The Error Surface of the simplest XOR Network has only global Minima

Ida G. Sprinkhuizen-Kuyper

Egbert J. W. Boers

Department of Computer Science, Leiden University

Niels Bohrweg 1, 2333 CA, Leiden, The Netherlands

The artificial neural network with one hidden unit and the input units connected to the output unit is considered. It is proven that the error surface of this network for the patterns of the XOR problem has minimum values with zero error and that all other stationary points of the error surface are saddle points. Also, the volume of the regions in weight space with saddle points is zero, hence training this network on the four patterns of the XOR problem using e.g. backpropagation with momentum, the correct solution with error zero will be reached in the limit with probability one.

1 Introduction

This paper studies the representation and learning aspect of the simplest feedforward artificial neural network with sigmoid transfer functions that can represent the logical e*X*clusive *OR* (XOR) function. This network consists of one hidden unit and has connections from the input units to the output unit (see figure 1a).

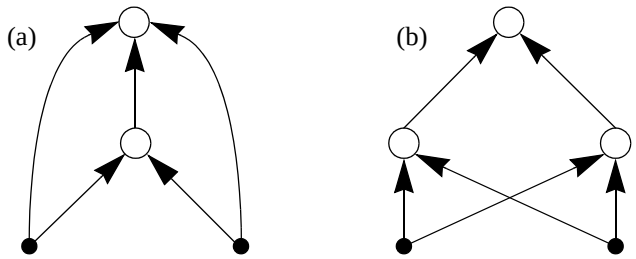

Figure 1. The simplest XOR network (a) and one with two hidden nodes (b).

The motivation to study this simple network and the XOR function is partly historical. Since Minsky and Papert (1969) wrote their *Perceptrons*, the XOR problem continued to be one of the most frequently used examples of a function needing a hidden layer for representation. Given the number of papers using the XOR problem as an example, it may seem strange that up to now, no complete analytical treatise has appeared. The complexity of even the smallest network capable to represent this function is so large, that we suspect that analytical solutions of larger, more complex networks, will not be feasible. Some general conclusions can however be drawn, generalizing the results of this small example.

The error of a network is here defined as the difference, in a least-squares sense, between the output *calculated* by the network and the *desired* output. The error of a network depends on its weights and the training patterns. With a fixed training set the error is a function of the weights: the *error surface*. The backpropagation algorithm reduces the error in the output by changing the weights—which are randomly initialized—in the direction opposite to the gradient of the error with respect to the weights and it stops when the gradient is zero. Distinction can be made between *batch* learning and *on-line* learning. During batch learning, the

weights are updated after seeing the whole training set. The errors of the individual training samples are summed to the total error. During on-line learning, the weights are corrected after *each* sample, with respect to the error for the sample just seen by the network.

1.1 Representation

First we looked at the representational power of the simplest XOR network. It is well known that this network with a threshold transfer function can represent the XOR function and that such a network with a sigmoid transfer function can approximate a solution of the XOR function. In this paper (section 3) we will show that such a network with a sigmoid transfer function can represent the XOR function exactly if TRUE ~ 0.9 and FALSE ~ 0.1 for the output unit[1]. This result is not trivial, since for a one-layer network[2] for the AND function, it is possible to find an approximate representation, but it is not possible to exactly solve the AND function, using a sigmoid transfer function.

1.2 Learning

When we assume that some kind of gradient-based learning algorithm is used, the shape of the error surface is very important for the ability of the network to learn the desired function. The ideal error surface has one minimum corresponding to an

[1]The values 0.9 and 0.1 are used, but all values $1-\delta$ and δ, for some small positive number δ, can also be used

[2]We do not count the input as a layer of the network.

acceptable solution with error zero, and in each other point in weight space a non-zero gradient. With such an error surface each gradient-based learning algorithm will approximate the minimum and find a reasonable solution. However, if the error surface has so-called local minima[1], then the learning algorithm can wind up in such a local minimum and reach a suboptimal solution. From experiments by Rumelhart *et al*. (1986) it seems that the simplest XOR network does not have local minima in contrast to the XOR network with two hidden units (see figure 1b). The problem of whether an error surface for a certain network that has to solve a certain problem, has local minima or not—and if they exist, how to avoid them—is investigated by many researchers (e.g. Gorse *et al*. 1993; Lisboa and Perantonis 1991; Rumelhart *et al*. 1986). Most researchers did numerical experiments, which gave a strong intuitive feeling of the existence of local minima, but not a real proof. Lisboa and Perantonis (1991) for example, found analytically all stationary points for both networks of figure 1, using a logarithmic error function. They claimed a local minimum for the XOR network of figure 1b, with the weights from the hidden units to the output unit equal to zero, while in the appendix we provide a proof that such a point is a saddle point and *not* a local minimum.

Blum (1989) states that the same network with the weights restricted to be symmetrical has a manifold of local minima. The same techniques as used in this paper

[1]By definition, a global minimum is also a local minimum. However, when speaking about local minima we mean here and in the rest of the paper those local minima that are *not global* minima.

prove that the points of the given manifold are saddle points and not local minima[1]. In contrast to Lisboa and Perantonis, who suggest that the simplest XOR network has local minima, this paper will analytically *prove* that the error surface of the simplest XOR network has *no* local minima.

The global minimum, with zero error, is not a strict minimum, since 3-dimensional regions in weight space exist with zero error. All points in a neighbourhood of each point in this region have error values which are *not less* than the error in that point. In a *strict* minimum, however, all points in a neighbourhood should give error values *larger* than the error value in that point. There exist more stationary points (i.e. points where the gradient of the error is zero), but we were able to prove that these points are saddle points. Saddle points are stationary points where for each neighbourhood both points with larger error values and with smaller error values can be found. Also we proved that the global minimum contains the only points with a gradient equal to zero for the error of all patterns individually. We call such a point a *stable* stationary point. The saddle points have a zero gradient for the error of a fixed training set of patterns, but not for the error of the patterns individually, so on-line learning can probably escape from these points.

The results of the analysis of the neighbourhood of these saddle points give valuable information which can be used to explain the behaviour of learning algorithms and to design learning algorithms that can escape from these saddle points.

[1]Even with the symmetric restrictions! A sketch of the proof is given in the appendix.

