The Error Surface of the simplest XOR Network has no local Minima

I.G. Sprinkhuizen-Kuyper
and
E.J.W. Boers
Department of Computer Science
Leiden University
email: {kuyper,boers}@wi.leidenuniv.nl

Abstract

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, using e.g. backpropagation with momentum, on the four patterns of the XOR problem, the correct solution with error zero will be reached in the limit with probability one.

1 Introduction

A central theme in neural network research is to find the right network (architecture and learning algorithm) for a problem. Some learning algorithms also influence the architecture (pruning and construction, see e.g. [5, 7]). In our research [1, 2, 3] we are trying to generate good architectures for neural networks using a genetic algorithm which works on strings containing coded production rules of a graph grammar (L-systems). These production rules result in an architecture and training of the architecture on a given problem results in a fitness for the given string, which is used by the genetic algorithm. In order to be able to decide objectively which architecture is better, a distinction is made between the following three aspects:

- representation,
- learning and
- generalization.

The representation aspect considers whether a network is able to represent a solution of the problem. The learning aspect concerns the ability of a network to learn a solution of the problem. If the network is able to learn a solution, how fast, with what probability and how accurate will that solution be learned? The third aspect is whether the network is able to generalize, i.e. does the network give reasonable output for patterns that were not in the training set?

In order to learn more about these aspects we considered some simple networks for boolean functions. This paper is concerned with the simplest network that can represent the XOR function: one hidden unit and connections from the input units to the output unit (see figure 1). As training algorithm we take some gradient-based algorithm, e.g. backpropagation with momentum. 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. Often, an extra momentum term is added to average out different directions, which speeds up learning. So the update of each weight w_{ij} is given by: $\Delta w_{ij}(t) = -\alpha\, \partial E/\partial w_{ij} + \beta\, \Delta w_{ij}(t-1)$, with learning parameter α and momentum parameter β. This has the effect that the weights are updated such that a point on the error surface is reached with a smaller error value. Distinction can be made between batch learning and on-line learning. During batch learning the weights are updated after the whole training set is seen and the errors of the individual samples are summed to the total error, while during on-line learning the weights are corrected after each sample, with respect to the error for the sample just seen by the network.

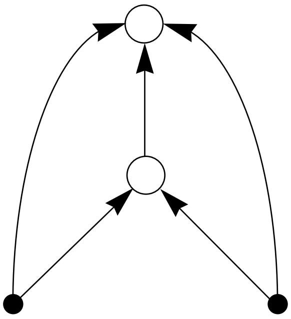

Figure 1. The simplest XOR network

1.1 Representation

First we looked at the representational power of the simplest XOR network. It is well known that this network with a treshold 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 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 (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). This result is not trivial, since for a one-layer network[1] for the AND function, it is possible to find an approximate representation, but it is not possible to solve the AND function exactly, using a sigmoid transfer function.

1.2 Learning

The next question that can be asked concerns learnability: under what conditions is a given network able to learn the desired function. When we assume that some kind of gradient-based learning algorithm is used, the shape of the error surface is very important. The ideal error surface has one minimum in weight space (ideally with error zero) corresponding to an acceptable solution and in each other point a nonzero gradient. With such an error surface each gradient-based learning algorithm will approximate the minimum, and so find a reasonable solution. However if the error surface has so-called local minima, then the learning algorithm can wind up in such a local minimum and reach a suboptimal solution. From experiments by Rumelhart et al. [9] it seems that the simplest XOR network does not have such local minima in contrast to the XOR network with two hidden units and without connections from the inputs to the output. The problem 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. 6, 7, 8, 9]. 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 [8] claim a local minimum, for example, for the XOR network with two hidden units and without connections from the inputs to the output, with the weights from the hidden units to the output unit equal to zero, while by similar techniques as used in this paper it can be shown that such a point is a saddle point and *not* a local minimum. 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 a 3-dimensional region in weight space exists with zero error. All points in a neighbourhood of each point of this region have error values which are *not*

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

less than the error in that point. In a *strict* minimum, however, it is necessary that all points in a neighbourhood 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 all 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.

For the standard XOR network with two hidden units, we already proved that it has a stable global minimum with error zero, and that other minima can not be stable. Results that on-line learning with a reasonably large learning parameter leads best to avoiding such minima [6], can be explained from this fact.

1.3 Generalization

The third point is the ability to generalize. The work of Denker, Schwartz, Solla et al. [4, 10, 11] suggests to investigate the a priori probability that a network represents a certain function when the weights are chosen randomly. We did some calculations for the XOR networks with treshold units and did some numerical experiments for the networks with a sigmoid transfer function, to determine the a priori probability of the network to represent (an approximation of) the XOR function, relatively to the probability of representing one of the other boolean functions of two inputs. Our results tell that this probability is very small (≈0.005 for the simplest network, and ≈0.0013 for the network with two hidden units). Thus, if less than four patterns of the XOR problem are used to train the network, almost always one of the other boolean functions corresponding to those patterns will be learned, and not the XOR function. However, this also suggests that more regular functions like AND and OR are preferred, if possible, above the XOR function. In a forthcoming paper we will publish our results of several measurements of a priori probabilities for several functions.

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. Section 3 contains some properties and equalities concerning the transfer function. In section 4 it is proven that a 3-dimensional region of the weight space exist with zero error. In section 5 it is proven that all finite points with nonzero error are unstable, i.e. the gradient of the error with respect to one single pattern is unequal to zero, and that local minima can occur only for finite values of the weights. Section 6 consists of the proof that all points with nonzero error and zero gradient (averaged for a training set) are saddle points. Finally section 7 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

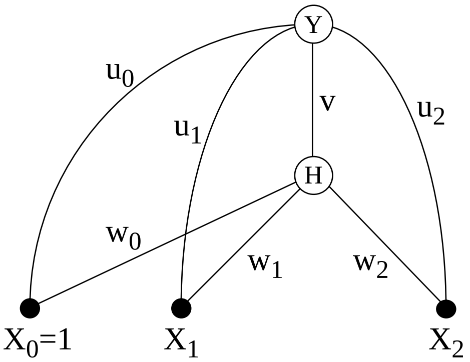

Figure 2. The simplest XOR network

The network in figure 2 with one hidden unit H is studied. This network consists of one treshold unit X_0, with constant value 1, two inputs X_1 and X_2, one hidden unit H and the output unit Y. There are seven weights which are labelled as follows:

- the weights w_i (i = 0..2) are the weights of the connections from the inputs X_i to the hidden unit H;
- the weights u_i (i = 0..2) belong to the connections from the inputs X_i to the output unit Y, and
- the weight v corresponds to the connection from the hidden unit H to the output unit Y.

