Communicated by Halbert White

Approximation of Boolean Functions by Sigmoidal Networks: Part I: XOR and Other Two-Variable Functions

E. K. Blum
Mathematics Department, University of Southern California, Los Angeles, CA 90089 USA

We prove the existence of a manifold of exact solutions (mean-square error $E = 0$) of weights and thresholds for sigmoidal networks for XOR and other 2-variable Boolean functions. We also prove the existence of a manifold of local minima of E where $E \neq 0$.

1 Introduction

It is well known that any Boolean function of n variables can be approximated arbitrarily closely by a suitable two-layer feed-forward network of sigmoidal "neurons." For example, for $n = 2$ the network shown in Figure 1 can be used (Rumelhart *et al.* 1986; Chauvin 1989; Soulie *et al.* 1987). x_1 and x_2 are the input nodes. The "hidden" layer outputs are $y_1 = \sigma(w_1x_1 + w_2x_2 - L_1)$ and $y_2 = \sigma(w_3x_1 + w_4x_2 - L_2)$, where $\sigma(\alpha) = 1/[1 + \exp(-\alpha)]$. The network output function is $z = \sigma(U_1y_1 + U_2y_2 - L_0)$. By appropriate choice of the weights w_i, U_i and thresholds L_i the function z can be made to approximate any of the 16 Boolean functions of x_1 and x_2. For the eight symmetric functions, it is sufficient to impose the following *symmetry constraints*: $w_3 = w_2, w_4 = w_1, L_2 = L_1, U_1 = U_2$. (To simplify notation we let $w = U_1 = U_2$ and $L = L_0$.) For example, applying these constraints with $w_1 = 4$, $w_2 = -4$, $L_1 = 3$, $w = 10$, and $L = 3$, we obtain the following values of $z(x_1, x_2)$: $z(0,0) = z(1,1) = 0.1139$, $z(0,1) = z(1,0) = 0.99869$, which yields a good approximation to XOR. To approximate OR we take $w_1 = 4$, $w_2 = -1$, $L_1 = 3$, $w = 10$, $L = 3$ to get $z(0,0) = 0.1139$, $z(0,1) = z(1,0) = 0.9889$ and $z(1,1) = 0.9991$.

Let $\xi_1 = (0,0)$. $\xi_2 = (0,1)$, $\xi_3 = (1,0)$, and $\xi_4 = (1,1)$. In a typical application of feed-forward networks of this kind, one specifies "target" values t_i in the real interval $(0,1)$ and seeks values of the weights and thresholds such that for all $i, z_i = z(\xi_i)$ approximates t_i sufficiently closely. The usual measure of approximation is the mean square error, $E = \sum_i E_i$, where $E_i = (z_i - t_i)^2/2$. In this example, letting $v = (w, w_1, w_2, L, L_1)$, we see that E is a function of v for given t_i. To find an acceptable value of v one usually applies a procedure to minimze $E(v)$. If an exact

