The Error Surface of the 2-2-1 XOR Network: The finite stationary Points[†]

Ida G. Sprinkhuizen-Kuyper
Egbert J.W. Boers

Abstract

We investigated the error surface of the XOR problem with a 2-2-1 network with sigmoid transfer functions. In this paper it is proved that all stationary points with finite weights are saddle points or absolute minima with error zero. So, for finite weights no local minima occur. The proof results from a careful analysis of the Taylor series expansion around the stationary points. For some points coefficients of third or even fourth order in the Taylor series expansion are used to complete the proof. The proofs give a deeper insight in the complexity of the error surface in the neighbourhood of saddle points. These results can guide the research in finding learning algorithms that can handle these kind of saddle points.

1 Introduction

In neural network research, the XOR problem is probably the most frequently used test problem to experiment with different training algorithms. Usually, a network with two hidden nodes (see figure 1) is used. Training such a network with sigmoid transfer functions can be seen as searching for a global minimum on the error surface in weight space. Most learning algorithms, like backpropagation, implement a gradient descent on this surface. Rumelhart and McClelland [3] give an example of a position in weight space where such a search process is trapped: the learning algorithm is not able to find points with less error in the direct neighbourhood of that position in weight space (i.e. a *local minimum* on the error surface is found).

For small networks only, it is possible to investigate all stationary points of the error surface and determine if they are (local) extremes or saddle points. The results of such an investigation can however be very valuable for

[†]. Technical Report 95-39, Dept. of Computer Science, Leiden University. Available as ftp://ftp.wi.leidenuniv.nl/pub/CS/TechnicalReports/1995/tr95-39.ps.gz

the research of learning algorithms. The “difficult” points are known, so it can be tested how a particular learning algorithm behaves in the neighbourhood of specific “difficult” points. Also knowledge about “why” a learning algorithm is trapped in some particular point of the error surface can guide the research for learning algorithms that can escape from such points.

We started investigating the simplest network that can learn the XOR problem with one hidden node and connections directly from the inputs to the output node and found that the error surface of this network does not contain local minima (see [5]). The techniques used to find these results for the simplest “XOR network” are extended and used to investigate the error surface of the XOR network with two hidden nodes (see figure 1). In this paper we prove that all stationary points with *finite* weights cannot be local minima: they are absolute minima with error zero or they are saddle points.

In a forthcoming paper [6] we will publish our further results on the error surface of this XOR network. We found that this network has regions with local minima when some weights from the inputs to the hidden nodes have infinite values. However, these regions of local minima have boundary points that are saddle points. So, we found that from each finite point in weight space a strictly decreasing path exists to a point with error zero.

In our analysis we used the quadratic error function

$$E = \frac{1}{2}\sum_{\alpha} (O_{\alpha} - t_{\alpha})^{2}$$

It is easily seen that all results also hold for the “cross-entropy” [1, 2] error function

$$L = -\sum_{\alpha} \ln\left((O_{\alpha})^{t_{\alpha}} (1 - O_{\alpha})^{1 - t_{\alpha}} \right)$$

since all stationary points of L are also stationary points of E. For E we found some extra stationary points (case II.4 in section 4).

The results that no local minima exist for points with finite weights is independently found by Hamey [1]. He used the error L and based his work on the work of Lisboa and Perantonis [2], who found the location of all stationary points for the network with error L. We also used part of the results and techniques in [2], but we met some more complications because we started from the error E. In their paper [2] Lisboa and Perantonis remark that they investigated all stationary points and found both saddle points and local minima, but they omitted their proofs. They give 5 examples of stationary points that are local minima by their analysis. One of these points has all weights finite and is a saddle point (see figure 2). The other points are numerically equivalent — using a finite precision floating point representa-

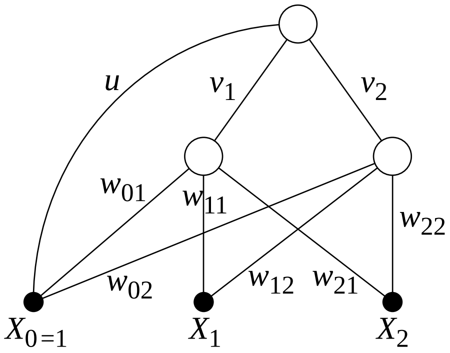

Figure 1. The XOR network with 2 hidden units

tion — to points with some of the weights having infinite values and correspond to local minima as will be proved in [6].

2 The 2-2-1 XOR network

In this paper we investigate the error surface of the network, see figure 1, with two hidden units and without direct connections from the input units to the output.

For the XOR function we assume that the following patterns should be learned/represented by the network:

Table 1: Patterns for the XOR problem

Patter n	***X* 1**	***X* 2**	**desired output**
P_{00}	0	0	0.1
P_{01}	0	1	0.9
P_{10}	1	0	0.9
P_{11}	1	1	0.1

The transfer function used is the sigmoid

$$f(x) = \frac{1}{1 + e^{-x}} \tag{2.1}$$

The input of the output unit is for the four patterns:

$$\begin{aligned} A_{00} &= u + v_1 f(w_{01}) + v_2 f(w_{02}) \\ A_{01} &= u + v_1 f(w_{01} + w_{21}) + v_2 f(w_{02} + w_{22}) \\ A_{10} &= u + v_1 f(w_{01} + w_{11}) + v_2 f(w_{02} + w_{12}) \\ A_{11} &= u + v_1 f(w_{01} + w_{11} + w_{21}) + v_2 f(w_{02} + w_{12} + w_{22}) \end{aligned} \tag{2.2}$$

So the four patters result in output values equal to $f(A_{00})$, $f(A_{01})$, $f(A_{10})$ and $f(A_{11})$, respectively.