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

$$\begin{aligned} y(X_1, X_2) &= f(u_0 + u_1X_1 + u_2X_2 + vf(w_0 + w_1X_1 + w_2X_2)) \\ &= f\left(\sum_{i=0}^{2} u_iX_i + vf\left(\sum_{j=0}^{2} w_jX_j\right)\right) \end{aligned} \tag{2.1}$$

Table 1 shows the patterns for the XOR problem which have to be learned. 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:

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

with $y(X_1,X_2)$ given in equation (2.1).

We will consider learning algorithms based on gradient-descent learning, both on-line and batch. During on-line learning the weights are updated after each training example, corresponding to the error for that example. With batch learning the updating of the weights is done after all training examples have been seen and all errors are summed. The weights will be adjusted proportional to $-\nabla E$ and the learning will end when $\nabla E = 0$.

Table 1: Patterns for the XOR problem

Pattern	**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

3 The transfer function *f*

This section contains formula's expressing properties of the sigmoid transfer function *f* and its derivatives needed in the remainder of this paper. Figure 3 below shows the shape of *f*, *f ′* and *f ″*.

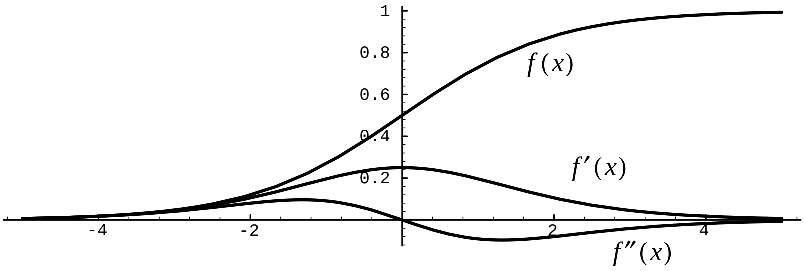

Figure 3. The transfer function and its derivatives

The transfer function used is:

$$f(x) = \frac{1}{1 + e^{-x}} \quad (3.1)$$

On the interval $(-\infty, \infty)$ this function is strictly monotonously increasing from 0 to 1. Hence

$$0 < f(x) < 1$$

$$\lim_{x \to -\infty} f(x) = 0$$

$$\lim_{x \to \infty} f(x) = 1$$

From the monotonicity of *f* it is clear that

$$f(a) = f(b) \Leftrightarrow a = b$$

Furthermore this function has the following properties:

$$f(-x) = 1 - f(x) \tag{3.2}$$

$$f'(x) = f(x)\,(1 - f(x)) \tag{3.3}$$

$$f''(x) = f'(x)\,(1 - 2f(x)) = f(x)\,(1 - f(x))\,(1 - 2f(x))$$

$$0 < f'(x) \le \frac{1}{4}$$

$$\lim_{x \to -\infty} f'(x) = \lim_{x \to \infty} f'(x) = 0$$

$$f'(0) = \frac{1}{4}$$

$$f'(x) = f'(-x) \tag{3.4}$$

$$f''(x) = 0 \Leftrightarrow x = 0 \tag{3.5}$$

$$f'''(0) \ne 0 \tag{3.6}$$

The derivative $f'(x)$ is strictly monotonously increasing on $(-\infty,0]$, and strictly monotonously decreasing on $[0,\infty)$, thus using the symmetry (3.12) gives:

$$f'(a) = f'(b) \Leftrightarrow a = b \vee a = -b \tag{3.7}$$

The function *f* has an inverse function:

$$f^{-1}(x) = \log\left(\frac{x}{1-x}\right) \quad \text{if } 0 < x < 1 \tag{3.8}$$

4 The minimum *E* = 0 can occur

In this section it is shown that a 3-dimensional region in the 7-dimensional weight space exists for which the error is exactly zero. The error *E* consists of four quadratic terms, so *E* = 0 holds only if all terms are zero. The four equations for the weights thus obtained are considered. From these equations four linear equations for the three weights u_0, u_1 and u_2 in terms of the other weights are found. It is shown that for almost all values of the three weights w_0, w_1 and w_2 it is possible to find a value of *v* such that the equations for u_0, u_1 and u_2 have a (unique) solution. This results in a 3-dimensional region depending on w_0, w_1 and w_2. We will distinguish two kinds of minima for the error *E*:

- Minima that remain 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 chance of the weights. These minima will be called *stable minima*.
- Minima that depend on the given training set. For batch learning this is a minimum, but 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 u_0, u_1, u_2, w_0, w_1, w_2 and v exist such that the following four equations hold:

$$\begin{aligned} f(u_0 + vf(w_0)) &= 0.1 \\ f(u_0 + u_2 + vf(w_0 + w_2)) &= 0.9 \\ f(u_0 + u_1 + vf(w_0 + w_1)) &= 0.9 \\ f(u_0 + u_1 + u_2 + vf(w_0 + w_1 + w_2)) &= 0.1 \end{aligned} \tag{4.1}$$

Application of the inverse function f^{-1} on both sides of these equations leads to:

$$\begin{aligned} u_0 + vf(w_0) &= f^{-1}(0.1) \approx -2.197 \\ u_0 + u_2 + vf(w_0 + w_2) &= f^{-1}(0.9) \approx 2.197 \\ u_0 + u_1 + vf(w_0 + w_1) &= f^{-1}(0.9) \approx 2.197 \\ u_0 + u_1 + u_2 + vf(w_0 + w_1 + w_2) &= f^{-1}(0.1) \approx -2.197 \end{aligned} \tag{4.2}$$

We will show below that for each value of the weights w_0, w_1 and w_2 where

$$f(w_0) - f(w_0 + w_1) - f(w_0 + w_2) + f(w_0 + w_1 + w_2) \neq 0 \tag{4.3}$$

unique values of the other weights u_0, u_1, u_2 and v can be found such that all equations of (4.2) hold. Let us first investigate the equation:

$$f(w_0) - f(w_0 + w_1) - f(w_0 + w_2) + f(w_0 + w_1 + w_2) = 0 \tag{4.4}$$

We calculated by using equation (3.1) and by substituting temporarily p_i for e^{-w_i} (i = 0..2) that this equation is equivalent to:

$$\begin{aligned} &\frac{1}{1+p_0} - \frac{1}{1+p_0p_1} - \frac{1}{1+p_0p_2} + \frac{1}{1+p_0p_1p_2} = \\ &\frac{p_0(p_1-1)(p_2-1)(p_0^2p_1p_2-1)}{(1+p_0)(1+p_0p_1)(1+p_0p_2)(1+p_0p_1p_2)} = \\ &\frac{e^{-w_0}\left(e^{-w_1}-1\right)\left(e^{-w_2}-1\right)\left(e^{-2w_0-w_1-w_2}-1\right)}{\left(1+e^{-w_0}\right)\left(1+e^{-w_0-w_1}\right)\left(1+e^{-w_0-w_2}\right)\left(1+e^{-w_0-w_1-w_2}\right)} = 0 \end{aligned} \tag{4.5}$$