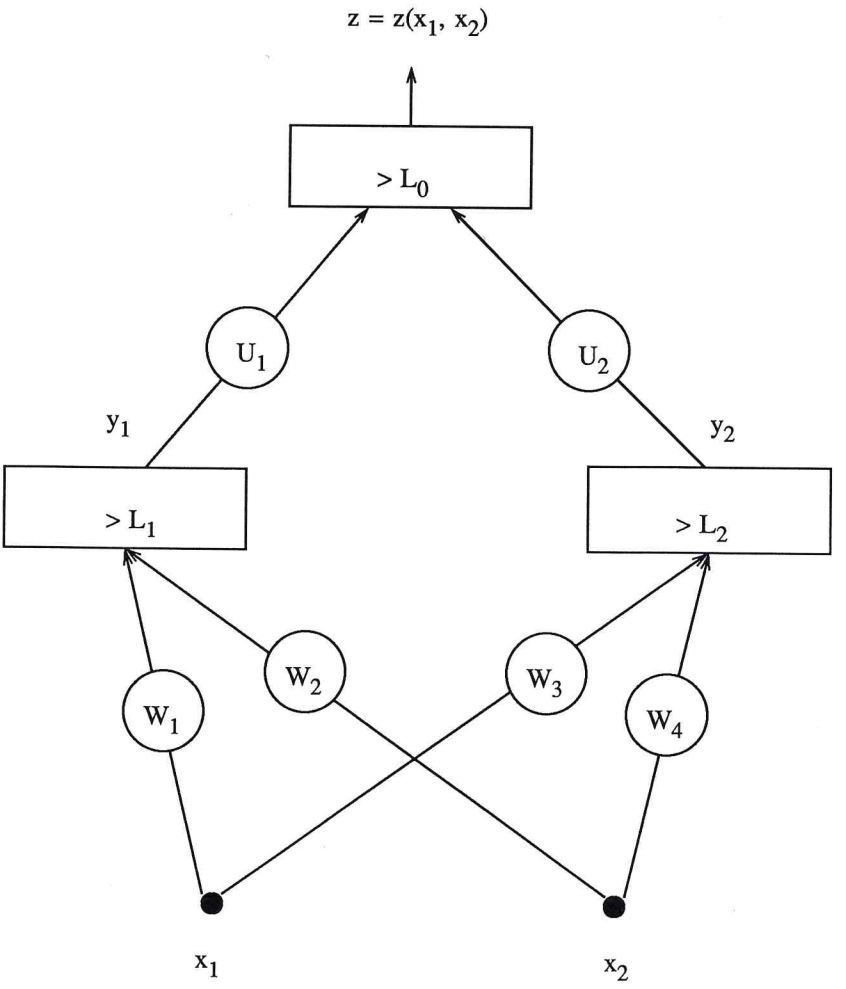

Figure 1: Network for Boolean functions of two variables.

solution, v^*, exists, then min $E(v) = E(v^*) = 0$ is the absolute minimum. Otherwise, min $E(v) > 0$. The necessary condition for a minimum is that $\nabla E(v^*) = 0$. This is not sufficient, since there may exist stationary points that are maxima or saddle points. More important, there may be stationary points that are local minima but not absolute minima.

Therefore, in the general feed-forward network problem, when a v is found for which $\nabla E(v) = 0$ (approximately) but for which $E(v)$ is too large (i.e., some z_i failing to be acceptably close to its t_i), one does not know whether to ascribe the failure to v being an unacceptable local minimum or to the nonexistence of an acceptable absolute minimum for the particular network. One does not know whether to continue to search for other minima or to change the structure of the network. In this paper, we analyze these questions for the symmetric $n = 2$ case, in particular, for the network in Figure 1. We show that a manifold of exact solutions exist, but also a manifold of local minima exists. In Part II (with Leong Li), we shall analyze the behavior of various gradient methods with respect to these manifolds and extend the results to $n > 2$.

2 Existence of Exact Solutions for Symmetric Functions

Let $0 < t_i < 1$, $1 \leq i \leq 4$, be target values for the respective inputs ξ_i as above. If $t_2 = t_3$, we call $\{t_i\}$ *a symmetric target set* (STS). Case 1: An STS with $t_2 > t_1$ and $t_2 > t_4$ is said to approximate XOR. Case 2: An STS with $t_2 = t_4$ and $t_2 > t_1$ is said to approximate OR. Case 3: An STS with $t_1 = t_4 = t_2$ is said to approximate a constant functiom (0 or 1). Case 4: An STS with $t_2 = t_1$ and $t_4 > t_1$ is said to approximate AND. (The other four symmetric functions are complements of these.) Note that $\sigma^{-1}(t_i) = -\ell n(1/t_i - 1)$ is the inverse function of σ.

Theorem 1. *Let $\{t_i\}$ be an STS. There exists a vector of parameter values, $v = (w, w_1, w_2, L, L_1)$, such that $E(v) = 0$ for the symmetric version of the network of Figure 1, that is, $z_i = t_i$.*

Proof. We give an explicit construction for computing v. Until further notice let L_1 be arbitrary but fixed. $y_1(\xi_1) = \sigma(-L_1)$. To satisfy $z_1 = t_1$ we must choose w and L such that

$$w = [L + \sigma^{-1}(t_1)]/2\sigma(-L_1) \tag{2.1}$$

We shall need $w \neq 0$. Hence, we require

(i) $L \neq -\sigma^{-1}(t_1)$

To satisfy $z_4 = t_4$ we must choose L and w so that also

$$y_1(\xi_4) = [L + \sigma^{-1}(t_4)]/2w \tag{2.2}$$

Since $y_1 > 0$, we require that

(ii) $L > -\sigma^{-1}(t_4)$ if $w > 0$ or $L < -\sigma^{-1}(t_4)$ if $w < 0$

Further, since $y_1 < 1$, equations 2.2 and 2.1 require that

(iii) $L > [\sigma^{-1}(t_4)\sigma(-L_1) - \sigma^{-1}(t_1)]/[1 - \sigma(-L_1)]$

if $w > 0$ and $1 > \sigma(-L_1)$ or $w < 0$ and $1 < \sigma(-L_1)$, with a reversed inequality in (iii) otherwise.

