a)^2 \frac{a_i}{\xi_i}\xi^a}{(1+\xi^a)^6}\right] \\ &\quad - \frac{2v_i a_i}{\xi_i^2}\left[\frac{q\xi^{2a} - p\xi^a}{(1+\xi^a)^3}\right] \\ &= 2\Big\langle \frac{v}{\xi}, a \Big\rangle \frac{a_i}{\xi_i}\left[\frac{-q\xi^{3a} + 2(p+q)\xi^{2a} - p\xi^a}{(1+\xi^a)^4}\right] \\ &\quad - \frac{2v_i a_i}{\xi_i^2}\left[\frac{q\xi^{2a} - p\xi^a}{(1+\xi^a)^3}\right]. \end{aligned} \tag{B.5}$$

If we multiply by v_i and add over i, we get

$$\begin{aligned} D_v^2 f &= 2\Big\langle \frac{v}{\xi}, a \Big\rangle^2 \left[\frac{-q\xi^{3a} + 2(p+q)\xi^{2a} - p\xi^a}{(1+\xi^a)^4}\right] \\ &\quad - 2\Big\langle \frac{v^2}{\xi^2}, a \Big\rangle \left[\frac{q\xi^{2a} - p\xi^a}{(1+\xi^a)^3}\right]. \end{aligned} \tag{B.6}$$

If we now set $q = 1$, $a = a^j$, and sum over j, we obtain Formula (3.26). ∎

References

[1] M. Brady, R. Raghavan, and J. Slawny, "Gradient descent fails to separate," in *Proc. IEEE International Conference on Neural Networks* **I**, San Diego, California, July 1988, 649–656.

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

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

[4] E.D. Sontag, *Some remarks on the backpropogation algorithm for neural net learning*, Rutgers Center for Systems and Control Technical Report 88-07, August 1988.

[5] E.D. Sontag and H.J. Sussmann, *Backpropagation Separates when Perceptrons Do*, Rutgers Center for Systems and Control Technical Report 88-12, November 1988.

[6] H.J. Sussmann, *On the convergence of learning algorithms for Boltzmann machines*, Rutgers Center for Systems and Control Technical Report 88-03, August 1988. Submitted to *Neural Networks*.