Since $e^x > 0$ equation (4.5) has the solutions:

$$w_1 = 0 \quad \text{or} \quad w_2 = 0 \quad \text{or} \quad 2w_0 + w_1 + w_2 = 0 \tag{4.6}$$

So equation (4.3) holds everywhere with exception from the three hyperplanes given in (4.6). The equations (4.2) are 4 linear equations in the weights u_0, u_1 and u_2. In order that these equations have a solution they have to be linearly dependent. This leads to the condition

$$v(f(w_0) - f(w_0 + w_1) - f(w_0 + w_2) + f(w_0 + w_1 + w_2)) = -4f^{-1}(0.9)$$

Thus given w_0, w_1 and w_2 such that (4.3) holds, we find for v:

$$v = \frac{-4f^{-1}(0.9)}{f(w_0) - f(w_0 + w_1) - f(w_0 + w_2) + f(w_0 + w_1 + w_2)}$$

and u_0, u_1 and u_2 can be uniquely solved from the first three equations in (4.2). Since the inequality (4.3) holds for all points w_0, w_1, w_2, which are not on the hyperplanes given in (4.6), we will find a 3-dimensional region in the 7-dimensional weight space, where $E = 0$. 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 since there are always points in a neighbourhood of a point with $E = 0$ where the error is also equal to zero. It is clear that $E = 0$ is a global minimum, since for all points E ≥ 0 holds, E being a positive sum of quadratic terms.

5 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. Consideration of the derivative of the error with respect to u_0 for finite values of the input of the output unit (thus the output is not equal to zero or one) shows that all patterns have to be learned exactly in this case, leading to an error value of zero, which is the absolute minimum. The derivative of the error with respect to u_0 can also go to zero if the output goes to zero or one for one or more patterns and the other patterns are learned exactly. It is shown that these cases do not result in a minimum, so the only stable minimum is the minimum with error zero for all patterns.

Let us consider the partial derivative of E with respect to u_0. Writing R_{ij} for the terms depending on pattern P_{ij} we obtain:

$$\frac{\partial E}{\partial u_0} = R_{00} + R_{01} + R_{10} + R_{11} \tag{5.1}$$

with

$$
\begin{aligned}
R_{00} &= a_{00}\,(f(u_0 + vf(w_0)) - 0.1)\,f'(u_0 + vf(w_0)) \\
R_{01} &= a_{01}\,(f(u_0 + u_2 + vf(w_0 + w_2)) - 0.9)\,f'(u_0 + u_2 + vf(w_0 + w_2)) \\
R_{10} &= a_{10}\,(f(u_0 + u_1 + vf(w_0 + w_1)) - 0.9)\,f'(u_0 + u_1 + vf(w_0 + w_1)) \\
R_{11} &= a_{11}\,(f(u_0 + u_1 + u_2 + vf(w_0 + w_1 + w_2)) - 0.1)\,\cdot \\
&\quad f'(u_0 + u_1 + u_2 + vf(w_0 + w_1 + w_2))
\end{aligned}
\tag{5.2}
$$

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

$$
R_{00} = R_{01} = R_{10} = R_{11} = 0\,. \tag{5.3}
$$

So all stable stationary points satisfy (5.3). The condition (5.3) is not only a necessary condition for a stable stationary point, but it is also sufficient, since if it holds then the partial derivatives of E with respect to the other weights will be zero too. Clearly the points such that the equations (4.1) hold and thus 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 derivative of the transfer function in (5.3) (see also (5.2)) approach $\pm\infty$. The corresponding outputs go to zero or one. We will show that if such a point is approached, it is always possible to leave the neighbourhood of such a point via a path with decreasing error.

Stationary points with output 0 or 1 for one or more patterns

First let us consider the case that three of the patterns are learned exactly in the limit and the fourth pattern has output 0 or 1. So consider e.g.:

$$
\begin{aligned}
u_0 + vf(w_0) &= q_{00} \to f^{-1}(0.1) \\
u_0 + u_2 + vf(w_0 + w_2) &= q_{01} \to f^{-1}(0.9) \\
u_0 + u_1 + vf(w_0 + w_1) &= q_{10} \to f^{-1}(0.9) \\
u_0 + u_1 + u_2 + vf(w_0 + w_1 + w_2) &= q_{11} \to \pm\infty
\end{aligned}
\tag{5.4}
$$

Since these equations in u_0, u_1 and u_2 are linearly dependent, it follows that

$$
\begin{aligned}
& v\,(f(w_0) - f(w_0 + w_1) - f(w_0 + w_2) + f(w_0 + w_1 + w_2)) = \\
& q_{00} - q_{01} - q_{10} + q_{11} \to \pm\infty
\end{aligned}
$$

in a point in the neighbourhood of the stationary point. Thus in such a point condition (4.3) holds. If $q_{11} < f^{-1}(0.1)$ or $q_{11} > 0$, then decreasing v in absolute value, while keeping w_0, w_1 and w_2 constant, and changing u_0, u_1 and u_2 such that the first three patterns keep resulting in the same output, will have the result that the error of the fourth pattern decreases, while those of the other patterns remain constant. So the total error decreases and a stationary point as given by the equations (5.4) is not a minimum. The same argument holds if the output of one of the other patterns goes to zero or one.

When more outputs are going to zero or one a similar argument can be given as long as condition (4.3) holds. So we have yet to consider the cases where it is not obvious that this condition holds.

For example consider the case that $q_{00} \to f^{-1}(0.1)$, $q_{01} \to f^{-1}(0.9)$, $q_{10} \to \pm\infty$ and $q_{11} \to \pm\infty$. The equations (5.4) make clear, that for $q_{10} > f^{-1}(0.9)$ and $q_{11} > f^{-1}(0.1)$ decreasing the value of u_1, while keeping the other weights constant, will result in a lower error value for both P_{10} and P_{11}. If $q_{10} < f^{-1}(0.9)$ and $q_{11} < f^{-1}(0.1)$ then increasing the value of u_2 leads to a decreasing total error too.