Next, having found L and w satisfying the above conditions and with $y_1(\xi_4)$ given by equation 2.2, we solve for $K = w_1 + w_2$ in

$$\sigma(K - L_1) = y_1(\xi_4) \tag{2.3}$$

It is useful to combine 2.1 and 2.2 to get $y_1(\xi_4) = \delta\sigma(-L_1)$, where

$$\delta = \delta(L) = [L + \sigma^{-1}(t_4)]/[L + \sigma^{-1}(t_1)] \tag{2.4}$$

Letting $\alpha = \exp(L_1)$, we have $\sigma(-L_1) = 1/(1+\alpha)$. Hence, from equation 2.3,

$$\alpha \exp(-K) = (1 + \alpha)/\delta - 1 \tag{2.5}$$

With L_1 given and L and w chosen as above, equation 2.5 is an equation in w_1 and w_2. To obtain a second equation we set $z_2 = t_2$ by solving for A in $\sigma(wA - L) = t_2$, where $A = y_1(\xi_2) + y_2(\xi_2)$. We get $A = [L + \sigma^{-1}(t_2)]/w = 2\gamma\sigma(-L) = 2\gamma/(1 + \alpha)$, where

$$\gamma = \gamma(L) = [L + \sigma^{-1}(t_2)]/[L + \sigma^{-1}(t_1)] \tag{2.6}$$

Since

$$\sigma(w_1 - L_1) + \sigma(w_2 - L_1) = A \tag{2.7}$$

we must require that $0 < A < 2$. $A > 0$ requires that

(iv) $L > -\sigma^{-1}(t_2)$ if $w > 0$ or $L < -\sigma^{-1}(t_2)$ if $w < 0$

$A < 2$ requires that L satisfy

(v) $\gamma(L) < 1 + \alpha$

Conditions (i)–(v) can be satisfied by taking L sufficiently large, since this makes $w > 0$. We shall also require that $A \neq 1$. So

(vi) $L \neq 2[(1 + \alpha)\sigma^{-1}(t_1)/2 - \sigma^{-1}(t_2)]/(1 - \alpha)$

Returning to equation 2.7, letting $x = \exp(w_1)$ and $y = \exp(w_2)$, we get

$$\begin{aligned} x/(x + \alpha) + y/(y + \alpha) &= A \\ x(y + \alpha) + y(x + \alpha) &= A(x + \alpha)(y + \alpha) \\ (1 - A)\alpha x + (1 - A)\alpha y &= A\alpha^2 + (A - 2)xy \end{aligned} \tag{2.8}$$

From equation 2.5 we get $\exp(K) = C$, where $C = \alpha\delta/(\alpha + 1 - \delta)$. Since $K = w_1 + w_2$, this yields $xy = C$. Substituting C for xy in 2.8, we get $x + y = -B$, where $B = [A\alpha^2 + C(A - 2)]/(A - 1)\alpha$. Hence, $x + C/x = -B$ and

$$x^2 + Bx + C = 0 \tag{2.9}$$

Equation 2.9 has a positive real root if $B < 0$ and $B^2 \geq 4C$. Now, $B = [A(\alpha^2 + C) - 2C]/(A-1)\alpha = 2[\gamma(\alpha^2 + C) - C(1+\alpha)]/\alpha(2\gamma - 1 - \alpha)$. As $L \to \infty$, $\gamma \to 1$, $\delta \to 1$, and $C \to 1$. Hence, $B \to -2$, that is, $B < 0$. Now, substituting the expression defining C, we get $B = 2\{\gamma[\alpha^2 + \alpha\delta/(\alpha + 1 - \delta)] - \alpha\delta(\alpha+1)/(\alpha+1-\delta)\}/\alpha(2\gamma - 1 - \alpha) = 2\{\gamma[\alpha^3 + (1-\delta)\alpha^2 + \delta\alpha] - \alpha\delta(\alpha + 1)\}/\alpha(\alpha + 1 - \delta)(-\alpha + 2\gamma - 1)$. Hence, $B^2 \geq 4C$ if and only if

$$\{\gamma[\alpha^3 + (1-\delta)\alpha^2 + \delta\alpha] - \alpha\delta(\alpha+1)\}^2 \geq \delta\alpha^3(\alpha+1-\delta)(\alpha+1-2\delta)^2 \quad (2.10)$$