The mean square error is equal to:

$$\begin{aligned} E = {} & \frac{1}{2}(f(A_{00}) - 0.1)^2 + \frac{1}{2}(f(A_{01}) - 0.9)^2 + \\ & \frac{1}{2}(f(A_{10}) - 0.9)^2 + \frac{1}{2}(f(A_{11}) - 0.1)^2 \end{aligned} \tag{2.3}$$

2.1 Some theorems

The following theorems are proved in [5] as theorem A2, A3, A4 and A1, respectively.

Theorem 2.1 *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 attains a saddle point and no extreme in that point.*

Theorem 2.2 *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 attains a saddle point and not an extreme in that point.*

Theorem 2.3 *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 $0 < i + j < 8$ and $\partial^4 q / \partial a^2 \partial b \partial c \neq 0$, then q attains a saddlepoint and not an extreme in that point.*

Theorem 2.4 *Let $g(x) = (f(x) - 0.1) f'(x)$, and let P_1 and P_2[†] be the nonzero solutions of the equation $g(x) - g(-3x) = 0$, then the set of equations*

$$g(a) = g(-a - b) = g(-a - c) = g(a + b + c) \tag{2.4}$$

has nine solutions which are given in table 2 (P_i stands for P_1 and P_2 respectively). For all solutions $g(a) \in \{g(0), g(P_1), g(P_2)\}$[‡] holds.

[†] $P_1 \approx -1.16139$ and $P_2 \approx -1.96745$

[‡] $\{g(0), g(P_1), g(P_2)\} \approx \{0.1, 0.025132, 0.0024389\}$

Table 2: Solutions of equation (2.4)

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

3 Stationary points

Let us introduce

$$\begin{aligned} R_{00} &= (f(A_{00}) - 0.1) f'(A_{00}) \\ R_{01} &= (f(A_{01}) - 0.9) f'(A_{01}) \\ R_{10} &= (f(A_{10}) - 0.9) f'(A_{10}) \\ R_{11} &= (f(A_{11}) - 0.1) f'(A_{11}) \end{aligned} \qquad (3.1)$$

with A_{ij} given as in (2.2). The components of the gradient of the error (2.3) are:

$$\frac{\partial E}{\partial u} = R_{00} + R_{01} + R_{10} + R_{11} \qquad (3.2)$$

$$\begin{aligned} \frac{\partial E}{\partial v_1} = {} & R_{00} f(w_{01}) + R_{01} f(w_{01} + w_{21}) + R_{10} f(w_{01} + w_{11}) + \\ & R_{11} f(w_{01} + w_{11} + w_{21}) \end{aligned} \qquad (3.3)$$

$$\begin{aligned} \frac{\partial E}{\partial v_2} = {} & R_{00} f(w_{02}) + R_{01} f(w_{02} + w_{22}) + R_{10} f(w_{02} + w_{12}) + \\ & R_{11} f(w_{02} + w_{12} + w_{22}) \end{aligned} \qquad (3.4)$$

$$\begin{aligned} \frac{\partial E}{\partial w_{01}} = {} & v_1 (R_{00} f'(w_{01}) + R_{01} f'(w_{01} + w_{21}) + \\ & R_{10} f'(w_{01} + w_{11}) + R_{11} f'(w_{01} + w_{11} + w_{21})) \end{aligned} \qquad (3.5)$$

$$\frac{\partial E}{\partial w_{11}} = v_1 \left(R_{10} f'(w_{01} + w_{11}) + R_{11} f'(w_{01} + w_{11} + w_{21}) \right) \tag{3.6}$$

$$\frac{\partial E}{\partial w_{21}} = v_1 \left(R_{01} f'(w_{01} + w_{21}) + R_{11} f'(w_{01} + w_{11} + w_{21}) \right) \tag{3.7}$$

$$\frac{\partial E}{\partial w_{02}} = v_2 \left(R_{00} f'(w_{02}) + R_{01} f'(w_{02} + w_{22}) + R_{10} f'(w_{02} + w_{12}) \right. \tag{3.8}$$
$$\left. + R_{11} f'(w_{02} + w_{12} + w_{22}) \right)$$

$$\frac{\partial E}{\partial w_{12}} = v_2 \left(R_{10} f'(w_{02} + w_{12}) + R_{11} f'(w_{02} + w_{12} + w_{22}) \right) \tag{3.9}$$

$$\frac{\partial E}{\partial w_{22}} = v_2 \left(R_{01} f'(w_{02} + w_{22}) + R_{11} f'(w_{02} + w_{12} + w_{22}) \right) \tag{3.10}$$

We will distinguish two kinds of stationary points:

- Stationary points with the property that the gradient of the error is zero for all patterns separately. These points we will call *stable stationary points*.
- The other stationary points we will call *unstable stationary points*. For these stationary points the total gradient of the error is zero, but the gradient of the error for at least one pattern is unequal to zero.

This distinction is useful both for the analysis of the error surface and for considering gradient based learning algorithms, since it explains why on-line learning can escape from some stationary points (unstable ones), while batch learning does not.

Stable stationary points are obtained when the gradient of the error is zero for each of the four patterns separately, thus if

$$R_{00} = R_{01} = R_{10} = R_{11} = 0 \tag{3.11}$$

For finite weights (3.11) only holds if

$$\begin{aligned} f(A_{00}) &= 0.1 \\ f(A_{01}) &= 0.9 \\ f(A_{10}) &= 0.9 \\ f(A_{11}) &= 0.1 \end{aligned}$$

and thus all patterns are learned exactly and the error (2.3) is zero. So the stable stationary points with finite weights are absolute minima with error zero. In [6] we will prove that a 5-dimensional region exists with error zero. In the next section the instable stationary points are investigated.