If $q_{00} \to \infty$, $q_{01} \to f^{-1}(0.9)$, $q_{10} \to f^{-1}(0.9)$ and $q_{11} \to -\infty$, then the error is decreased by decreasing u_0 and increasing u_1 and u_2 equally, such that u_0+u_1 and u_0+u_2 remain constant.

The other cases with two patterns leading to an output 0 or 1 are treated similar. If three patterns have an output in the neighbourhood of 0 or 1 the weights can similarly be adjusted to reach points with lower error values. The cases where all four patterns give an output almost 0 or 1 can only be reached via a path with increasing error: these points are (local) maxima.

Conclusion: *The unique stable minimum for the considered network for the XOR problem is a 3-dimensional region in weight space with E = 0.*

6 All unstable stationary points are saddle points

In this section it is proved that all unstable stationary points are saddle points. Examination of the equations for $\nabla E = 0$ leads to three equations (6.8), (6.9) and (6.10) which have to be satisfied by the considered points. The proof is separated into the cases where the weight $v = 0$ and the cases where $v \neq 0$. In the cases where $v = 0$ all partial derivatives of the error with respect to w_0, w_1 and w_2 are zero. It is proved that the first partial derivative of the error of the form $\partial^{i+j+1}E/\partial w_1{}^i \partial w_2{}^j \partial v$ which is unequal to zero determines that these points are saddle points. The cases where $v \neq 0$ are solved by considering the behaviour of the error on some carefully selected curves. Also some pictures are added showing some of the saddle points, visualized with Mathematica.

We have to investigate all points in the weight space with $\nabla E = 0$, not treated in the previous section. The components of ∇E are:

$$\frac{\partial E}{\partial u_0} = R_{00} + R_{01} + R_{10} + R_{11} \tag{6.1}$$

$$\frac{\partial E}{\partial u_1} = R_{10} + R_{11} \tag{6.2}$$

$$\frac{\partial E}{\partial u_2} = R_{01} + R_{11} \tag{6.3}$$

$$\begin{aligned}\frac{\partial E}{\partial w_0} = {} & R_{00} v f'(w_0) + R_{01} v f'(w_0 + w_2) + R_{10} v f'(w_0 + w_1) + \\ & R_{11} v f'(w_0 + w_1 + w_2)\end{aligned} \tag{6.4}$$

$$\frac{\partial E}{\partial w_1} = R_{10} v f'(w_0 + w_1) + R_{11} v f'(w_0 + w_1 + w_2) \tag{6.5}$$

$$\frac{\partial E}{\partial w_2} = R_{01} v f'(w_0 + w_2) + R_{11} v f'(w_0 + w_1 + w_2) \tag{6.6}$$

$$\frac{\partial E}{\partial v} = R_{00} f(w_0) + R_{01} f(w_0 + w_2) + R_{10} f(w_0 + w_1) + R_{11} f(w_0 + w_1 + w_2) \tag{6.7}$$

If $\nabla E = 0$ then it is concluded from equations (6.1), (6.2) and (6.3) that

$$R_{00} = -R_{01} = -R_{10} = R_{11} \tag{6.8}$$

From equations (6.4), (6.5), (6.6) and (6.8) it follows that

$$\begin{aligned}& R_{00} v f'(w_0) = R_{00} v f'(w_0 + w_2) = \\ & R_{00} v f'(w_0 + w_2) = R_{00} v f'(w_0 + w_1 + w_2)\end{aligned} \tag{6.9}$$

Equation (6.7), finally, leads together with equation (6.8) to:

$$R_{00} \left(f(w_0) - f(w_0 + w_2) - f(w_0 + w_1) + f(w_0 + w_1 + w_2) \right) = 0 \tag{6.10}$$

Since we are looking for unstable minima, we only have to consider here the case $R_{00} \neq 0$. (The cases where $R_{00} = 0$ are already considered in the previous section.) Equations (6.9) and (6.10) then simplify to the following equations:

$$v f'(w_0) = v f'(w_0 + w_2) = v f'(w_0 + w_1) = v f'(w_0 + w_1 + w_2) \tag{6.11}$$

and

$$f(w_0) - f(w_0 + w_2) - f(w_0 + w_1) + f(w_0 + w_1 + w_2) = 0 \tag{6.12}$$

So we will investigate all points satisfying equations (6.8), (6.11) and (6.12), which in addition are such that $E \neq 0$, in order to prove that all points where $\nabla E = 0$ and where no stable minimum is attained, are saddle points. Remark that equation (6.12) is identical to equation (4.4) and has the solutions $w_1 = 0$ or $w_2 = 0$ or $2w_0 + w_1 + w_2 = 0$. From equation (6.11) it is clear that it makes sense to distinguish between points where $v = 0$ and points where $v \neq 0$.

6.1 The case $v = 0$

In this case equations (6.8), (6.11) and (6.12) lead to:

$$\begin{aligned} &R_{00} = a_{00}(f(u_0) - 0.1)f'(u_0) = \\ &-a_{01}(f(u_0+u_1) - 0.9)f'(u_0+u_1) = \\ &-a_{10}(f(u_0+u_2) - 0.9)f'(u_0+u_2) = \\ &a_{11}(f(u_0+u_1+u_2) - 0.1)f'(u_0+u_1+u_2) \neq 0 \end{aligned} \tag{6.13}$$

$$f(w_0) - f(w_0+w_2) - f(w_0+w_1) + f(w_0+w_1+w_2) = 0 \tag{6.14}$$

If $a_{00} = a_{01} = a_{10} = a_{11} = 1$ then it follows from equations (3.2) and (3.4) that equation (6.13) is equivalent to:

$$\begin{aligned} &R_{00} = (f(u_0) - 0.1)f'(u_0) = \\ &(f(-u_0-u_1) - 0.1)f'(-u_0-u_1) = \\ &(f(-u_0-u_2) - 0.1)f'(-u_0-u_2) = \\ &(f(u_0+u_1+u_2) - 0.1)f'(u_0+u_1+u_2) \neq 0 \end{aligned} \tag{6.15}$$

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, 0.407392 and 0.403321 (see the remark after theorem A.1). From theorem A.1 it is also clear that $R_{00} > 0$ and that $E = 0$ cannot occur if $v = 0$.

Let us consider the behaviour of the error in the neighbourhood of a point with $v = 0$ satisfying (6.13) and (6.14) for small variations of w_1 and v (the other weights are kept constant). Considering $\partial E/\partial w_1$ (equation (6.5)), it is clear that each term contains a factor v, which will not disappear by taking the partial derivative with respect to w_1 again. Thus it is clear that also $\partial^2 E/\partial w_1^2 = 0$. Computation of $\partial^2 E/\partial w_1 \partial v$, using equation (6.13) results in:

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

Hence $\partial^2 E/\partial w_1 \partial v \neq 0$ if $f'(w_0+w_1) \neq f'(w_0+w_1+w_2)$. Given equation (3.7), this holds if and only if $w_2 \neq 0$ and $w_2 \neq -2w_0-2w_1$. From theorem A.2, which is given and proved in the appendix, with $a = w_1$ and $b = v$ it follows that E attains a saddle point if $\partial^2 E/\partial w_1 \partial v \neq 0$.