On the left is a polynomial in α with leading term $\gamma^2\alpha^6$ and on the right a polynomial with leading term $\delta\alpha^6$. For α large the sign of $B^2 - 4C$ is determined by $(\gamma^2 - \delta)\alpha^6$. Hence, $B^2 > 4C$ if $\gamma^2 > \delta$. Thus, by equations 2.4 and 2.6, for 2.9 to have a positive real root it suffices to take α large and $[L + \sigma^{-1}(t_2)]^2 > [L + \sigma^{-1}(t_4)][L + \sigma^{-1}(t_1)]$, that is,

$$L[2\sigma^{-1}(t_2) - \sigma^{-1}(t_1) - \sigma^{-1}(t_4)] > \sigma^{-1}(t_4)\sigma^{-1}(t_1) - [\sigma^{-1}(t_2)]^2 \quad (2.11)$$

Case 1. Approximate XOR ($t_2 > t_4$ and $t_2 > t_1$). Thus, $\sigma^{-1}(t_2) > \sigma^{-1}(t_4)$ and $\sigma^{-1}(t_2) > \sigma^{-1}(t_1)$. So equation 2.9 has a positive real root in this case if L_1 (until now arbitrary) is sufficiently large and L satisfies (i)–(v) and 2.11, which in this case becomes

$$\text{(vii)}\quad L > \{\sigma^{-1}(t_4)\sigma^{-1}(t_1) - [\sigma^{-1}(t_2)]^2\}/[2\sigma^{-1}(t_2) - \sigma^{-1}(t_1) - \sigma^{-1}(t_4)].$$

Case 2. Approximate OR ($t_2 = t_4$ and $t_2 > t_1$). Thus, $\sigma^{-1}(t_2) > \sigma^{-1}(t_1)$. Again, equation 2.11 becomes (vii), which now is just $L > -\sigma^{-1}(t_2)$.

Case 3. Approximate constant ($t_1 = t_2 = t_4$). This makes $\gamma = \delta = 1$ and both sides of equation 2.10 reduce to $\alpha^4(\alpha - 1)^2$, that is, $B^2 = 4C$ and $B = -2$, $C = 1$. So $x = 1$ and $w_1 = 0 = w_2$.

Case 4. Approximate AND ($t_1 = t_2, t_4 > t_1$). In this case, equation 2.11 becomes

$$\text{(viii)}\quad L < -\sigma^{-1}(t_2)$$

It is easily verified that (viii) is compatible with (i)–(v) in this case.

We have proved that an exact solution is obtained by first choosing an L_1 large enough, then choosing L to satisfy (i)–(v) and (vii) or (viii) as the case may be, computing w by equation 2.1 and $w_1 = \ell n\ x$ where x is a positive root of equation 2.9 and $w_2 = K - w_1$, where K satisfies equation 2.5.

Example 1. Let $t_1 = t_4 = 0.01$ and $t_2 = t_3 = 0.99$ approximate XOR (case 1). Since $\sigma(-5.5) = 0.00407 < t_1$, a reasonable guess is $L_1 = 5.5$. This should produce acceptable values for w and L in equation 2.1. L must satisfy (vii). Since $\sigma^{-1}(t_1) = \sigma^{-1}(0.01) = -4.595$, and $\sigma^{-1}(t_2) = \sigma^{-1}(0.99) = -\sigma^{-1}(0.01) = 4.595$, we see that (vii) requires only that $L > 0$. To satisfy (v), assume $L > -\sigma^{-1}(t_1)$. Then (v) becomes $L + \sigma^{-1}(t_2) < [L + \sigma^{-1}(t_1)](1 + \alpha)$, whence $L > [\sigma^{-1}(t_2) - (1+\alpha)\sigma^{-1}(t_1)]/\alpha$. Here, $\alpha = \exp(L_1) = 244.69$. So $L > [4.595 - 245.69(-4.595)]/244.69 = 4.632$. We take $L = 4.7$. By equation 2.1, $w = (4.7 - 4.595)/0.00814 = 12.884$. Next, by equation 2.6 $\gamma =$