The remainder of the paper consists of the following sections: In section 2 the XOR problem and the network that is used to implement it are given. In section 3 it is proven that 3-dimensional regions in weight space exist with zero error. In section 4 it is proven that all stable stationary points with nonzero error are saddle points for finite values of the weights and are either saddle points or local maxima for infinite values of the weights. Section 5 consists of the proof that all unstable stationary points are saddle points. Finally section 6 contains our conclusions. An appendix is added with some more theorems and proofs used in the paper.

2 The XOR problem and the simplest network solving it

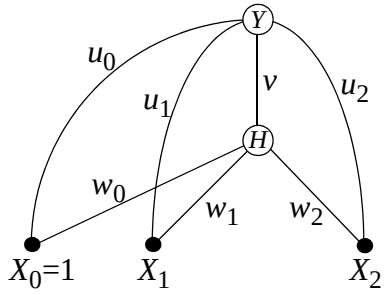

Figure 2. The simplest XOR network

The network in figure 2 with one hidden unit H is studied. This network consists of one threshold unit X_0, with constant value 1, two inputs X_1 and X_2, one hidden unit H and the output unit Y. The seven weights are labelled w_0, w_1, w_2, u_0, u_1, u_2 and v, see figure 2.

If each unit uses a sigmoid transfer function f—the commonly used transfer function $f(x) = 1/(1 + e^{-x})$ is discussed at the end of this section—the output of this network is, as function of the inputs X_1 and X_2:

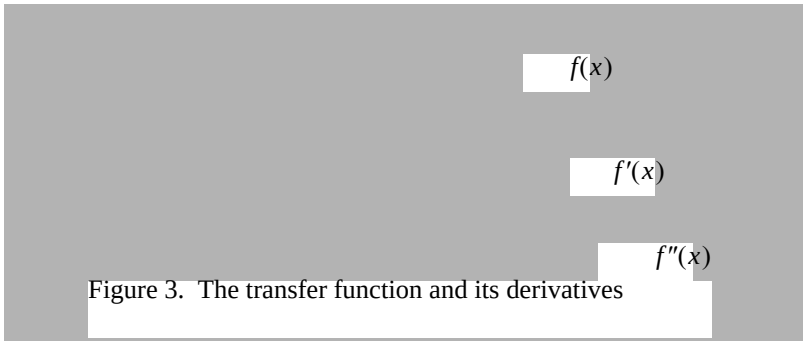

Figure 3. The transfer function and its derivatives

$$y(X_1, X_2) = f(u_0 + u_1X_1 + u_2X_2 + vf(w_0 + w_1X_1 + w_2X_2)) \tag{2.1}$$

The patterns for the XOR problem which have to be learned are $P_{X_1X_2} = ((X_1, X_2), t_{X_1X_2})$ with input (X_1, X_2) and desired output $t_{X_1X_2}$. For X_1, $X_2 \in \{0, 1\}$ the desired outputs are $t_{00} = t_{11} = 0.1$ and $t_{01} = t_{10} = 0.9$. The error E of the network when training a training set containing a_{ij} times the pattern P_{ij}, $a_{ij} > 0$, $i, j \in \{0,1\}$ is:

$$\begin{aligned} E = {} & \frac{1}{2}a_{00}(y(0, 0) - 0.1)^2 + \frac{1}{2}a_{01}(y(0, 1) - 0.9)^2 + \\ & \frac{1}{2}a_{10}(y(1, 0) - 0.9)^2 + \frac{1}{2}a_{11}(y(1, 1) - 0.1)^2 \end{aligned} \tag{2.2}$$

In the remainder of this paper it is assumed that $a_{ij} = 1$, $i, j \in \{0,1\}$. All proofs that stationary points are saddle points do not depend on the values of a_{ij}. Only the error levels corresponding to these stationary points depend on the explicit values of the a_{ij}'s.

The transfer function used is $f(x) = 1/(1 + e^{-x})$. Figure 3 shows the shape of f, f' and f''. On the interval $]-\infty,\infty[$ this function is strictly monotonously increasing from 0 to 1. In this paper we will use that $0 < f'(x)$, $\lim_{x \to \pm\infty} f'(x) = 0$, $f''(x) = 0 \Leftrightarrow x = 0$, and $f'''(0) \neq 0$, and the properties:

$$f(-x) = 1 - f(x) \tag{2.3}$$

$$f'(x) = f(x)(1 - f(x)) \tag{2.4}$$

$$f'(a) = f'(b) \Leftrightarrow a = b \vee a = -1 \tag{2.5}$$

and that the function f has an inverse function:

$$f^{-1}(x) = \ln\left(\frac{x}{1-x}\right) \quad \text{if } 0 < x < 1 \tag{2.6}$$

3 The minimum $E = 0$ can occur

The error E consists of four quadratic terms, so $E = 0$ only holds if all terms are zero. We will distinguish two kinds of minima for the error E:

- Minima remaining stable during on-line learning independent of the chosen training sequence; these minima have the property that no pattern will lead to an error that can be decreased by a local change of the weights. These minima will be called *stable minima*.
- Minima that are not stable during on-line learning, but *are* minima for batch learning. During on-line learning the weights will continue to change in a neighbourhood of such a minimum, since it is not a minimum for all patterns separately. These minima will be called *unstable minima*.