Conclusion: *If $E \neq 0$, $\nabla E = 0$, $w_2 \neq 0$ and $w_2 \neq -2w_0-2w_1$ then E has a saddle point, and not a minimum.*

From symmetry with respect to w_1 and w_2 we conclude that:

Conclusion: *If $E \neq 0$, $\nabla E = 0$, $w_1 \neq 0$ and $w_1 \neq -2w_0-2w_2$ then E has a saddle point, and not a minimum.*

The remaining case that has to be investigated if $v = 0$ is thus the case with $E \neq 0$, $\nabla E = 0$, ($w_2 = 0$ or $w_2 = -2w_0-2w_1$) and ($w_1 = 0$ or $w_1 = -2w_0-2w_2$). In order to solve this case we will consider E as function of w_1, w_2 and v (keeping w_0, u_0, u_1 and u_2 constant). If $v = 0$ all partial derivatives of E of the form $\partial^{i+j}E/\partial w_1{}^i \partial w_2{}^j$ with $i+j > 0$ are zero because of at least one factor v. Also $\partial^2 E/\partial w_1 \partial v = \partial^2 E/\partial w_2 \partial v = 0$. But computation of the third order derivative $\partial^3 E/\partial w_1 \partial w_2 \partial v$ leads to:

$$\partial^3 E/\partial w_1 \partial w_2 \partial v\Big|_{v=0} = R_{00} f''(w_0 + w_1 + w_2)$$

So because of equation (3.5) clearly $\partial^3 E/\partial w_1 \partial w_2 \partial v \neq 0$ if $v = 0$ and $w_0+w_1+w_2 \neq 0$. If, however, ($w_2 = 0$ or $w_2 = -2w_0-2w_1$) and ($w_1 = 0$ or $w_1 = -2w_0-2w_2$) and $w_0+w_1+w_2 = 0$ then $w_0 = w_1 = w_2 = 0$. From theorem A.3 (see appendix) with $a = w_1$, $b = w_2$ and $c = v$ it follows thus that all cases where $\nabla E = 0$ and $v = 0$ are saddle points except for the case $w_0 = w_1 = w_2 = 0$, which has to be studied further.

This last case is decided by considering

$$\partial^4 E/\partial w_1^2 \partial w_2 \partial v\Big|_{v = w_0 = w_1 = w_2 = 0} = R_{00} f'''(0) \neq 0 \tag{6.16}$$

Application of equation (3.6) and theorem A.4 proofs that even this point is a saddle point. 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.

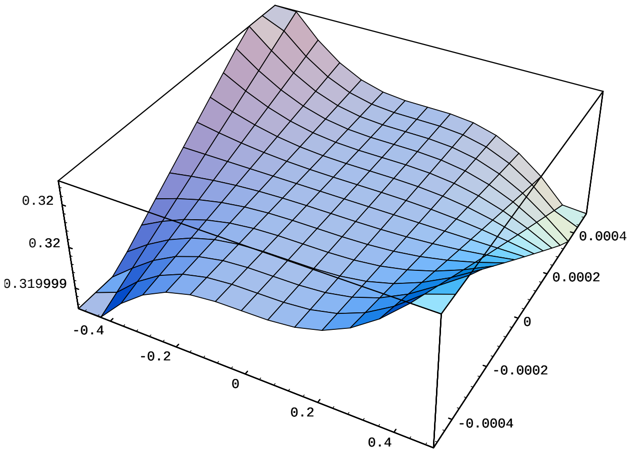

Figure 4. The error surface in the neighbourhood of $u_0 = u_1 = u_2 = w_0 = w_1 = w_2 = v = 0$. This picture is obtained by varying w_0, w_1 and w_2 equally from –0.5 to 0.5 and v from –0.0005 to 0.0005.

So we conclude that the case $v = 0$ can only lead to saddle points, since $E = 0$ is not possible if $v = 0$. Thus we have proved the following theorem:

Theorem 6.0 *If $v = 0$ then all points where $\nabla E = 0$ are saddle points.*

6.1 The case *v* unequal to zero

If $E \neq 0$, $\nabla E = 0$ and $v \neq 0$, equation (6.11) leads to:

$$f'(w_0) = f'(w_0 + w_2) = f'(w_0 + w_1) = f'(w_0 + w_1 + w_2) \tag{6.17}$$

and we have to consider solutions of equations (6.8), (6.12) and (6.17). The solutions of equation (6.12) are given by $w_1 = 0$ or $w_2 = 0$ or $w_1 + w_2 + 2w_0 = 0$. Substituting these solutions in equations (6.17) and applying the relation (3.7) results in the following four cases satisfying both (6.17) and (6.12):

- *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 first three cases lead to possibilities to attain points in the neighbourhood with smaller values for E. The fourth case follows directly from the third case by using the symmetry in w_1 and w_2. Since it is also easy to show that points exist in the neighbourhood of such points with greater values, it follows that these points are saddle points (and no extremes).

In order to prove that the points corresponding to cases 1 to 3 are saddle points, we started to investigate the stationary points corresponding to case 1. If $a_{00} = a_{01} = a_{10} = a_{11} = 1$, the points with $w_0 = w_1 = w_2 = u_1 = u_2 = 0$ and $u_0 = -vf(0)$ belong to this case, since for these points also equation (6.8) holds. For this special case we found the following expression for the second order part of the Taylor series expansion of the error E:

$$\begin{aligned} \Delta E / \{f'(0)\}^2 \approx {} & (\Delta u_1 + vf'(0)\Delta w_1)^2 + (\Delta u_2 + vf'(0)\Delta w_2)^2 + \\ & (2\Delta u_0 + \Delta u_1 + \Delta u_2 + 2vf'(0)\Delta w_0 + vf'(0)\Delta w_1 + vf'(0)\Delta w_2 + 2f(0)\Delta v)^2 \end{aligned} \tag{6.18}$$

This second order part contains three quadratic terms, but that is not enough to prove that E has a minimum here; the Hessian is not positive definite. Contrarily we looked for and indeed found ways to prove that E has a saddle point here.