$(4.7+4.595)/(4.7-4.595) = 88.625$. By equation 2.7, $A = 2(88.625)/245.69 = 0.7214$. By equation 2.4, $\delta = 1$. Hence, $C = 1$. The formula for B gives $B = -[0.7214(244.69)^2 - 1.2786]/0.2786 \times 244.69 = -633.576$. Solving equation 2.10, we find $x = (633.576 + 633.573)/2 = 633.574$. Thus, $w_1 = \ell n(633.574) = 6.451$. By equation 2.5, with $\delta = 1$, we get $K = 0$. So $w_2 = -w_1$. With nine-digit precision we calculate $w_1 = 6.45158607$ and $w = 12.8840961$. With these values, the network yields $z_4 - t_4 = z_1 - t_1 = -5.8 \times 10^{-10}$ and $z_2 - t_2 = -1.1 \times 10^{-7}$, an exact solution up to rounding errors.

Example 2. Let $t_1 = 0.01$ and $t_2 = t_4 = 0.99$ approximate OR (case 2). As before, take $L_1 = 5.5$. Hence, α is as in example 1. One readily verifies that $L = 4.7$ also satisfies (i)–(vii) in this case. Thus, $w = 12.8840961$ again. The value of γ is as in example 1 but now $\delta = \gamma$. Hence, we get $C = 138.07$ and $A = 0.72144061$. This yields $B = -631.13743$. Solving equation 2.9, we get $x = 630.91859$ and $w_1 = \ell n\ x = 6.4471768$. By equation 2.5, $K = 4.9277608$ and $w_2 = -1.5194160$. With these weights the network yields $z_1 - t_1 = -5 \times 10^{-10}$, $z_2 - t_2 = -1 \times 10^{-7}$, and $z_4 - t_4 = -1 \times 10^{-7}$, an exact solution up to rounding errors.

From the proof of Theorem 1 and from the two examples it is clear that there is a manifold of exact solutions for any symmetric target set approximating the four Boolean functions XOR, OR, AND, and 1. (Their complements are approximated by reversing the signs of weights and thresholds.) The manifold of exact solutions can be regarded as parametrized by L and L_1 in the region of the (L, L_1) plane defined by inequalities (i)–(viii). As L and L_1 vary over this *feasible region*, the corresponding values of w, w_1, and w_2 are determined by equations 2.1, 2.5, and 2.9. Thus, the exact solutions (the absolute minima of E) lie on a surface in (w, w_1, w_2) space parametrized by L and L_1 restricted to the feasible region. Since equation 2.9 has two solutions, the surface may have two disconnected pieces. Assuming that (L, L_1) are already in the feasible region and held fixed, a gradient descent or other iterative minimization procedure to find the weights should generate a path in weight space approaching this surface. However, in the next section, we shall show that there is another manifold, consisting of stationary points and local minima, which may attract such iterative trajectories. It has been frequently observed in numerical experiments that such points exist in problems of this kind (for example, see McInerney *et al.* 1988). Here, we are able to prove their existence and characterize the manifold of such points in relation to the manifold of absolute minima. In Part II, we shall consider their respective regions of attraction for gradient methods.

3 Stationary Points and Local Minima

For a given STS $\{t_i\}$ appproximating one of the Boolean functions we consider the mean square error E as a function of $v = (w, w_1, w_2, L, L_1)$.

We calculate $\nabla E(v) = \sum \nabla E_i(v)$, where $\nabla E_i = (\partial E_i/\partial w, \partial E_i/\partial w_1, \partial E_i/\partial w_2, \partial E_i/\partial L, \partial E_i/\partial L_1)$. Note that $\sigma'(u) = \sigma(u)[1 - \sigma(u)]$. Thus, for $1 \leq i \leq 4$,

$$\partial E_i/\partial w = (z_i - t_i)\partial z_i/\partial w = (z_i - t_i)z_i(1 - z_i)(y_{1i} + y_{2i}) = r_i\zeta_i Y_i,$$

where $r_i = (z_i - t_i)$, $\zeta_i = z_i(1 - z_i)$, $y_{ji} = y_j(\xi_i)$, and $Y_i = y_{1i} + y_{2i}$. Similarly, letting $\eta_{ji} = y_{ji}(1 - y_{ji})$, we have

$$\partial E_i/\partial w_1 = r_i\zeta_i w(\eta_{1i}\xi_{1i} + \eta_{2i}\xi_{2i}), \partial E_i/\partial w_2 = r_i\zeta_i w(\eta_{1i}\xi_{2i} + \eta_{2i}\xi_{1i}).$$