If E is equal to zero for all patterns that are in the training set, given a certain set of weights, a stable minimum is found. E can become equal to zero if and only if values of the weights exist such that all four terms in equation (2.2) are equal to zero, resulting in four equations. Application of the inverse function f^{-1} (see (2.6)) on both sides of these equations leads to:

$$
\begin{aligned}
u_0 + vf(w_0) &= f^{-1}(0.1) \approx -2.197 \\
u_0 + u_2 + vf(w_0 + w_2) &= f^{-1}(0.9) \approx 2.197 \\
u_0 + u_1 + vf(w_0 + w_1) &= f^{-1}(0.9) \approx 2.197 \\
u_0 + u_1 + u_2 + vf(w_0 + w_1 + w_2) &= f^{-1}(0.1) \approx -2.197
\end{aligned}
\tag{3.1}
$$

The equations (3.1) are linear in the variables u_0, u_1, u_2, and v. The determinant of this set of equations is equal to

$$
-f(w_0) + f(w_0 + w_1) + f(w_0 + w_2) - f(w_0 + w_1 + w_2) \tag{3.2}
$$

So for each combination of values of the weights w_0, w_1 and w_2 with this determinant unequal to zero unique values of the other weights u_0, u_1, u_2 and v can be found such that all equations of (3.1) hold. Investigation of the equation:

$$
f(w_0) - f(w_0 + w_1) - f(w_0 + w_2) + f(w_0 + w_1 + w_2) = 0 \tag{3.3}
$$

shows that this equation is equivalent to

$$
\frac{e^{-w_0}(e^{-w_1} - 1)(e^{-w_2} - 1)(e^{-2w_0 - w_1 - w_2} - 1)}{(1 + e^{-w_0})(1 + e^{-w_0 - w_1})(1 + e^{-w_0 - w_2})(1 + e^{-w_0 - w_1 - w_2})} = 0 \tag{3.4}
$$

Since $e^x > 0$ equation (3.4) and thus (3.3) has the solutions:

$$
w_1 = 0 \quad \text{or} \quad w_2 = 0 \quad \text{or} \quad 2w_0 + w_1 + w_2 = 0 \tag{3.5}
$$

Since the equations (3.1) are uniquely solvable for all values of w_0, w_1, w_2, which are not on the hyperplanes given in (3.5), we will find 3-dimensional regions in the 7-dimensional weight space, where $E = 0$. The region where $E = 0$ is a global minimum, since for all points $E \geq 0$ holds, E being a positive sum of quadratic terms. Since the dimension of the region where $E = 0$ is higher than zero, it is clear that

the minimum value $E = 0$ cannot be a strict minimum and there are always points in a neighbourhood of a point with $E = 0$ where the error is also equal to zero.

4 The minimum $E = 0$ is the unique stable minimum

In order to obtain a stable minimum, it is necessary that the gradient of the error for each pattern is zero. Writing R_{ij} for the terms depending on pattern P_{ij} we obtain:

$$\frac{\partial E}{\partial u_0} = R_{00} + R_{01} + R_{10} + R_{11} \tag{4.1}$$

with

$$\begin{aligned} R_{X_1X_2} = {} & (f(u_0 + u_1X_1 + u_2X_2 + vf(w_0 + w_1X_1 + w_2X_2)) - t_{X_1X_2}) \cdot \\ & f'(u_0 + u_1X_1 + u_2X_2 + vf(w_0 + w_1X_1 + w_2X_2)) \end{aligned} \tag{4.2}$$

The derivative $\partial E / \partial u_0$ is equal to zero for each pattern in the training set if

$$R_{00} = R_{01} = R_{10} = R_{11} = 0\,. \tag{4.3}$$

So all stable stationary points satisfy (4.3). The condition (4.3) is also a sufficient condition for a stable stationary point, since if it holds then the partial derivatives of E with respect to the other weights will be zero too. Clearly the points satisfying (3.1), i.e. the points with $E = 0$, are stable stationary points. Other stable stationary points can be found when one or more of the arguments of the derivatives of the transfer function in (4.3) (see also (4.2)) approach $\pm\infty$. The corresponding outputs go to zero or one. Let us consider

$$
\begin{aligned}
u_0 + vf(w_0) &= q_{00} \\
u_0 + u_2 + vf(w_0 + w_2) &= q_{01} \\
u_0 + u_1 + vf(w_0 + w_1) &= q_{10} \\
u_0 + u_1 + u_2 + vf(w_0 + w_1 + w_2) &= q_{11}
\end{aligned}
\tag{4.4}
$$

with one or more of the terms q_{ij} in the neighbourhood of plus or minus infinity. From these equations it follows that

$$
v(f(w_0) - f(w_0 + w_1) - f(w_0 + w_2) + f(w_0 + w_1 + w_2)) = q_{00} - q_{01} - q_{10} + q_{11} \tag{4.5}
$$

If the determinant (3.2) is unequal to zero it is possible to move the patterns one by one to their desired value by altering *v* such that the right hand side of (4.5) moves in the right direction for the considered pattern, while altering u_0, u_1 and u_2 correspondingly to keep the output of the other three patterns constant.

If the determinant (3.2) is equal to zero then $q_{00} - q_{01} - q_{10} + q_{11} = 0$ and at least two patterns have a q_{ij} in the neighbourhood of plus or minus infinity for the considered stationary points. The weights u_0, u_1 and u_2 can be used to decrease the error of two patterns at the same time. For example if q_{00} and q_{01} are both in the neighbourhood of plus infinity, decreasing u_0 and increasing u_1, keeping u_2 and $u_0 + u_1$ constant, results in a path with decreasing error moving q_{00} and q_{01} away from infinity. Other combinations can be treated similarly.

So all stationary points with one or more patterns having an output equal to 0 or 1 are not local minima. The stationary points where all four patterns give an output 0 or 1 can only be reached via a path with increasing error: these points are (local) maxima.

Conclusion: *The unique <u>stable</u> minima for the considered network for the XOR problem are 3-dimensional regions in weight space with E = 0.*

The unstable stationary points are treated in the next section.

5 All unstable stationary points are saddle points

Here all points in weight space with $\nabla E = 0$, not treated in the previous section are investigated. The components of ∇E are (4.1) and

$$\frac{\partial E}{\partial u_1} = R_{10} + R_{11} \tag{5.1}$$

$$\frac{\partial E}{\partial u_2} = R_{01} + R_{11} \tag{5.2}$$

$$\frac{\partial E}{\partial w_0} = R_{00} v f'(w_0) + R_{01} v f'(w_0 + w_2) + R_{10} v f'(w_0 + w_1) + R_{11} v f'(w_0 + w_1 + w_2) \tag{5.3}$$

$$\frac{\partial E}{\partial w_1} = R_{10} v f'(w_0 + w_1) + R_{11} v f'(w_0 + w_1 + w_2) \tag{5.4}$$

$$\frac{\partial E}{\partial w_2} = R_{01} v f'(w_0 + w_2) + R_{11} v f'(w_0 + w_1 + w_2) \tag{5.5}$$

$$\frac{\partial E}{\partial v} = R_{00} f(w_0) + R_{01} f(w_0 + w_2) + R_{10} f(w_0 + w_1) + R_{11} f(w_0 + w_1 + w_2) \tag{5.6}$$

If $\nabla E = 0$ then it is concluded from equations (4.1), (5.1) and (5.2) that

$$R_{00} = -R_{01} = -R_{10} = R_{11} \tag{5.7}$$

Since we are looking for unstable minima, we only have to consider here the case $R_{00} \neq 0$. From equations (5.3), (5.4) and (5.5) it is clear that it makes sense to distinguish between points where $v = 0$ and points where $v \neq 0$.