Inspired by (6.18) we investigated the error surface for all stationary points of cases 1 to 3 by considering curves in the weight space through those points 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 $\Delta u_0 + vf'(w_0)\Delta w_0 = 0$. Finally we found the following three curves, which together proof that E has saddle points for all stationary points in the considered cases 1, 2 and 3.

Curve 1: This curve is parametrized by x such that $\Delta w_0 = x$, $\Delta w_2 = -x$, $\Delta u_0 = \alpha x$, and $\Delta u_2 = -\alpha x$, while $\alpha = -vf'(w_0)$. Using equation (2.2) this leads to:

$$\begin{aligned} E = {} & \frac{1}{2}a_{00}(f(u_0 + \alpha x + vf(w_0 + x)) - 0.1)^2 + \\ & \frac{1}{2}a_{01}(f(u_0 + u_2 + vf(w_0 + w_2)) - 0.9)^2 + \\ & \frac{1}{2}a_{10}(f(u_0 + u_1 + \alpha x + vf(w_0 + w_1 + x)) - 0.9)^2 + \\ & \frac{1}{2}a_{11}(f(u_0 + u_1 + u_2 + vf(w_0 + w_1 + w_2)) - 0.1)^2 \end{aligned} \tag{6.19}$$

Calculation using equation (6.8) results in:

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

Thus for the stationary points of case 3 ($w_1 = -2w_0$) the sign of $\partial^2 E/\partial x^2$ on curve 1 is equal to the sign of $vf''(w_0)$ (note that $R_{00} > 0$ because of theorem A.1). So if $vf''(w_0) < 0$ in case 3 then E has points in the neighbourhood with smaller values, since if in a point the first derivative of a function is zero and the second derivative is negative, this function attains a maximum in such a point.

Curve 2: This curve is parametrized by x such that $\Delta w_1 = x$, $\Delta w_2 = -x$, $\Delta u_1 = \alpha x$, and $\Delta u_2 = -\alpha x$, while $\alpha = -vf'(w_0)$. This leads analogously to:

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

In case 2 ($w_1 = w_2 = 0$) the sign of $\partial^2 E/\partial x^2$ on curve 2 is equal to the sign of $-vf''(w_0)$. So if $vf''(w_0) > 0$ in case 2 then E has points in the neighbourhood with smaller values.

Curve 3: This curve is parametrized by x such that $\Delta w_1 = \Delta w_2 = x$, $\Delta u_1 = \Delta u_2 = \alpha x$, where $\alpha = -vf'(w_0)$. This leads analogously to:

$$\left.\frac{\partial^2 E}{\partial x^2}\right|_{x=0} = R_{00}v(-f''(w_0 + w_1) - f''(w_0 + w_2) + 4f''(w_0 + w_1 + w_2))$$

In case 3 ($w_2 = 0$, $w_1 = -2w_0$) the sign of $\partial^2 E/\partial x^2$ on curve 3 is equal to the sign of $-vf''(w_0)$. So if $vf''(w_0) > 0$ in case 3 then E has points in the neighbourhood with smaller values. Combination with the results from curve 1, tells us that in case 3 always points can be found where the error becomes smaller.

In case 2 ($w_1 = w_2 = 0$) the sign of $\partial^2 E/\partial x^2$ on curve 3 is equal to the sign of $vf''(w_0)$. So if $vf''(w_0) < 0$ in case 2 then E has points in the neighbourhood with smaller values. This completes the proof that case 2 is not a minimum. In figure 5 one of these saddle points is visualized: the downward bow is obtained from the parametrizing of curve 3, the upward bow is inspired by equation (6.18).

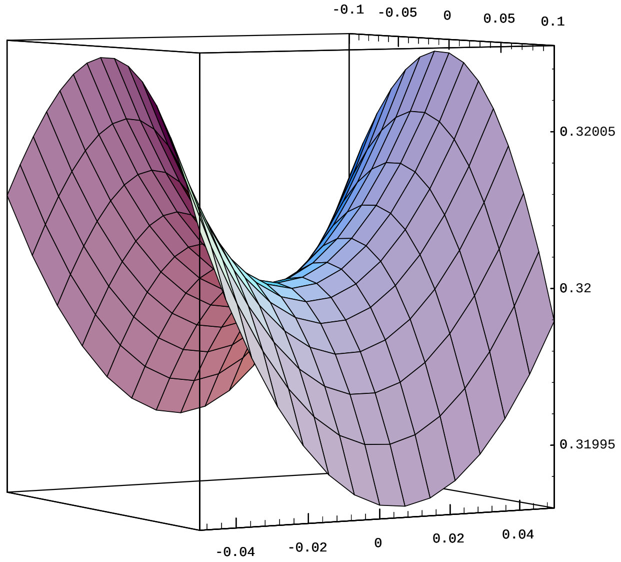

Figure 5. The error surface in the neighbourhood of the point $u_0 = -f(0.5)$, $u_1 = u_2 = 0$, $w_0 = 0.5$, $w_1 = w_2 = 0$, $v = 1$. The downward bow of the saddle is obtained by varying w_1, w_2, u_1 and u_2 such that $\Delta u_1 = \Delta u_2 = -f'(0)\Delta w_1 = -f'(0)\Delta w_2$. The other direction is given by varying w_0 and u_0 such that $\Delta u_0 = f'(0)\Delta w_0$.

The remaining case that has to be investigated is case 1 with $w_0 = w_1 = w_2 = 0$. Since $f''(0) = 0$ it follows that in this case on curve 3:

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

Computation of the third derivative on curve 3 results in:

$$\left.\frac{\partial^3 E}{\partial x^3}\right|_{x=0} = 6R_{00}vf'''(0)$$

This is unequal to zero and thus also in case 1 it is possible to find a direction to obtain lower values for E. One of these saddle points is shown in figure 6. In figure 7 it is shown that if the error surface is only considered in the direc-

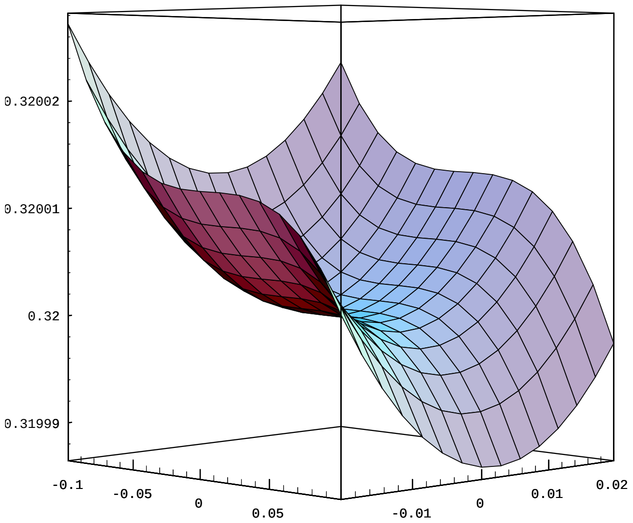

Figure 6. The saddle point in the neighbourhood of $u_0 = -f(0)$, $u_1 = u_2 = 0$, $w_0 = w_1 = w_2 = 0$ and $v = 1$. This picture is obtained by plotting the error against $u_1 = u_2 = -f'(0)w_1 = -f'(0)w_2$ and $u_0 = f'(0)w_0$. The weight w_1 runs from –0.1 to 0.1 and the weight w_0 runs from –0.02 to 0.02.