Since $y_{12} = \sigma(w_2 - L_1) = y_{23}$ and $y_{22} = \sigma(w_1 - L_1) = y_{13}$, it follows that $Y_2 = Y_3$ and $\eta_{22} = \eta_{13}$ and $\eta_{12} = \eta_{23}$. Hence, $\partial E_2/\partial w_1 = \partial E_3/\partial w_1$ and $\partial E_2/\partial w_2 = \partial E_3/\partial w_2$. Since $y_{14} = y_{24}$, $\partial E_4/\partial w_1 = \partial E_4/\partial w_2$. Further, $\partial E_i/\partial L = -r_i\zeta_i$ and $\partial E_i/\partial L_1 = -r_i\zeta_i w(\eta_{1i} + \eta_{2i})$. Sum over i to get

$$\begin{aligned}
\partial E/\partial L &= -r_1\zeta_1 - 2r_2\zeta_2 - r_4\zeta_4 \\
\partial E/\partial L_1 &= -2r_1\zeta_1 w\eta_{11} - 2r_2\zeta_2 w(\eta_{12} + \eta_{22}) - 2r_4\zeta_4 w\eta_{14} \\
\partial E/\partial w &= 2r_1\zeta_1 y_{11} + 2r_2\zeta_2 Y_2 + 2r_4\zeta_4 y_{14} \\
\partial E/\partial w_1 &= r_1 \cdot 0 + 2r_2\zeta_2 w\eta_{22} + 2r_4\zeta_4 w\eta_{14} \\
\partial E/\partial w_2 &= r_1 \cdot 0 + 2r_2\zeta_2 w\eta_{12} + 2r_4\zeta_4 w\eta_{14}
\end{aligned}$$

Setting the five partial derivatives equal to 0, we obtain five linear homogeneous equations for the r_i. Theorem 1 proves the existence of exact solutions with all $r_i = 0$. A necessary and sufficient condition that v be a stationary point that is not an absolute minimum is that these equations have a nontrivial solution $r = (r_1, r_2, r_3)$. Let $r' = (\zeta_1 r_1, \zeta_2 r_2, \zeta_3 r_3)$. Since all $\zeta_i \neq 0$, there is a a nontrivial solution r if there is a nontrivial solution r' of $Ar' = 0$, where $A = A(v)$ is the 5×3 matrix

$$A = \begin{bmatrix} 1 & 2 & 1 \\ \eta_{11} & \eta_{12} + \eta_{22} & \eta_{14} \\ y_{11} & y_{12} + y_{22} & y_{14} \\ 0 & \eta_{22} & \eta_{14} \\ 0 & \eta_{12} & \eta_{14} \end{bmatrix}$$

Theorem 2. *For any STS that approximates XOR the corresponding error function E(v) has a manifold of stationary points that are relative minima but not absolute minima.*

Proof. By the preceding discussion, it suffices to prove that there exist v such that $A(v)r' = 0$ has a nontrivial solution. This is true if and only if rank $(A) < 3$. A necessary condition on A is that $\eta_{12} = \eta_{22}$. Otherwise rows 4 and 5 are independent and, therefore, also rows 1, 4, and 5. Now, $\eta_{12} = \eta_{22}$ if and only if

(a) $y_{12} + y_{22} = 1$

that is, $\sigma(w_2 - L_1) + \sigma(w_1 - L_1) = 1$. This is equivalent to

(b) $w_1 + w_2 = 2L_1$

Thus, $y_{14} = \sigma(w_1 + w_2 - L_1) = \sigma(L_1)$. Since $y_{11} = \sigma(-L_1)$, it follows that $y_{11} + y_{14} = 1$ and hence $\eta_{11} = \eta_{14}$. Therefore, A must have the form

$$\begin{bmatrix} 1 & 2 & 1 \\ \eta_{11} & 2\eta_{12} & \eta_{11} \\ y_{11} & 1 & y_{14} \\ 0 & \eta_{12} & \eta_{14} \\ 0 & \eta_{12} & \eta_{14} \end{bmatrix}$$

and rank $(A) < 3$ if and only if the determinant, D, of the first three rows is 0. $D = 2(1 - 2y_{11})(\eta_{12} - \eta_{11})$. Hence,

(c) $\eta_{11} = \eta_{12}$ or (d) $y_{11} = 1/2$

Condition (b) with (c) or (d) is necessary and sufficient for rank $(A) < 3$.