5.1 The case $v = 0$

If $v = 0$ and $R_{00} \neq 0$ the equations (4.1), and (5.1) till (5.6) are equivalent to (5.7) and (3.3) and it follows from equation (2.3) that equation (5.7) is equivalent to:

$$\begin{aligned} &R_{00} = (f(u_0) - 0.1)f'(u_0) = (f(-u_0 - u_1) - 0.1)f'(-u_0 - u_1) = \\ &(f(-u_0 - u_2) - 0.1)f'(-u_0 - u_2) = (f(u_0 + u_1 + u_2) - 0.1)f'(u_0 + u_1 + u_2) \neq 0 \end{aligned} \tag{5.8}$$

In theorem A.1 of the appendix we derive that this equation has exactly nine solutions for u_0, u_1 and u_2. There are three possible error levels: 0.32, 0.786045 and 0.805872. From theorem A.1 it is also clear that $R_{00} > 0$ and from (3.1) it follows that $E = 0$ cannot occur if $v = 0$.

Let us consider the partial derivatives of the error with respect to v, w_1 and w_2 in the stationary points with $v = 0$. Considering $\partial E/\partial w_1$ and $\partial E/\partial w_2$ (equations (5.4) and (5.5)), it is clear that each term contains a factor v, which will not disappear by taking the partial derivative with respect to w_1 or w_2 again. Thus also $\partial^{i+j}E/\partial w_1^i \partial w_2^j = 0$ if $i + j > 0$. Computation of some partial derivatives of E, using equation (5.7), results in:

$$\left.\frac{\partial^2 E}{\partial w_1 \partial v}\right|_{v=0} = R_{00}(-f'(w_0 + w_1) + f'(w_0 + w_1 + w_2)),$$

$$\left.\partial^3 E/\partial w_1 \partial w_2 \partial v\right|_{v=0} = R_{00} f''(w_0 + w_1 + w_2), \text{ and}$$

Figure 4. The error surface in the neighbourhood of $u_0 = u_1 = u_2 = w_0 = w_1 = w_2 = v = 0$. This picture is obtained by varying w_0, w_1 and w_2 equally from –0.5 to 0.5 and v from –0.0005 to 0.0005.

$$\partial^4 E / \partial w_1^2 \partial w_2 \partial v \Big|_{v=0} = R_{00} f'''(w_0 + w_1 + w_2) \ .$$

It is clear that at least one of these terms is unequal to zero, so theorems A.2, A.3 and A.4 prove that all stationary points with $v = 0$ are saddle points.

Figure 4 shows that indeed the error surface behaves as a saddle point when in a neighbourhood of the point with all weights zero, the weights w_0, w_1, w_2 and v are varied such that $\Delta w_0 = \Delta w_1 = \Delta w_2$ and Δv is very small with respect to Δw_i.

Thus we have proven the following theorem:

Theorem 5.1 *If $v = 0$ then all points where $\nabla E = 0$ are saddle points.*

5.2 The case v unequal to zero

If $v \neq 0$, equations (4.1), and (5.1) till (5.6) are equivalent to (5.7), (3.3) and

$$f'(w_0) = f'(w_0 + w_2) = f'(w_0 + w_1) = f'(w_0 + w_1 + w_2) \qquad (5.9)$$

Substituting the solutions of equation (3.3), given by (3.5), in equation (5.9) and applying the relation (2.5) results in the following four cases satisfying both (5.9) and (3.3):

- *Case 1:* $w_0 = w_1 = w_2 = 0$,
- *Case 2:* $w_1 = w_2 = 0$, $w_0 \neq 0$,
- *Case 3:* $w_2 = 0$, $w_1 = -2w_0$, $w_0 \neq 0$,
- *Case 4:* $w_1 = 0$, $w_2 = -2w_0$, $w_0 \neq 0$.

We will show that the stationary points of the first three cases are saddle points. The fourth case follows directly from the third case by using the symmetry in w_1 and w_2.

For the points with $w_0 = w_1 = w_2 = u_1 = u_2 = 0$ and $u_0 = -vf(0)$ belonging to case 1 the second order part of the Taylor series expansion of the error E is:

$$\begin{aligned}\Delta E / \{f'(0)\}^2 \approx\ & (\Delta u_1 + vf'(0)\Delta w_1)^2 + (\Delta u_2 + vf'(0)\Delta w_2)^2 + \\ & (2\Delta u_0 + \Delta u_1 + \Delta u_2 + 2vf'(0)\Delta w_0 + vf'(0)\Delta w_1 + vf'(0)\Delta w_2 + 2f(0)\Delta v)^2\end{aligned} \tag{5.10}$$

This second order part contains three quadratic terms, but the Hessian is not positive definite.