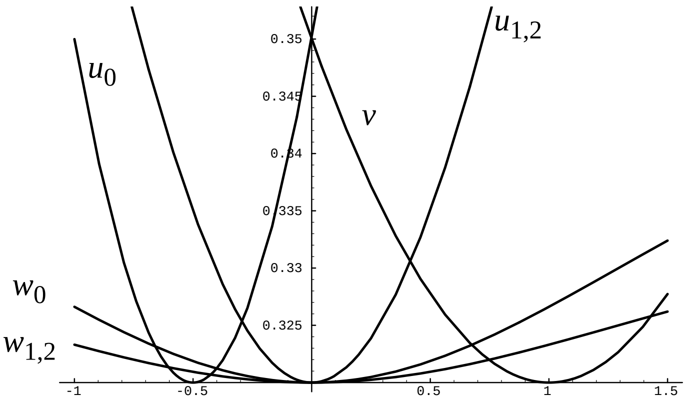

Figure 7. The error as function of each of the weights in the neighbourhood of $u_0 = -0.5$, $u_1 = u_2 = 0$, $w_0 = w_1 = w_2 = 0$ and $v = 1$. The curves for w_1 and w_2 and those for u_1 and u_2 are identical. This picture gives the (false) impression that the error has a local minimum if $u_0 = -0.5$, $u_1 = u_2 = 0$, $w_0 = w_1 = w_2 = 0$ and $v = 1$. Figure 6 showed already that this point is a saddle point.

tion of each of the weights it is suggested that such a point is a local minimum. So it is essential to vary the weights in the right combination, as is done in figure 6 in order to be able to conclude that this point is a saddle point.

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

Theorem 6.1 *If $E \neq 0$ and $v \neq 0$, then all points where $\nabla E = 0$ are saddle points.*

7 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 attained in a 3-dimensional region of the 7-dimensional weight space. Also a number of low dimensional regions exist where the error surface behaves as a saddle point (dimension 2 for the case $v = 0$, and dimension 1 for the other cases). The levels of the error surface in the saddle points are 0.32, 0.407392 and 0.403321, 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 at most 2, so its volume as part of the 7-dimensional weight space is zero.

When a saddle point is encountered, a batch learning process with zero momentum term can wind up in such a saddle point, but an on-line learning process can probably escape from such a saddle 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. Using batch learning no progress was made to escape from the saddle point.

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. Also the shape of the error surface at a minimum (narrow or wide) might determine how easy it is to escape from this minimum. Further research is necessary to examine this.

Another possible use of looking at the shape of the error surface when an exact representation of a problem is not possible (e.g. when noise is present), is finding an estimator for the generalization of a given weight configuration. We expect that very narrow minima will show a worse generalization than very wide minima with the same residual error on the training patterns. Further research is necessary to find a good measure for the shape of a minimum (especially considering the fact that several scaling factors are present) and to obtain experimental and theoretical results in this direction.

In this paper we used the quadratic error function

$$E = 1/2\sum_{\alpha}\left(y\left(X_1^{\alpha}, X_2^{\alpha}\right) - t^{\alpha}\right)^2$$

where α is the index of the pattern and t^{α} is the desired output. In the literature [e.g. 7, 8] also the error function

$$E' = \sum_{\alpha} t^{\alpha}\log\left(\frac{y\left(X_1^{\alpha}, X_2^{\alpha}\right)}{t^{\alpha}}\right) + \left(1 - t^{\alpha}\right)\log\left(\frac{1 - y\left(X_1^{\alpha}, X_2^{\alpha}\right)}{1 - t^{\alpha}}\right)$$

is used. All computations needed to prove that the points with $E \neq 0$ are saddle points also hold when using E' instead of E. Thus also with this error function it is true that only one global minimum value of the error exists. The only difference 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.

References

[1] E.J.W. Boers and H. Kuiper; *Biological Metaphors and the Design of Modular Artificial Neural Networks*, Master’s Thesis, Department of Computer Science, Leiden University, 1992.

[2] E.J.W. Boers, H. Kuiper, B.L.M. Happel and I.G. Sprinkhuizen-Kuyper; “Biological metaphors in designing modular artificial neural networks”. In: S. Gielen and B. Kappen (eds.); *Proceedings of the International Conference on Artificial Neural Networks*, Springer-Verlag, Berlin, 1993.

[3] E.J.W. Boers, H. Kuiper, B.L.M. Happel and I.G. Sprinkhuizen-Kuyper; “Designing Modular Artificial Neural Networks”. In: H.A. Wijshoff (ed.); *Proceedings of Computing Science in the Netherlands CSN’93*, pp. 87–96, 1993.

[4] J. Denker, D. Schwartz, B. Wittner, S. Solla, R. Howard, L. Jackel and J. Hopfield; “Large Automatic Learning, Rule Extraction, and Generalization”. *Complex Systems 1*, pp. 877–922, 1987.

[5] S.E. Fahlman and C. Lebiere; “The Cascade-Correlation Learning Architecture”. In: D.S. Touretzky (ed.); *Advances in Neural Information Processing Systems II,* Morgan Kaufmann, San Mateo, pp. 542–532, 1989.

[6] D. Gorse, A. Shepherd and J.G. Taylor; “Avoiding Local Minima by Progressive Range Expansion”. In: S. Gielen and B. Kappen (eds.); *Proceedings of the International Conference on Artificial Neural Networks*, Springer-Verlag, Berlin, 1993. (added)

[7] J. Herz, A. Krogh and R.G. Palmer; *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood, CA, 1991.

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

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

[10] D.B. Schwartz, V.K. Samalam, S.A. Solla and J.S. Denker: “Exhaustive Learning”. *Neural Computation 2*, pp. 374–385, 1990.

[11] S. A. Solla; “Supervised Learning: A Theoretical Framework”. In: M. Casdagli, S. Eubank (eds.), *Nonlinear Modeling and Forecasting, SFI Studies in the Science of Complexity, Proc. Vol. XII*, Addison-Wesley, Redwood, CA, 1992.