4 Instable stationary points with finite weights

Using equations (3.2) up to (3.10) in $\nabla E = 0$ results in 9 equations for R_{00}, R_{01}, R_{10} and R_{11}. Clearly $R_{00} = R_{01} = R_{10} = R_{11} = 0$ is a solution, resulting in the stable stationary points with error zero for finite weights. In this section we are investigating solutions with at least one of the terms R_{ij} unequal to zero. If v_1 and/or v_2 is equal to zero, then the number of equations is reduced by at least three. So let us first consider the stationary points where v_1 and/or v_2 is equal to zero.

4.1 Stationary points with v_1 and/or v_2 equal to zero

Analogously to the case for the simplest XOR network where the weight from the hidden node to the output unit is equal to zero, we can prove the following theorem:

Theorem 4.1 *If in a stationary point* $v_1 = 0$ *or* $v_2 = 0$ *or both, then this point is a saddle point and not an extreme.*

Proof Suppose $v_1 = 0$. Then clearly $\partial^{i+j}E/\partial w_{11}^i \partial w_{21}^j = 0$ if $i+j>0$, since all these derivatives contain at least one factor v_1 (see (3.6) and (3.7)). If $R_{11} \neq 0$, then at least one of the following inequalities is true:

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

or

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

Thus if $R_{11} \neq 0$, it follows from the theorems 2.2 and 2.3 that a stationary point with $v_1 = 0$ is a saddle point.

Suppose $v_1 = 0$ and $v_2 \neq 0$. For these points we find from (3.8), (3.9) and (3.10) that if at least one of the terms R_{ij} is unequal to zero, then all terms R_{ij} are unequal to zero. If both $v_1 = 0$ and $v_2 = 0$ we find $A_{00} = A_{01} = A_{10} = A_{11} = u$. So $R_{00} = R_{11}$ and $R_{01} = R_{10}$. From (3.2) it follows that $R_{00} = -R_{11}$ and thus $u = 0$ has to hold (see [5]). Thus also for these stationary points all terms R_{ij} are unequal to zero. So, especially R_{11} is unequal to zero and no local minimum will be attained if $v_1 = 0$ and all weights are finite. The case $v_2 = 0$ is proved completely analogously. ❑

Figure 2 shows that the point in weight space with $u = 0.0$, $v_1 = 0.0$, $v_2 = 0.0$, $w_{01} = 1.50931$, $w_{11} = 0.0$, $w_{21} = 0.48349$, $w_{02} = -0.89611$,

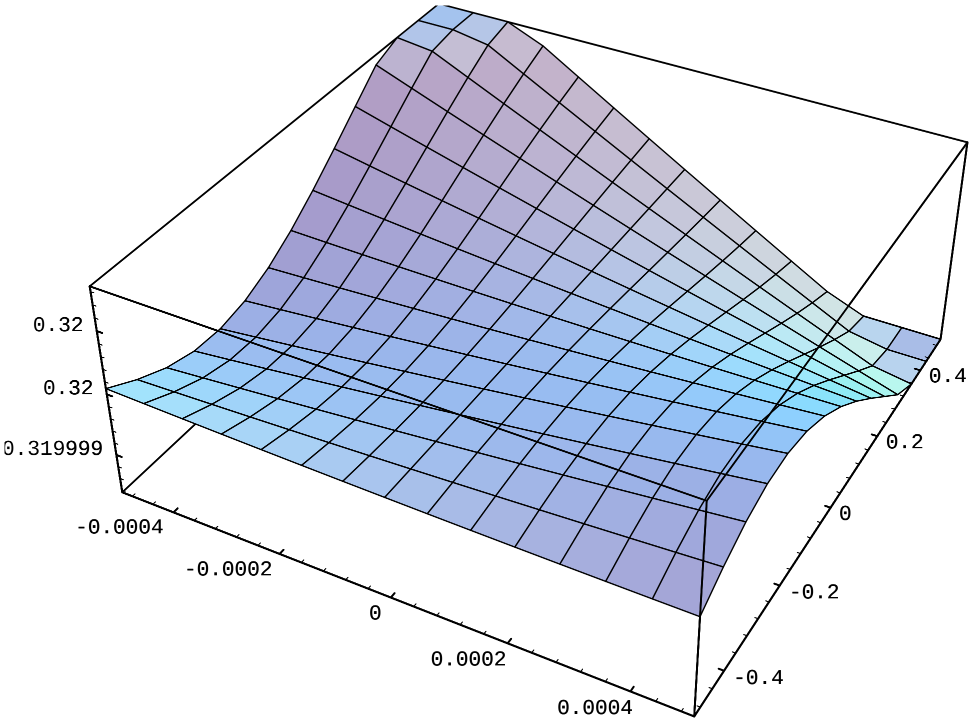

Figure 2. The error surface in the neighbourhood of the point with weights $u = v_1 = v_2 = w_{11} = w_{22} = 0$, $w_{01} = 1.50931$, $w_{21} = 0.48349$, $w_{02} = -0.89611$, $w_{12} = -0.57221$. This saddle point view is obtained by varying Δv_1 from -0.0005 to 0.0005 and $\Delta w_{01} = \Delta w_{11} = \Delta w_{21}$ from -0.5 to 0.5.

$w_{12} = -0.57221$ and $w_{22} = 0.0$, which is classified by Lisboa and Perantonis in [2] as a local minimum, really is a saddle point.