Inspired by (5.10) we investigated the error surface for all stationary points of cases 1 to 3 in directions such that $\Delta u_1 + vf'(w_0)\Delta w_1 = 0$, $\Delta u_2 + vf'(w_0)\Delta w_2 = 0$ and $\Delta u_0 + vf'(w_0)\Delta w_0 = 0$. Parameterizing these directions with x, y and z such that $\Delta w_0 = x$, $\Delta w_1 = y + z$, $\Delta w_2 = -x - y + z$, $\Delta u_0 = \alpha x$, $\Delta u_1 = \alpha y + \alpha z$, and $\Delta u_2 = -\alpha x - \alpha y + \alpha z$ with $\alpha = -vf'(w_0)$ gives the following expression for the error:

Figure 5. The saddle point in the neighbourhood of $u_0 = -f(0)$, $u_1 = u_2 = 0$, $w_0 = w_1 = w_2 = 0$ and $v = 1$. This picture is obtained by plotting the error against $u_1 = u_2 = -f'(0)w_1 = -f'(0)w_2$ and $u_0 = f'(0)w_0$. The weight w_1 runs from –0.1 to 0.1 and the weight w_0 runs from –0.02 to 0.02.

$$
\begin{aligned}
E = \; & \frac{1}{2}(f(u_0 + \alpha x + vf(w_0 + x)) - 0.1)^2 + \\
& \frac{1}{2}(f(u_0 + u_2 - \alpha y + \alpha z + vf(w_0 + w_2 - y + z)) - 0.9)^2 + \\
& \frac{1}{2}(f(u_0 + u_1 + \alpha x + \alpha y + \alpha z + vf(w_0 + w_1 + x + y + z)) - 0.9)^2 + \\
& \frac{1}{2}(f(u_0 + u_1 + u_2 + 2\alpha z + vf(w_0 + w_1 + w_2 + 2z)) - 0.1)^2
\end{aligned}
\tag{5.11}
$$

For case 1 ($w_0 = w_1 = w_2 = 0$) calculation of partial derivatives of E with respect to z using equation (5.7) leads to $\partial^2 E/\partial z^2 = 0$ and $\partial^3 E/\partial z^3 = 6R_{00}vf'''(0) \neq 0$ for $x = y = z = 0$, and thus the stationary points of case 1 are saddle points.

One of the saddle points of case 1 is shown in figure 5. Figure 6 shows that consideration of the error surface in the direction of each of the weights could suggest that such a point is a local minimum. So it is essential to vary the weights in the right combination in order to be able to visualize that this point is a saddle point.

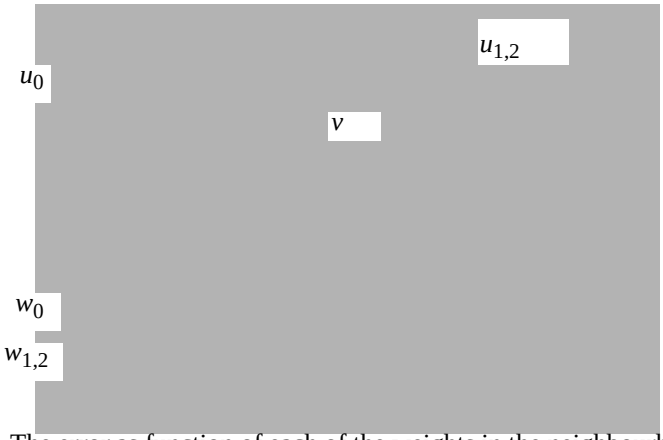

Figure 6. The error as function of each of the weights in the neighbourhood of $u_0 = -0.5$, $u_1 = u_2 = 0$, $w_0 = w_1 = w_2 = 0$ and $v = 1$. This picture gives the (false) impression that the error has a local minimum if $u_0 = -0.5$, $u_1 = u_2 = 0$, $w_0 = w_1 = w_2 = 0$ and $v = 1$. Figure 5 showed already that this point is a saddle point.

For case 2 ($w_1 = w_2 = 0$, $w_0 \neq 0$) calculation of partial derivatives of E with respect to y and z leads to $\partial^2 E / \partial y^2 = -\partial^2 E / \partial z^2 = -2R_{00} v f''(w_0)$ for $x = y = z = 0$, which is unequal to zero, so either the second order partial derivative with respect to y or that with respect to z is negative, while the other is positive, and thus also the points of case 2 are saddle points.

For case 3 ($w_2 = 0$, $w_1 = -2w_0$, $w_0 \neq 0$) calculation of partial derivatives of E with respect to x and z leads to $\partial^2 E / \partial z^2 = -2\partial^2 E / \partial x^2 = -4R_{00} v f''(w_0)$ for $x = y = z = 0$, and thus also the points of case 3 are saddle points.

Thus also the case $v \neq 0$ will not result in local minima, and we have proven the following theorem:

Theorem 5.2 *If $E \neq 0$ and $v \neq 0$, then all points where $\nabla E = 0$ are saddle points.*

6 Conclusions

The error surface of the network with one hidden unit for the XOR function has no local minima, only one global minimum with zero error. This minimum value is realized in 3-dimensional regions of the 7-dimensional weight space. Also a number of 2-dimensional regions exist where the error surface behaves as a saddle point. The levels of the error surface in the saddle points are 0.32, 0.78645 and 0.805872 respectively, for a training set with exactly one example of each pattern. When training is started with small weights, only a saddle point with error level 0.32 is possibly reached. The probability that the learning process will start in a saddle point or will end up in a saddle point is (theoretically) zero since the dimension of the region consisting of saddle points is 2, so its volume as part of the 7-dimensional weight space is zero.

A batch learning process without momentum can wind up in a saddle point, but an on-line learning process can probably escape from such a point, since the error surface is not horizontal for each individual pattern, only the average error surface for all patterns is horizontal. So a small change of the weights in the right direction will decrease the error, moving away from the saddle point. We did some experiments starting on-line learning exactly in the saddle point with all weights equal to zero and found that even with a small value of the learning parameter (0.01) and no momentum term the learning algorithm escaped from the saddle point and reached a solution with (almost) zero error in finite time.

In this paper distinction is made between stable minima (minima for each pattern) and unstable minima (minima for a training set of patterns, but not for each pattern separately). This distinction is relevant, since if an exact solution can be represented by the network, then only the absolute minima with $E = 0$ are stable minima and all other (local) minima are unstable. The fact that all local minima are unstable can be exploited by the learning algorithm to escape from these minima. However, it is more attractive to have an architecture of the network such that no local minima occur at all, as is the case for the network studied in this paper for the XOR problem. As is shown by Lisboa and Perantonis (1991), the direct connections from the inputs to the output are important for getting a good architecture for learning the XOR problem. This can be extended to modular network architectures for more difficult problems (see e.g. Boers *et al.* 1995). Finding the right architecture and learning algorithm for a problem will remain an important domain of neural network research.