APPENDIX: Some proofs and theorems

This section starts with a theorem concerning the transfer function f that gives more insight in the values for which saddle points can be found. As a result this theorem gives as possible levels for the error corresponding to the saddle points the levels 0.32, 0.407392 and 0.403321. Furthermore some theorems are proved on the behaviour of a function in the neighbourhood of a point where the gradient is equal to zero. These results are derived by considering the Taylor series expansion of the function in the neighbourhood of such a point. Only results that are needed for the proofs in section 6 are considered.

A result on the transfer function and the error levels of the saddle points

If $a_{00} = a_{01} = a_{10} = a_{11} = 1$ then R_{00}, R_{01}, R_{10} and R_{11} all have the form $(f(x)–0.1)f'(x)$ with x depending on the weights. So we defined $g(x) = (f(x)–0.1)f'(x)$. Carefully considering the cases where $\nabla E = 0$ makes clear that in all these cases equation (6.8) results in:

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

with a, b and c functions of the weights. So we investigated this equation a bit deeper and 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 2: 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 are 0.32, 0.407392 and 0.403321, respectively.

Proof of theorem A.1:
The values $a = b = c = 0$ certainly result in a solution. Consider the graph of $g(x)$, given in figure 8. From the definition of $g(x)$ and equation (3.3) 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 $f(x) = 0$, and one maximum and one minimum and that $\lim_{x \to \pm\infty} g(x) = 0$.

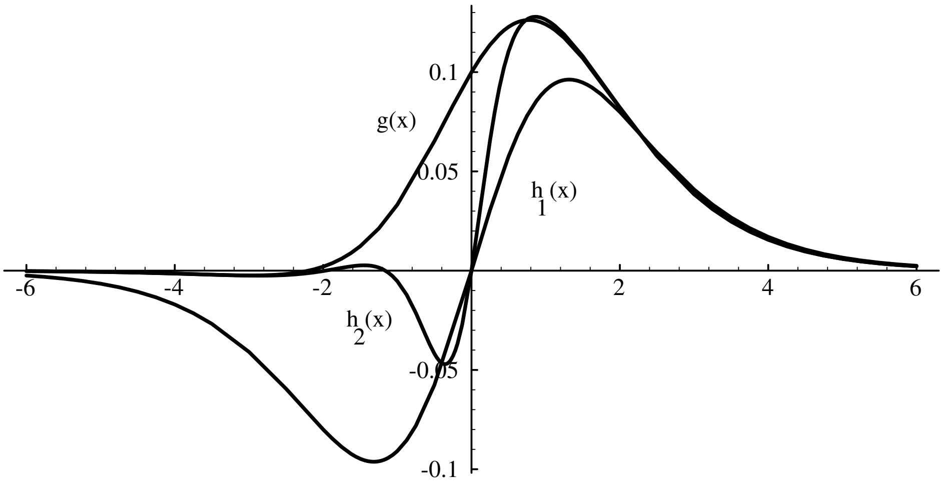

Figure 8. The functions $g(x) = (f(x)–0.1)f'(x)$, $h_1(x) = g(x) - g(-x)$ and $h_2(x) = g(x) - g(-3x)$.

For each value of $g(a)$ at most two different points P and Q exist such that $g(P) = g(Q) = g(a)$ (and thus a, $-a-b$, $-a-c$, $a+b+c$ all have to be equal either to P or to Q). Let us investigate all possibilities except for those that do not result in new possible solutions. So only the possible solutions with $a = P$ are considered, and possible solutions which are obtained by interchanging the values of b and c in an already considered solution are not studied separately. 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 obliged 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 8) 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 8) is equal to zero if and only if x is equal to one of the values in the set $\{0,P_1,P_2\} = \{0, -1.16139, -1.96745\}$. Essentially only the region $f^{-1}(0.1) < x < -\frac{1}{3} f^{-1}(0.1)$ has to be investigated.
So we have the following possibilities:

- $a = P, -a-b = P, -a-c = P, a+b+c = P \Leftrightarrow 2P = -2P \Leftrightarrow P = 0 \Leftrightarrow a = b = c = 0$.
- $a = P, -a-b = P, -a-c = P, a+b+c = Q \Leftrightarrow Q = -3P \Leftrightarrow \exists x \mid g(x) = g(-3x)$. This again leads to the solution with $P = Q = 0$ and to the solutions in the second row of the table. Thus two extra solutions are obtained.
- $a = P, -a-b = P, -a-c = Q, a+b+c = P \Leftrightarrow Q = -3P$. This leads to the solutions in the third row of table 2, on page 22. From symmetry in b and c also the solutions in the fourth row of the table are found. Thus four extra solutions are obtained.
- $a = P, -a-b = P, -a-c = Q, a+b+c = Q \Leftrightarrow P = -Q \Leftrightarrow \exists x \mid g(x) = g(-x)$. This results in the known solution with $P = Q = 0$.

- *a = P, –a–b = Q, –a–c = Q, a+b+c = P* ⇔ *P = –Q*. No new solution is obtained.
- *a = P, –a–b = Q, –a–c = Q, a+b+c = Q* ⇔ *P = –3Q*. This results in the solutions in the fourth row of the table. So the last two solutions of the nine solutions represented in the table are found.

Conclusion: *The nine solutions represented in table 2 are the only solutions of g(a) = = g(–a–b) = g(–a–c) = g(a+b+c).* ❑

Some theorems which prove that certain points are saddle points

Now some theorems follow deciding from certain higher order partial derivatives not being zero that some points are saddle points. The theorems give those results that were needed in section 6.

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

Proof The behaviour of a function in the direct neighbourhood of a certain point is determined by the first (partial) derivative(s) that is (are) unequal to zero. So if for a function $q(a,b)$ of two variables $\nabla q = 0$ holds in a certain point, then the first approximation of this function is given by the second and third order terms of the Taylor series expansion:

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

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

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

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 attain values with opposite sign for $x < 0$ and $x > 0$. ❑

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

Proof We will consider the behaviour of the Taylor series expansion as function of x, with $\Delta a = \alpha x$, $\Delta b = \beta x$ and $\Delta c = \gamma x^3$. This results in:

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

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 thus nonzero values of α, β and γ can be found such that the coefficient of x^5 is unequal to zero and thus q can attain both higher and lower values for small values of x and thus the point considered is a saddle point. ❑

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 attains a saddlepoint 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:

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

If the term with x^5 is unequal to zero, theorem A.2 can be applied. The term with x^6 being not identically zero is solved by application of theorem A.3. 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. ❑