4.2 Stationary points with both v_1 and v_2 unequal to zero

From equations (3.5) till (3.10) it follows that:

$$R_{00}f'(w_{01}) = -R_{01}f'(w_{01}+w_{21}) = -R_{10}f'(w_{01}+w_{11}) = R_{11}f'(w_{01}+w_{11}+w_{21}) \tag{4.1}$$

and

$$R_{00}f'(w_{02}) = -R_{01}f'(w_{02}+w_{22}) = -R_{10}f'(w_{02}+w_{12}) = R_{11}f'(w_{02}+w_{12}+w_{22}) \tag{4.2}$$

implying:

$$\begin{aligned} R_{01} &= -\frac{f'(w_{01})}{f'(w_{01}+w_{21})}R_{00} = -\frac{f'(w_{02})}{f'(w_{02}+w_{22})}R_{00} \\ R_{10} &= -\frac{f'(w_{01})}{f'(w_{01}+w_{11})}R_{00} = -\frac{f'(w_{02})}{f'(w_{02}+w_{12})}R_{00} \\ R_{11} &= \frac{f'(w_{01})}{f'(w_{01}+w_{11}+w_{21})}R_{00} = \frac{f'(w_{02})}{f'(w_{02}+w_{12}+w_{22})}R_{00} \end{aligned} \tag{4.3}$$

Thus either all $R_{ij} = 0$ or all $R_{ij} \neq 0$. So let us suppose that all $R_{ij} \neq 0$. Substitution of the first parts of the equalities (4.3) in equation (3.3) results in:

$$\frac{f(w_{01})}{f'(w_{01})} - \frac{f(w_{01}+w_{21})}{f'(w_{01}+w_{21})} - \frac{f(w_{01}+w_{11})}{f'(w_{01}+w_{11})} + \frac{f(w_{01}+w_{11}+w_{21})}{f'(w_{01}+w_{11}+w_{21})} = 0 \tag{4.4}$$

Using $f(x)/f'(x) = 1 + e^x$, this equation is equivalent to:

$$e^{w_{01}} - e^{w_{01}+w_{21}} - e^{w_{01}+w_{11}} + e^{w_{01}+w_{11}+w_{21}} = e^{w_{01}}(1-e^{w_{11}})(1-e^{w_{21}}) = 0 \tag{4.5}$$

and thus we have $w_{11} = 0$ or $w_{21} = 0$ in an instable stationary point with $v_1 \neq 0$ and $v_2 \neq 0$ and finite weights. Similarly we find from the substitution of the second parts of (4.3) in (3.4) that $w_{12} = 0$ or $w_{22} = 0$.

So we have to consider the following four cases for instable stationary points with $v_1 \neq 0$ and $v_2 \neq 0$ and finite weights:

- Case I: $w_{21} = 0$ and $w_{22} = 0$,
- Case II: $w_{21} = 0$ and $w_{12} = 0$,
- Case III: $w_{11} = 0$ and $w_{22} = 0$,
- Case IV: $w_{11} = 0$ and $w_{12} = 0$.

Remark that case I and case IV are essentially the same because of the symmetry in the network obtained from interchanging X_1 and X_2. Case II and case III are equivalent as well.

Let us consider cases I and II carefully, taking into account that equations (4.3) have to hold with R_{00}, R_{01}, R_{10} and R_{11} given by (3.1).

4.2.1 Case I: $w_{21} = 0$ and $w_{22} = 0$

In this case equation (4.3) becomes:

$$
\begin{aligned}
R_{01} &= -R_{00} \\
R_{10} &= -\frac{f'(w_{01})}{f'(w_{01}+w_{11})}R_{00} = -\frac{f'(w_{02})}{f'(w_{02}+w_{12})}R_{00} \\
R_{11} &= \frac{f'(w_{01})}{f'(w_{01}+w_{11})}R_{00} = \frac{f'(w_{02})}{f'(w_{02}+w_{12})}R_{00} = -R_{10}
\end{aligned}
$$

while equation (2.2) results in $A_{00} = A_{01}$ and $A_{10} = A_{11}$. Combining this with $R_{01} = -R_{00}$ and $R_{10} = -R_{11}$ leads to (using the results from [5])

$$A_{00} = u + v_1 f(w_{01}) + v_2 f(w_{02}) = 0 \tag{4.6}$$

and

$$A_{11} = u + v_1 f(w_{01} + w_{11}) + v_2 f(w_{02} + w_{12}) = 0 \tag{4.7}$$

So also $R_{00} = R_{11}$ and the error for case I is 0.32 (all patterns give output 0.5). Further, $R_{00} = R_{11}$ implies $f'(w_{01}) = f'(w_{01} + w_{11})$ and $f'(w_{02}) = f'(w_{02} + w_{12})$. Since $f'(a) = f'(b)$ if and only if $a = b$ or $a = -b$, we find that ($w_{11} = 0$ or $w_{11} = -2w_{01}$) and ($w_{12} = 0$ or $w_{12} = -2w_{02}$). So we can split case I into the four cases I.1 to I.4:

- Case I.1: $w_{21} = 0$, $w_{22} = 0$, $w_{11} = 0$ and $w_{12} = 0$,
- Case I.2: $w_{21} = 0$, $w_{22} = 0$, $w_{11} = 0$ and $w_{12} = -2w_{02}$,
- Case I.3: $w_{21} = 0$, $w_{22} = 0$, $w_{11} = -2w_{01}$ and $w_{12} = 0$,
- Case I.4: $w_{21} = 0$, $w_{22} = 0$, $w_{11} = -2w_{01}$ and $w_{12} = -2w_{02}$,

and will investigate these cases further.

Case I.1: $w_{21} = 0$, $w_{22} = 0$, $w_{11} = 0$ and $w_{12} = 0$

Equations (4.6) and (4.7) become both equal to

$$u + v_1 f(w_{01}) + v_2 f(w_{02}) = 0$$

In this case we can choose the weights w_{01}, w_{02}, v_1 ($\neq 0$) and v_2 ($\neq 0$). Then the value of u can be determined such that the former equation hold. So these stationary points form a 4-dimensional region in the weight space with error 0.32.