In this paper we used the quadratic error function. In literature (e.g. Lisboa and Perantonis 1991) also the error function

$$E' = -\sum_{\alpha} \ln\left(y(X_1^{\alpha}, X_2^{\alpha})^{t^{\alpha}} (1 - y(X_1^{\alpha}, X_2^{\alpha}))^{1 - t^{\alpha}} \right) \tag{6.1}$$

is used, where α is the index of the pattern and t^{α} is the desired output. The stationary points for the error function E' are a subset of those for the quadratic error function. All computations needed to prove that the stationary points with the quadratic error unequal to zero are saddle points also hold when using E'. The only differ-

ence in the computations is that in the coefficients R_{ij} the factor containing the derivative of the transfer function disappears. A consequence of this alteration is that the equation $R_{00} = -R_{01} = -R_{10} = R_{11}$ for $\nabla E = 0$ has exactly one solution and not 9 as in the case considered here.

Acknowledgement

We would like to thank the referees for their valuable suggestions to improve this paper.

References

Blum, E. K. 1989. Approximation of boolean functions by sigmoidal networks: Part I: XOR and other two-variable functions. In *Neural Computation*, 1, pp. 532–540.

Boers, E. J. W., Borst, M. V., and Sprinkhuizen-Kuyper, I. G. 1995. Evolving artificial neural networks using the "Baldwin effect". In *Artificial Neural Nets and Genetic Algorithms*, D. W. Pearson, N. C. Steel, and R. F. Albrecht, eds., pp. 333–336. Springer-Verlag, Wien, New York.

Gorse, D., Shepherd, A., and Taylor, J. G. 1993. Avoiding local minima by progressive range expansion. In *Proceedings of the International Conference on Artificial Neural Networks*, S. Gielen and B. Kappen, eds., added. Springer-Verlag, Berlin.

Lisboa, P. J. G., and Perantonis, S. J. 1991. Complete solution of the local minima in the XOR problem. In *Network*, 2, pp. 119–124.

Minsky, M., and Papert, S. 1969. *Perceptrons*. MIT Press, Cambridge, MA.

Rumelhart, D. E., Hinton, G. E., and Williams, R. J. 1986. Learning internal representations by error propagation. In *Parallel Distributed Processing*, J. L. McClelland, D. E. Rumelhart,

and the PDP research group, eds., vol. 1, pp. 318–362. MIT Press, Cambridge, MA.

Sprinkhuizen-Kuyper, I.G., and Boers, E.J.W. 1994. *A Comment on a Paper of Blum: Blum's "local minima" are saddle points*. Technical Report 94-34, Leiden University, Dept. of Computer Science, The Netherlands.

APPENDIX: Some proofs and theorems

A.1 A result on error levels of the saddle points

The coefficients R_{00}, R_{01}, R_{10} and R_{11} as defined in equation (4.2) have the form $g(x) = (f(x) - 0.1)f'(x)$ with x some function of the weights. Carefully considering the cases where $\nabla E = 0$ makes clear that in all these cases equation (5.7) results in:

$$g(a) = g(-a-b) = g(-a-c) = g(a+b+c) \qquad \text{(A.1)}$$

with a, b and c functions of the weights. Investigating this equation a bit deeper we derived the following theorem:

Theorem A.1 *Let $g(x) = (f(x)–0.1)f'(x)$, and let $P_1 \approx -1.16139$ and $P_2 \approx -1.96745$ be the nonzero solutions of the equation $h_2(x) = g(x) - g(–3x)$, then the set of equations (A.1) has nine solutions which are given in table 2 (P_i stands for P_1 and P_2 respectively). For all solutions $g(a) \in \{g(0), g(P_1), g(P_2)\} = \{0.1, 0.025132, 0.0024389\}$ holds.*

Table 1: Solutions of equation (A.1)

a	b	c	$-a-b$	$-a-c$	$a+b+c$
0	0	0	0	0	0
P_i	$-2P_i$	$-2P_i$	P_i	P_i	$-3P_i$
P_i	$-2P_i$	$2P_i$	P_i	$-3P_i$	P_i
P_i	$2P_i$	$-2P_i$	$-3P_i$	P_i	P_i
$-3P_i$	$2P_i$	$2P_i$	P_i	P_i	P_i

The error levels corresponding to points with values for a, b and c given in terms of 0, P_1 and P_2

Figure 7. The functions $g(x) = (f(x)-0.1)f'(x)$, $h_1(x) = g(x) - g(-x)$ and $h_2(x) = g(x) - g(-3x)$.

are 0.32, 0.786045 and 0.805872, respectively.

Proof The values $a = b = c = 0$ certainly result in a solution. From the definition of $g(x)$ and equation (2.4) it follows that

$$g(x) = (f(x) - 0.1)f(x)(1 - f(x)) \tag{A.2}$$

Since $f(x)$ is monotonously increasing from 0 to 1, it is clear that $g(x)$ has exactly one zero point, where $f(x) = 0.1$, and one maximum and one minimum and that $\lim_{x \to \pm\infty} g(x) = 0$ (see figure 7). For each value of $g(a)$ at most two different points P and Q exist such that $g(P) = g(Q) = g(a)$. Since a, $-a-b$, $-a-c$, $a+b+c$ cannot all be negative, only the region with $g(a) > 0$, thus $a > f^{-1}(0.1)$, has to be investigated. All possibilities are tested on the equality $a + (a+b+c) = -((-a-b) + (-a-c))$, which results in conditions on P and Q. In order to obtain an extra solution it is necessary that either for some value of $x \neq 0$ the relation $g(x) = g(-x)$ or $g(x) = g(-3x)$ holds. From the graph of $h_1(x) = g(x)-g(-x)$ (see figure 7) it is clear that $h_1(x)$ is not equal to zero if $x \neq 0$. The function $h_2(x)=g(x)-g(-3x)$ (see figure 7) is equal to zero if and only if x is equal to one of the values in the set $\{0,P_1,P_2\} = \{0, -1.16139, -1.96745\}$.