Consider (d). This implies $L_1 = 0$ and so $w_1 = -w_2$ by (b). Also, $\eta_{11} = 1/4$. Since $y_{14} = 1 - y_{11}$, we must also have $y_{14} = 1/2$, and so $\eta_{14} = 1/4$. In this case, A must be of the form

$$\begin{bmatrix} 1 & 2 & 1 \\ 1/4 & 2\eta_{12} & 1/4 \\ 1/2 & 1 & 1/2 \\ 0 & \eta_{12} & 1/4 \\ 0 & \eta_{12} & 1/4 \end{bmatrix}$$

and to have rank $(A) < 3$ it is further necessary that the determinant of rows 1, 2, and 4 be 0, which implies $\eta_{12} = 1/4$. But then $y_{12} = 1/2$ and since $L_1 = 0$, we must have $w_2 = 0$ and therefore $w_1 = 0$.

Now, consider (c). This implies $y_{11} + y_{12} = 1$. Thus, $\sigma(-L_1) + \sigma(w_2 - L_1) = 1$, which implies $L_1 = w_2 - L_1$, that is, $w_2 = 2L_1$ and so $w_1 = 0$, by (b). However, (a) and (c) imply $y_{11} + y_{22} = 1$ also. Thus, $w_1 = 2L_1$, which implies $L_1 = w_2 = 0$ as well.

Hence, we arrive at the necessary and sufficient condition $w_1 = w_2 = L_1 = 0$. With these values, $y_{ji} = 1/2$ and $z_i = \sigma(w - L)$ for all i. All ζ_i are equal and there is a unique nontrivial solution of $Ar = 0$, namely, $r = (r_1, -r_1, r_1)$. [To satisfy $r_i = z_i - t_i$ we need to take $r_1 = (t_2 - t_1)/2$.] This implies that $t_4 = t_1$ and $t_2 = 2\sigma(w - L) - t_1$. Thus, the $\{t_i\}$ must approximate XOR, and then there are many stationary points that are not absolute minima. They correspond to w and L satisfying the equation $\sigma(w - L) = (t_1 + t_2)/2$. Hence, they lie on a line $w = L +$ const. in the (w, L) plane. Actually, these points are local minima of E, the value of E being $(t_2 - t_1)^2/2$. To prove these are local minima we show that $dE/ds = dE(v^* + s\Delta v)/ds \geq 0$ for any $v^* = (w^*, 0, 0, L^*, 0)$, where (w^*, L^*) is on the stationary line, $\Delta v = (\Delta w, \Delta w_1, \Delta w_2, \Delta L, \Delta L_1)$ has norm $\| \Delta v \| = \epsilon$ sufficiently small and $0 < s < 1$.

Since $z_i(v^*)$ and $y_{ji}(v^*)$ are all equal, all $\nabla z_i(v^*)$ are equal. Let $v = v^* + s\Delta v$. Then $z_i(v) = z_i(v^*) + s\nabla z_i(v^*) \cdot \Delta v + O(\epsilon^2)$. Letting $z = z_i(v^*) + s\nabla_i(v^*) \cdot \Delta v$, all $z_i(v) = z + O(\epsilon^2)$. Similarly, all $\zeta_i(v) = \zeta$, $y_{ji}(v) = y$, and $\eta_{ji}(v) = \eta$ up to $O(\epsilon^2)$. Henceforth, we ignore terms of size $O(\epsilon^2)$. Thus, we calculate