We will prove that these points are saddle points. In order to do so we consider the second order part of the Taylor expansion of the error:

$$
\begin{aligned}
\Delta E \; = \; & f'(0)^2 \Bigg(4 \Bigg(\Delta u + f(w_{01})\Delta v_1 + f(w_{02})\Delta v_2 + v_1 f'(w_{01})\Delta w_{01} \\
& + \frac{1}{2} v_1 f'(w_{01})\Delta w_{11} + \frac{1}{2} v_1 f'(w_{01})\Delta w_{21} + v_2 f'(w_{02})\Delta w_{02} \\
& + \frac{1}{2} v_2 f'(w_{02})\Delta w_{12} + \frac{1}{2} v_2 f'(w_{02})\Delta w_{22} \Bigg)^2 \\
& + \left(v_1 f'(w_{01})\Delta w_{11} + v_2 f'(w_{02})\Delta w_{12} \right)^2 \\
& + \left(v_1 f'(w_{01})\Delta w_{21} + v_2 f'(w_{02})\Delta w_{22} \right)^2 \Bigg) \\
& + 2 v_1 \left(f(0) - 0.1 \right) f'(0) f''(w_{01})\Delta w_{11}\Delta w_{21} \\
& + 2 v_1 \left(f(0) - 0.1 \right) f'(0) f''(w_{02})\Delta w_{12}\Delta w_{22}
\end{aligned}
\tag{4.8}
$$

We will study variations of the weights such that

$$v_1 f'(w_{01})\Delta w_{11} + v_2 f'(w_{02})\Delta w_{12} \; = \; 0 \tag{4.9}$$

$$v_1 f'(w_{01})\Delta w_{21} + v_2 f'(w_{02})\Delta w_{22} \; = \; 0 \tag{4.10}$$

and

$$
\begin{aligned}
& \Delta u + f(w_{01})\Delta v_1 + f(w_{02})\Delta v_2 + v_1 f'(w_{01})\Delta w_{01} \\
& \; + v_2 f'(w_{02})\Delta w_{02} = 0
\end{aligned}
\tag{4.11}
$$

If we choose Δw_{ij} such that equations (4.9), (4.10) and (4.11) are satisfied it is easily seen from the nonquadratic terms in equation (4.8) that a saddle point is attained in those points where

$$\alpha^2 v_1 f''(w_{01}) + v_2 f''(w_{02}) \neq 0 \tag{4.12}$$

with

$$\alpha \; = \; -\frac{v_2 f'(w_{02})}{v_1 f'(w_{01})} \tag{4.13}$$

Consider the stationary points of case I.1 with

$$\alpha^2 v_1 f''(w_{01}) + v_2 f''(w_{02}) \; = \; 0 \tag{4.14}$$

Since this is a thin set of the 4-dimensional region of stationary points with $w_{11} = w_{21} = w_{12} = w_{22} = u + v_1 f(w_{01}) + v_2 f(w_{02}) = 0$, these points also have to be saddle points (see [1]).

In the following we will give a proof that these points are saddle points by showing that certain partial derivatives are unequal to zero. We can choose 6 variables w, x_0, x_1, x_2, y and z for the variations of the weights in the hyperplane given by the equations (4.9), (4.10) and (4.11) in order to investigate the neighbourhood of points on the error surface in case I.1 more precisely:

$$\begin{aligned}
&\Delta w_{12} = x_1, \Delta w_{11} = \alpha x_1 \\
&\Delta w_{22} = x_2, \Delta w_{21} = \alpha x_2 \\
&\Delta w_{02} = x_0 + z, \Delta w_{01} = \alpha x_0 - \alpha z \\
&\Delta v_2 = y + w, \Delta v_1 = \beta y - \beta w \\
&\Delta u = f(w_{01})\beta w - f(w_{02})w + \alpha v_1 f'(w_{01})z - v_2 f'(w_{02})z
\end{aligned} \tag{4.15}$$

where α is given in (4.13) and β is given by:

$$\beta = -\frac{f(w_{02})}{f(w_{01})} \tag{4.16}$$

If (4.14) holds then all first and second order derivatives of the error E with respect to the six variables w, x_0, x_1, x_2, y and z are zero. Calculation of some third order derivatives in the stationary points of case I.1 in the directions given by (4.15) results in the following formulas, where $\Psi|_0$ stands for Ψ restricted to $w = x_0 = x_1 = x_2 = y = z = 0$:

$$\left.\frac{\partial^3 E}{\partial x_1^2 \partial x_2}\right|_0 = (f(0) - 0.1)f'(0)\,(\alpha^3 v_1 f'''(w_{01}) + v_2 f'''(w_{02})) \tag{4.17}$$

$$\left.\frac{\partial^3 E}{\partial x_1 \partial x_2 \partial y}\right|_0 = (f(0) - 0.1)f'(0)\,(\alpha^2 \beta f''(w_{01}) + f''(w_{02})) \tag{4.18}$$

$$\left.\frac{\partial^3 E}{\partial x_1 \partial x_2 \partial z}\right|_0 = (f(0) - 0.1)f'(0)\,(-\alpha^3 v_1 f'''(w_{01}) + v_2 f'''(w_{02})) \tag{4.19}$$

$$\left.\frac{\partial^3 E}{\partial x_1 \partial x_2 \partial w}\right|_0 = (f(0) - 0.1)f'(0)\,(-\alpha^2 \beta f''(w_{01}) + f''(w_{02})) \tag{4.20}$$