Checking these possibilities results in the conclusion that the nine solutions represented in table

2 are the only solutions of $g(a) = g(-a-b) = g(-a-c) = g(a+b+c)$. q

A.2 Some theorems proving that certain points are saddle points

Theorem A.2 *Consider the function q of two variables a and b in the neighbourhood of a point where $\nabla q = 0$. If $\partial^2 q/\partial a^2 = 0$ and $\partial^2 q/\partial a \partial b \neq 0$, then the function q has a saddle point and no extreme in that point.*

Proof If for a function $q(a,b)$ of two variables $\nabla q = 0$ holds in a certain point, then the function is approximated by the second and third order terms of the Taylor series expansion:

$$\Delta q = \frac{1}{2!}\left(\frac{\partial^2 q}{\partial a^2}(\Delta a)^2 + 2\frac{\partial^2 q}{\partial a \partial b}(\Delta a)(\Delta b) + \frac{\partial^2 q}{\partial b^2}(\Delta b)^2\right) +$$
$$\frac{1}{3!}\left(\frac{\partial^3 q}{\partial a^3}(\Delta a)^3 + 3\frac{\partial^3 q}{\partial a^2 \partial b}(\Delta a)^2(\Delta b) + 3\frac{\partial^3 q}{\partial a \partial b^2}(\Delta a)(\Delta b)^2 + \frac{\partial^3 q}{\partial b^3}(\Delta b)^3\right)$$

Assuming $\partial^2 q/\partial a^2 = 0$ and $\partial^2 q/\partial a \partial b \neq 0$, and taking $\Delta a = \alpha x$ and $\Delta b = \beta x^2$ results in:

$$q = \frac{\partial^2 q}{\partial a \partial b}\alpha\beta x^3 + \frac{1}{3!}\frac{\partial^3 q}{\partial a^3}\alpha^3 x^3 + O(x^4) = \alpha\left(\frac{\partial^2 q}{\partial a \partial b}\beta + \frac{1}{3!}\frac{\partial^3 q}{\partial a^3}\alpha^2\right)x^3 + O(x^4$$

If $\partial^2 q/\partial a \partial b \neq 0$ then values of $\alpha \neq 0$ and $\beta \neq 0$ can be found such that the coefficient of x^3 is unequal to zero. Thus Δq will have values with opposite sign for $x < 0$ and $x > 0$.ëëq

Theorem A.3 *Let q be a function of three variables a, b and c. If in a point with $\nabla q = 0$, $\partial^{i+j} q/\partial a^i \partial b^j = 0$, for $0 < i+j < 6$ and $\partial^3 q/\partial a \partial b \partial c \neq 0$ (or $\partial^3 q/\partial a^2 \partial c \neq 0$ or $\partial^3 q/\partial b^2 \partial c \neq 0$), then q has a saddle point and not an extreme in that point.*

Proof Consideration of the Taylor series expansion as function of x, with $\Delta a = \alpha x$, $\Delta b = \beta x$ and $\Delta c = \gamma x^3$ results in:

$$\Delta q = \frac{1}{2!}\left(2\frac{\partial^2 q}{\partial a \partial c}\alpha\gamma + 2\frac{\partial^2 q}{\partial b \partial c}\beta\gamma\right)x^4 +$$
$$\frac{1}{3!}\left(3\frac{\partial^3 q}{\partial a^2 \partial c}\alpha^2\gamma + 6\frac{\partial^3 q}{\partial a \partial b \partial c}\alpha\beta\gamma + 3\frac{\partial^3 q}{\partial b^2 \partial c}\beta^2\gamma\right)x^5 + O(x^6$$

If $\partial^2 q/\partial a\partial c \neq 0$ or $\partial^2 q/\partial b\partial c \neq 0$ then theorem A.2 tells that the considered point is a saddle point. If both terms are equal to zero, then the coefficient of x^5 is decisive if it is unequal to zero. If $\partial^3 q/\partial a^2\partial c \neq 0$, or $\partial^3 q/\partial a\partial b\partial c \neq 0$, or $\partial^3 q/\partial b^2\partial c \neq 0$ the coefficient of x^5 is not identically zero and so nonzero values of α, β and γ can be found such that the coefficient of x^5 is unequal to zero. Thus q can attain both higher and lower values for small values of x and the point considered is a saddle point. q

Theorem A.4 *Let q be a function of three variables a, b and c. If in a point with $\nabla q = 0$, $\partial^{i+j} q/\partial a^i \partial b^j = 0$, for $0 < i+j < 8$ and $\partial^4 q/\partial a^2\partial b\partial c \neq 0$, then q has a saddle point and not an extreme in that point.*

Proof The proof is analogously to that of the previous theorem. We will take $\Delta a = \alpha x$, $\Delta b = \beta x$ and $\Delta c = \gamma x^4$, leading to the expansion:

$$
\begin{aligned}
q = {} & \frac{1}{2!}\left(2\frac{\partial^2 q}{\partial a\partial c}\alpha\gamma + 2\frac{\partial^2 q}{\partial b\partial c}\beta\gamma\right)x^5 + \\
& \frac{1}{3!}\left(3\frac{\partial^3 q}{\partial a^2\partial c}\alpha^2\gamma + 6\frac{\partial^3 q}{\partial a\partial b\partial c}\alpha\beta\gamma + 3\frac{\partial^3 q}{\partial b^2\partial c}\beta^2\gamma\right)x^6 + \\
& \frac{1}{4!}\left(4\frac{\partial^4 q}{\partial a^3\partial c}\alpha^3\gamma + 12\frac{\partial^4 q}{\partial a^2\partial b\partial c}\alpha^2\beta\gamma + 12\frac{\partial^4 q}{\partial a\partial b^2\partial c}\alpha\beta^2\gamma + 4\frac{\partial^4 q}{\partial b^3\partial c}\beta^3\gamma\right)x^7 + O(x^8
\end{aligned}
$$

If the terms with x^5 or x^6 are unequal to zero, theorems A.2 or A.3 can be applied. The coefficient of x^7 is not identically zero if $\partial^4 q/\partial a^2\partial b\partial c \neq 0$, and thus the theorem is proved. q