$$\begin{aligned}
\partial E/\partial L &= -\zeta(r_1 + 2r_2 + r_4) = -\zeta(4z - t), \text{ where } t = t_1 + 2t_2 + t_4 \\
\partial E/\partial L_1 &= -2\zeta\eta w(4z - t), \partial E/\partial w = 2\zeta y(4z - t) \\
\partial E/\partial w_1 &= \partial E/\partial w_2 = 2\zeta\eta w(2z - t_2 - t_4) = \zeta\eta w(4z - t), \text{ when } t_1 = t_4 \\
dE/ds &= \partial E/\partial w\Delta w + \partial E/\partial w_1\Delta w_1 + \partial E/\partial w_2\Delta w_2 + \partial E/\partial L\Delta L \\
&\quad + \partial E/\partial L_1\Delta L_1 \\
&= \zeta(4z - t)[2y\Delta w - 2w\eta\Delta L_1 - \Delta L + \eta w(\Delta w_1 + \Delta w_2)]
\end{aligned}$$

Since $z = \sigma[w(y_{14} + y_{24}) - L] + O(\epsilon^2)$, we have up to $O(\epsilon^2)$ $\Delta z = \zeta[2y\Delta w - 2w\eta\Delta L_1 - \Delta L + \eta w(\Delta w_1 + \Delta w_2)]$ and $dE/ds = \zeta(4z - t)\Delta z$. Since $4z - t = (4z^* - t) + 4\Delta z = 4\Delta z$, it follows that for ϵ small enough, $dE/ds \geq 0$ for all Δz with $\| \Delta z \| = \epsilon$ and $0 < s < 1$. This completes the proof.

These analytic results suggest further research, both analytical and computational (McInerney *et al.* 1988), on the nature of the error surfaces in back propagation and on the regions of attraction for the local and global minima manifolds.

Acknowledgments

This research was partially supported by AFOSR Grant 88-0245 and NSF Grant CCR-8712192.

References

Chauvin, Y. 1989. A back-propagation algorithm with optimal use of hidden units. In *Advances in Neural Information Processing Systems I*, D. Touretzky, ed. Morgan Kauffmann, San Mateo, CA.

McInerney, J., Haines, K., Biafore, S., and Hecht-Nielsen, R. 1988. Can back-propagation error surfaces have non-global minima? Dept. of Electrical Engineering and Computer Engineering, UCSD, August 1988. Abstract in *IJCNN 89 Proc.* **II**, 627.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. 1986. Learning internal representations by error propagation. In *Parallel Distributed Processing*, Vol. 1. MIT Press, Cambridge, MA.

Soulie, F. F., Gallinari, P., Le Cun, Y., and Thiria, S. 1987. Automata networks and artificial intelligence. In *Automata Networks in Computer Science*, F. F. Soulie, Y. Robert, and M. Tchuente, eds. Princeton University Press, Princeton, NJ.

Received 27 June 1989; accepted 15 September 1989.

Communicated by Dana Ballard

Backpropagation Applied to Handwritten Zip Code Recognition

Y. LeCun
B. Boser
J. S. Denker
D. Henderson
R. E. Howard
W. Hubbard
L. D. Jackel
AT&T Bell Laboratories, Holmdel, NJ 07733 USA

The ability of learning networks to generalize can be greatly enhanced by providing constraints from the task domain. This paper demonstrates how such constraints can be integrated into a backpropagation network through the architecture of the network. This approach has been successfully applied to the recognition of handwritten zip code digits provided by the U.S. Postal Service. A single network learns the entire recognition operation, going from the normalized image of the character to the final classification.

1 Introduction

Previous work performed on recognizing simple digit images (LeCun 1989) showed that good generalization on complex tasks can be obtained by designing a network architecture that contains a certain amount of a priori knowledge about the task. The basic design principle is to reduce the number of free parameters in the network as much as possible without overly reducing its computational power. Application of this principle increases the probability of correct generalization because it results in a specialized network architecture that has a reduced entropy (Denker *et al.* 1987; Patarnello and Carnevali 1987; Tishby *et al.* 1989; LeCun 1989), and a reduced Vapnik–Chervonenkis dimensionality (Baum and Haussler 1989).

In this paper, we apply the backpropagation algorithm (Rumelhart *et al.* 1986) to a real-world problem in recognizing handwritten digits taken from the U.S. Mail. Unlike previous results reported by our group on this problem (Denker *et al.* 1989), the learning network is directly fed with images, rather than feature vectors, thus demonstrating the ability of backpropagation networks to deal with large amounts of low-level information.