Thus a saddlepoint is found if inequality (4.12) or one of the following inequalities hold:

$$\alpha^3 v_1 f'''(w_{01}) + v_2 f'''(w_{02}) \neq 0 \tag{4.21}$$

or

$$\alpha^2 \beta f''(w_{01}) + f''(w_{02}) \neq 0 \tag{4.22}$$

or

$$-\alpha^3 v_1 f'''(w_{01}) + v_2 f'''(w_{02}) \neq 0 \tag{4.23}$$

or

$$-\alpha^2 \beta f''(w_{01}) + f''(w_{02}) \neq 0 \tag{4.24}$$

At least one of these inequalities has to be fulfilled, since the set of equations:

$$\begin{aligned} \alpha^2 v_1 f''(w_{01}) + v_2 f''(w_{02}) &= 0 \\ \alpha^3 v_1 f'''(w_{01}) + v_2 f'''(w_{02}) 0 &= 0 \\ \alpha^2 \beta f''(w_{01}) + f''(w_{02}) &= 0 \\ -\alpha^3 v_1 f'''(w_{01}) + v_2 f'''(w_{02}) &= 0 \\ -\alpha^2 \beta f''(w_{01}) + f''(w_{02}) &= 0 \end{aligned} \tag{4.25}$$

has no solutions in the considered case with all weights finite and v_1 and v_2 unequal to zero: From the second and the fourth equation of (4.25) it follows that $f'''(w_{01}) = 0$, while from the third and fifth equation it has to be concluded that $f''(w_{01}) = 0$. Since the last equality implies that $w_{01} = 0$ and since $f'''(0) \neq 0$ these equalities are contradictory.

Conclusion 4.1 *All points of case I.1 are saddle points.*

In figures 3 and 4 it is visualized that the point with weights $u = -1$, $v_1 = v_2 = 1$, $w_{01} = w_{11} = w_{21} = w_{02} = w_{12} = w_{22} = 0$, which is an example of this case, is indeed a saddle point.

Case I.2: $w_{21} = 0$, $w_{22} = 0$, $w_{11} = 0$ and $w_{12} = -2w_{02}$

Equations (4.6) and (4.7) result in

$$u + v_1 f(w_{01}) + v_2 f(w_{02}) = 0$$

and

$$u + v_1 f(w_{01}) + v_2 f(-w_{02}) = 0$$

implying that $v_2(2f(w_{02}) - 1) = 0$, thus $w_{02} = 0$, and thus $w_{12} = 0$. Thus Case I.2 is a special case of Case I.1.

Case I.3: $w_{21} = 0$, $w_{22} = 0$, $w_{11} = -2w_{01}$ and $w_{12} = 0$

Analogously to the previous case, this case is also a special case of Case I.1.

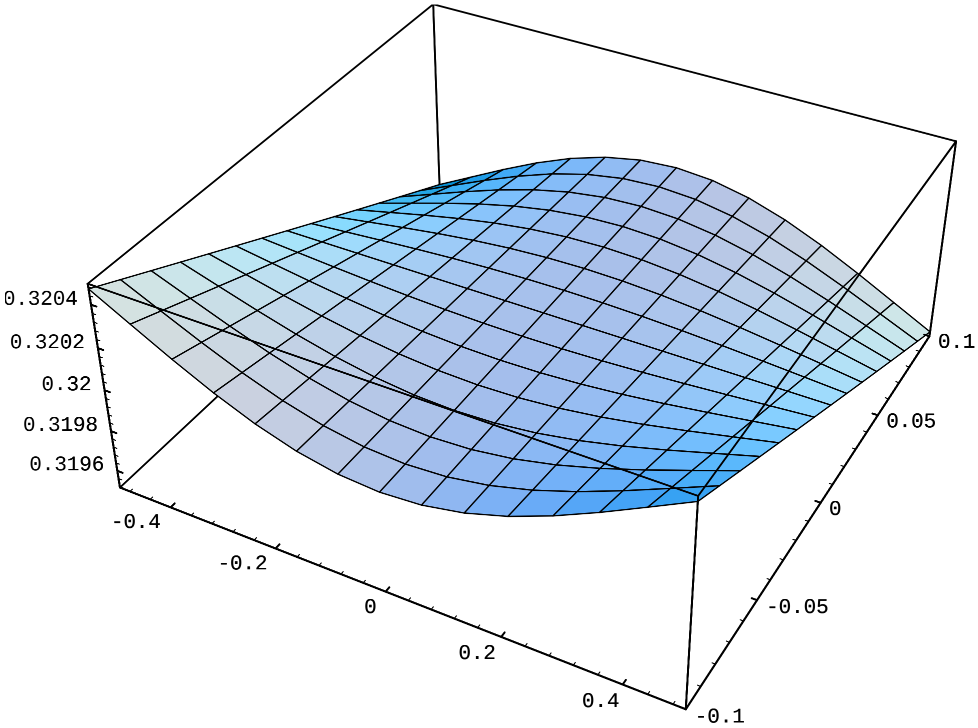

Figure 3. The neighbourhood of the weights $u = -1$, $v_1 = v_2 = 1$, $w_{01} = w_{11} = w_{21} = w_{02} = w_{12} = w_{22} = 0$. This picture is obtained by varying $-\Delta w_{11} = -\Delta w_{21} = \Delta w_{12} = \Delta w_{22}$ from –0.5 to 0.5 and $-2\Delta u = \Delta w_{01} = \Delta w_{02}$ from –0.1

Case I.4: $w_{21} = 0$, $w_{22} = 0$, $w_{11} = -2w_{01}$ and $w_{12} = -2w_{02}$

Equations (4.6) and (4.7) result in

$$u + v_1 f(w_{01}) + v_2 f(w_{02}) = 0$$

and

$$u + v_1 f(-w_{01}) + v_2 f(-w_{02}) = 0$$

implying that $v_1(2f(w_{01}) - 1) + v_2(f(w_{02}) - 1) = 0$. So choosing $w_{01} \neq 0$, $w_{02} \neq 0$ and v_1 ($\neq 0$) determines u and v_2. Thus these stationary points form a 3-dimensional region in the weight space.