A.3 Some results for the XOR network with two hidden units

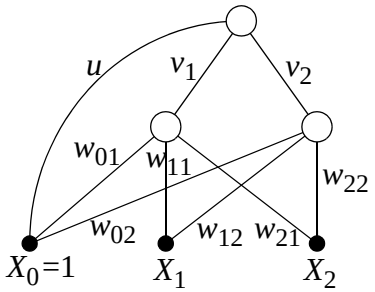

Figure 8. The XOR network with 2 hidden units

We will prove that stationary points with $v_1 = 0$ and/or $v_2 = 0$ are saddle points. For the quadratic error function, the partial derivative with respect to w_{11} is equal to:

$$\frac{\partial E}{\partial w_{11}} = S_{10} v_1 f'(w_{01} + w_{11}) + S_{11} v_1 f'(w_{01} + w_{11} + w_{21})$$

with ($t_{00} = t_{11} = 0.1$ and $t_{01} = t_{10} = 0.9$)

$$S_{X_1X_2} = (f(u + v_1 f(w_{01} + w_{11}X_1 + w_{21}X_2) + v_2 f(w_{02} + w_{12}X_1 + w_{22}X_2)) - t_{X_1X_2}) \cdot$$
$$f'(u + v_1 f(w_{01} + w_{11}X_1 + w_{21}X_2) + v_2 f(w_{02} + w_{12}X_1 + w_{22}X_2))$$

Thus $\partial^{i+j} E / \partial w_{11}^i \partial w_{12}^j = 0$ if $i + j > 0$ and $v_1 = 0$. However,

$$\left.\frac{\partial^2 E}{\partial w_{11} \partial v_1}\right|_{v_1 = 0} = S_{10} f'(w_{01} + w_{11}) + S_{11} f'(w_{01} + w_{11} + w_{21})$$

$$\left.\frac{\partial^3 E}{\partial w_{11} \partial w_{21} \partial v_1}\right|_{v_1 = 0} = S_{11} f''(w_{01} + w_{11} + w_{21})$$

and

$$\left.\frac{\partial^4 E}{\partial w_{11}^2 \partial w_{21} \partial v_1}\right|_{v_1 = 0} = S_{11} f'''(w_{01} + w_{11} + w_{21})$$

Thus, if $S_{11} \neq 0$ at least one of these terms is unequal to zero and the considered points are saddle points by theorem A.2, A.3 and A.4. If $S_{11} = 0$ and $S_{10} \neq 0$ consideration of $\partial^3 E / \partial w_{11}^2 \partial v_1$ and $\partial^4 E / \partial w_{11}^3 \partial v_1$ leads to the same conclusion. If $S_{11} = 0$, $S_{10} = 0$ and $S_{01} \neq 0$ the partial derivatives with respect to w_{21} instead of w_{11} can be considered to prove that these points are saddle points and if $S_{11} = S_{10} = S_{01} = 0$ and $S_{00} \neq 0$ the partial derivatives with respect to w_{01} lead to the same result. Finally, a point with finite weights and $S_{11} = S_{10} = S_{01} = S_{00} = 0$ both should have error zero and cannot occur if $v_1 = 0$. Thus all stationary points with finite weights and $v_1 = 0$ are saddle points. From symmetry it is clear that the same holds if $v_2 = 0$.

Lisboa and Perantonis (1991) mentioned that the point in weight space with $u = v_1 = v_2 = w_{11} =$

$w_{22} = 0.0$, $w_{01} = 1.50931$, $w_{21} = 0.48349$, $w_{02} = -0.89611$, and $w_{12} = -0.57221$ is a local minimum. They used the error function E' given in equation (6.1), but the proof given above that this point is a saddle point can be transformed into a proof for this error function by skipping the factor with the derivative of f in $S_{X_1X_2}$. Figure 9 visualizes a neighbourhood of this point in such a way that it is clear that this is a saddle point indeed.

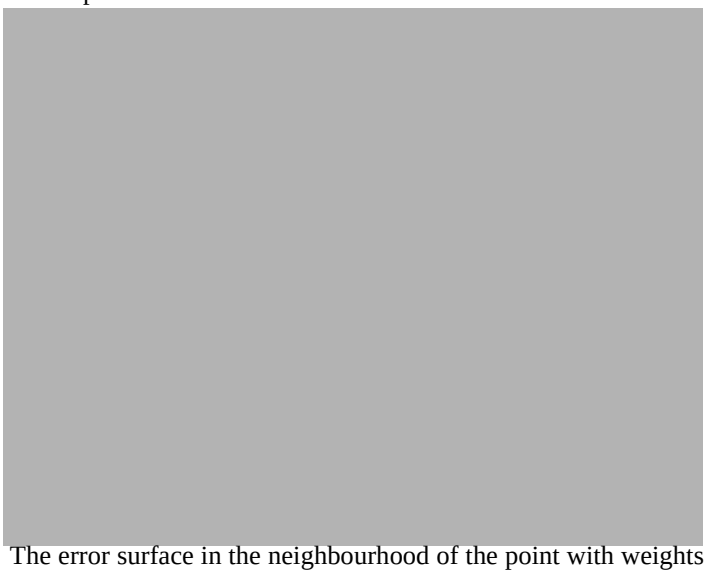

Figure 9. The error surface in the neighbourhood of the point with weights $u = v_1 = v_2 = w_{11} = w_{22} = 0$, $w_{01} = 1.50931$, $w_{21} = 0.48349$, $w_{02} = -0.89611$, $w_{12} = -0.57221$. This saddle point view is obtained by varying Δv_1 from –0.0005 till 0.0005 and $\Delta w_{01} = \Delta w_{11} = \Delta w_{21}$ from –0.5 till 0.5.

Blum (1989), found a manifold of local minima for the network of figure 8 where the weights are restricted to be symmetrical, so $v_1 = v_2 = v$, $w_{01} = w_{02} = w_0$, $w_{11} = w_{22} = w_1$ and $w_{21} = w_{12} = w_2$. The desired output of the patterns P_{00} and P_{11} is equal to t_1 and that of the patterns P_{01} and P_{10} is equal to t_2. The manifold mentioned by Blum is given by $w_0 = w_1 = w_2 = 0$ and $f(w-u) = (t_1+t_2)/2$. For the point with $v = 0$ on this manifold consideration of the partial derivatives with respect to w_1, w_2 and v result in the proof that this point is a saddle point, analogously to the proof given above. A careful analysis of the Taylor series expansion up to order 3 in directions where the second order part of this expansion is zero, leads to the result that also the other points on this manifold are saddle points. A complete proof can be found in (Sprinkhuizen-Kuyper and Boers, 1994).