To prove that these points are saddle points we introduce the variables x and y such that:

$$\Delta w_{01} = x, \Delta w_{21} = y, \Delta u = -v_1 f'(w_{01})x - \frac{1}{2}v_1 f'(w_{01})y$$

Thus we find for the inputs of the output unit for the 4 patterns, see equation (2.2):

$$A_{00} = u - v_1 f'(w_{01})x - \frac{1}{2}v_1 f'(w_{01})y + v_1 f(w_{01} + x) + v_2 f(w_{02})$$

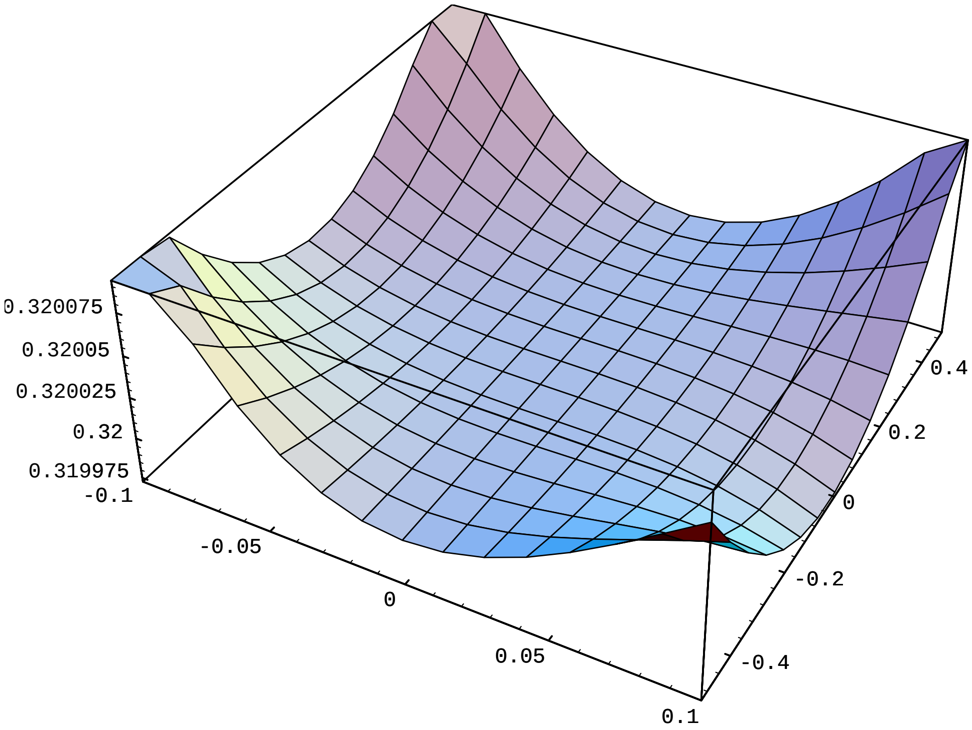

Figure 4. The neighbourhood of the same point in weight space as figure 3. This picture is obtained by varying $-2\Delta u = \Delta w_{01} = -\Delta w_{11} = -\Delta w_{21} = \Delta w_{02} = \Delta w_{12} = \Delta w_{22}$ from –0.1 to 0.1 and $\Delta v_1 = -\Delta v_2$ from –0.5 to 0.5.

$$A_{01} = u - v_1 f'(w_{01})\,x - \frac{1}{2} v_1 f'(w_{01})\,y + v_1 f(w_{01} + w_{21} + x + y) + v_2 f(w_{02} + w_{22})$$

$$A_{10} = u - v_1 f'(w_{01})\,x - \frac{1}{2} v_1 f'(w_{01})\,y + v_1 f(w_{01} + w_{11} + x) + v_2 f(w_{02} + w_{12})$$

$$A_{11} = u - v_1 f'(w_{01})\,x - \frac{1}{2} v_1 f'(w_{01})\,y + v_1 f(w_{01} + w_{11} + w_{21} + x + y) + v_2 f(w_{02} + w_{12} + w_{22})$$

For the second order derivatives of the error E with respect to x and y we find (using $w_{21} = w_{22} = 0$, $w_{11} = -2w_{01}$, $w_{12} = -2w_{02}$, $f'(x) = f'(-x)$ and $f''(x) = -f''(-x)$):

$$\left.\frac{\partial^2 E}{\partial x^2}\right|_0 = 0$$

and

$$\left.\frac{\partial^2 E}{\partial x \partial y}\right|_0 = (f(0) - 0.1) f'(0) \, (-2 v_1 f''(w_{01}))$$

Thus using theorem 2.1 it follows that the stationary points of case I.4 are saddle points if $f''(w_{01}) \neq 0$. The case that $w_{01} = 0$ (and thus also $w_{02} = 0$) is already considered in case I.1.

Conclusion 4.2 *All points of case I.4 are saddle points.*

4.2.2 Case II: w_{21} = 0 and w_{12} = 0

Equations (4.3) become in this case:

$$\begin{aligned} R_{01} &= -R_{00} = -\frac{f'(w_{02})}{f'(w_{02}+w_{22})} R_{00} \\ R_{10} &= -\frac{f'(w_{01})}{f'(w_{01}+w_{11})} R_{00} = -R_{00} \\ R_{11} &= \frac{f'(w_{01})}{f'(w_{01}+w_{11})} R_{00} = \frac{f'(w_{02})}{f'(w_{02}+w_{22})} R_{00} \end{aligned} \tag{4.26}$$

resulting in $R_{00} = -R_{01} = -R_{10} = R_{11}$, ($w_{22} = 0$ or $w_{22} = -2w_{02}$) and ($w_{11} = 0$ or $w_{11} = -2w_{01}$).

Thus analogously to case I we can split this case into the four cases:

- Case II.1: $w_{21} = 0$ and $w_{12} = 0$ and $w_{22} = 0$ and $w_{11} = 0$,
- Case II.2: $w_{21} = 0$ and $w_{12} = 0$ and $w_{22} = 0$ and $w_{11} = -2w_{01}$,
- Case II.3: $w_{21} = 0$ and $w_{12} = 0$ and $w_{22} = -2w_{02}$ and $w_{11} = 0$,
- Case II.4: $w_{21} = 0$ and $w_{12} = 0$ and $w_{22} = -2w_{02}$ and $w_{11} = -2w_{01}$.

Case II.1 is equal to case I.1; Case II.2 is equal to Case I.3 and thus a special case of Case I.1; Case II.3 is also a special case of Case I.1. So these cases result in saddle points. Let us consider Case II.4:

Case II.4: w_{21} = 0 and w_{12} = 0 and w_{22} = –2w_{02} and w_{11} = –2w_{01}

Equation (3.1) results in this case in:

$$\begin{aligned} R_{00} &= (f(u + v_1 f(w_{01}) + v_2 f(w_{02})) - 0.1) \cdot \\ &\quad f'(u + v_1 f(w_{01}) + v_2 f(w_{02})) \\ R_{01} &= (f(u + v_1 f(w_{01}) + v_2 f(-w_{02})) - 0.9) \cdot \\ &\quad f'((u + v_1 f(w_{01}) + v_2 f(-w_{02}))) \\ R_{10} &= (f(u + v_1 f(-w_{01}) + v_2 f(w_{02})) - 0.9) \cdot \\ &\quad f'(u + v_1 f(-w_{01}) + v_2 f(w_{02})) \end{aligned}$$

$$R_{11} = (f(u + v_1 f(-w_{01}) + v_2 f(-w_{02})) - 0.1) \cdot \\ f'(u + v_1 f(-w_{01}) + v_2 f(-w_{02}))$$

Combining this with $R_{00} = -R_{01} = -R_{10} = R_{11}$ and applying theorem 2.4 (using $f(x) = 1 - f(-x)$) with $a = u + v_1 f(w_{01}) + v_2 f(w_{02})$, $b = v_2(1 - 2f(w_{02}))$ and $c = v_1(1 - 2f(w_{01}))$ shows that there exist exactly 9 different solutions for (a, b, c). For each of the eight solution points not equal to $(0, 0, 0)$ we can chose $w_{02} \neq 0$ and $w_{01} \neq 0$, and then v_1, v_2 and u are determined by this choice. Thus these stationary points form 2-dimensional regions in the weight space. The corresponding error values are 0.786045 and 0.805872 (see [5]). The solution $a = b = c = 0$ results in $w_{01} = w_{02} = 0$ and thus $w_{22} = w_{11} = 0$ and is part of case I.1.

The proof that all stationary points of case II.4 are saddle points is completely equivalent to that of case I.4. The only difference is that in the second order derivatives of the error with respect to *x* and *y*, the factor $(f(0) - 0.1)f'(0)$ has to be replaced by the more general factor R_{00}.

Conclusion 4.3 *We have shown that all instable stationary points with* $v_1 \neq 0$ *and* $v_2 \neq 0$ *form regions of dimension at least 2 in the weight space. This implies that the Hessian matrix of the second order derivatives has at least two eigenvalues equal to zero. The Hessian can not be positive definite for these points. We have proved that all instable stationary points are saddle points.*

5 Conclusions

In this paper we investigated the error surface of the XOR network with two hidden nodes (see figure 1). We proved that stationary points of this error surface with finite weights are either absolute minima with error zero or saddle points. So no local minima exist for finite weights.

In this paper we used the quadratic error function

$$E = \frac{1}{2}\sum_{\alpha}(O_\alpha - t_\alpha)^2$$

All proofs hold also for the “cross-entropy” error function, used in [1, 2]:

$$L = -\sum_{\alpha} \ln\left((O_\alpha)^{t_\alpha}(1 - O_\alpha)^{1 - t_\alpha}\right)$$

which can be seen immediately by replacing the terms R_{ij} (see (3.1)) for the quadratic error function by the (simpler) terms $R_{ij}' = O_\alpha - t_\alpha$. Since all stationary points for L are stationary points for E it is clear that also the error surface for L will not result in local minima for finite values of the weights.

The stationary points from case II.4 for the error function E do not occur for the error L. However, the proof that these points are saddle points can be almost copied from other cases.

In a forthcoming paper [6] we will publish our results on stationary points for infinite values of the weights. We found that this network has regions with local minima for some weights from the inputs to the hidden nodes having infinite values. However, since boundary points of these regions are saddle points, a strictly decreasing path exists from each finite point in weight space to a point with error zero. In the neighbourhood of the found local minima learning algorithms can be trapped, as is the case for the point given by Rumelhart and McClelland [3] and four of the five points given by Lisboa and Perantonis [2].

6 References

[1] L.G.C. Hamey; *Analysis of the Error Surface of the XOR Network with Two Hidden Nodes*, Computing Report 95/167C, Department of Computing, Macquarie University, NSW 2109 Australia, 1995.

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

[3] D.E. Rumelhart, J.L. McClelland and the PDP Research Group; *Parallel Distributed Processing, Volume 1*. The MIT Press, Cambridge, Massachusetts, 1986.

[4] I.G. Sprinkhuizen-Kuyper and E.J.W. Boers; “Classification of all stationary points on a neural network error surface”. In: J.C. Bioch and S.H. Nienhuys-Cheng, eds., *Proceedings of Benelearn-94: 4th Belgian-Dutch Conference on Machine Learning*, pp. 192-201, Report EUR-CS-94-05, Dept. of Computer Science, Fac. of Economics, Erasmus University, Rotterdam, The Netherlands, June 1994. Also as Technical Report 94-19, Leiden University, Dept. of Computer Science, The Netherlands.

[5] I.G. Sprinkhuizen-Kuyper and E.J.W. Boers; *The Error Surface of the simplest XOR Network has no local Minima*. Technical Report 94-21, Leiden University, Dept. of Computer Science, The Netherlands.

[6] I.G. Sprinkhuizen-Kuyper and E.J.W. Boers; *The Error Surface of the XOR Network with two hidden Nodes*. To appear.