# A Comment on a Paper of Blum: Blum's "local minima" are saddle points

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

## Abstract

*Blum [1] states that the error surface of a one layer network with two hidden units for the XOR function where the weights are restricted to be symmetrical (so 5 independent weights remain) has a manifold of local minima with error unequal to zero. Blum's proof, however, is not correct. In this comment we will show that all points of the given manifold are saddle points and not local minima.*

## Introduction

The error of a neural network can be expressed as the sum of the errors of the patterns to be represented. Given an input, the output of the network is determined by its weights. Thus for a fixed training set, the error is a function of the weights. This function is called the *error surface*. The goal of training the neural network is finding a weight combination with minimal error. The existence of local minima, i.e. minima with a value above the value of the absolute minimum, on error surfaces of neural networks for certain problems is essential for the probability to learn solutions or to be trapped in a local minimum.

We recently investigated the error surfaces for the simplest networks that have the ability to learn the XOR function [4, 5, 6]. Our results are that the error surface of the simplest network, see figure 1, has no local minima.The error surface of the network with two hidden units, see figure 2, only has local minima for values of some of the weights $w_{ij}$, from the inputs to the hidden units, equal to plus or minus infinity. So the latter network has no local minima when all weights are finite.Investigations of the literature lead us via the work of Gori and Tesi [3] to a paper of Blum [1] that gives local minima for one of the XOR networks. Blum [1] states that the error surface of a one layer network with two hidden units for the XOR function where the weights are restricted to be symmetrical (so 5 independent weights remain) has a manifold of local minima with error unequal to zero.

Checking Blum's proof, we found that it could not be correct. Moreover, also Blum's result that a certain manifold of points is a local minimum, does not hold: we were able to proof that all points on the given manifold are saddle points. In [2]

Frasconi et al. generalize, as one of their examples of local minima, Blum's result to the network with 9 independent weights and they give two figures showing that these points are local minima. One of their conclusions is that the point with all weights equal to zero is a local minimum for this network and the XOR function, while it is a saddle point for a linearly separable problem. In contrast, from our results [6], it follows that also the XOR function results in a saddle point.

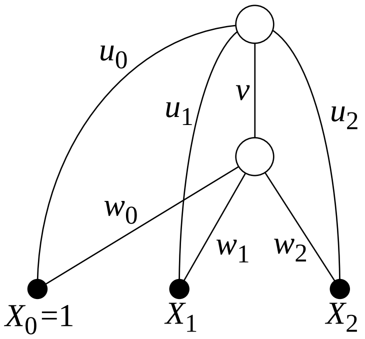

Figure 1. The simplest XOR network

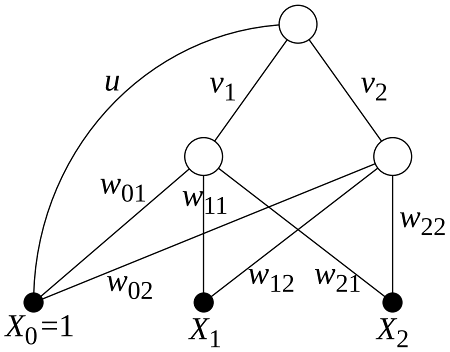

Figure 2. The XOR network with 2 hidden units

## What is wrong with Blum's proof?

Blum's proof is based on two incorrect assumptions. Firstly, Blum assumes that the outputs $z_i$ for different patterns, which are equal in the stationary points, are also equal in the neighbourhood of these stationary points. Secondly, Blum assumes that the gradient of the error depends linearly on the distance to a stationary point. This second assumption is true only for those points, lying in a direction where the second order derivative is unequal to zero.

## The proof that all points are saddle points

We will use the same notation as Blum, who considers the following network:

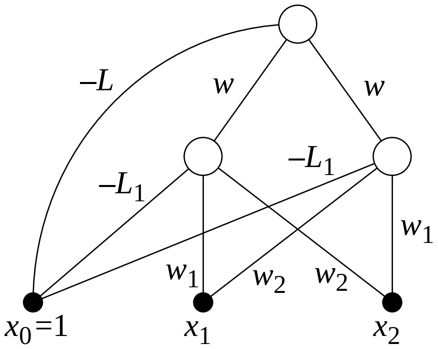

Figure 2. The network and its weights

and the standard sigmoid function: $\sigma(x) = \frac{1}{1 + e^{-x}}$.

So the 5 weights $w$, $w_1$, $w_2$, $L$ and $L_1$ determine the output of the network, as function of the inputs $x_1$ and $x_2$, as:

$$z_{x_1, x_2} = \sigma(-L + w\sigma(-L_1 + w_1 x_1 + w_2 x_2) + w\sigma(-L_1 + w_2 x_1 + w_1 x_2))$$

The XOR problem resulting in local minima after Blum is specified in table 1, where $\sigma(w\text{-}L) > t_1 > 0$. Thus a special case is that $w-L = 0$ and $t_2 = 1-t_1$, e.g. $t_1 = 0.1$ and $t_2 = 0.9$. Since $z_{0,1} = z_{1,0}$ because of the symmetry of the weights, the mean square error corresponding to these values is:

$$E = \frac{1}{2}(z_{0,0} - t_1)^2 + (z_{0,1} - t_2)^2 + \frac{1}{2}(z_{1,1} - t_1)^2$$

**Table 1: Patterns for the XOR problem**

| Pattern | $x_1$ | $x_2$ | desired output |
|---|---|---|---|
| $\xi_1$ | 0 | 0 | $t_1$ |
| $\xi_2$ | 0 | 1 | $t_2 = 2\sigma(w-L)-t_1$ |
| $\xi_3$ | 1 | 0 | $t_3 = t_2$ |
| $\xi_4$ | 1 | 1 | $t_4 = t_1$ |

With these desired values for the patterns of the XOR problem, Blum finds stationary points with $L_1 = w_1 = w_2 = 0$ and remarks about these stationary points: "*They correspond to w and L satisfying the equation* $\sigma(w-L) = (t_1+t_2)/2$. *Hence, they lie on a line w = L + const. in the* $(w,L)$ *plane. Actually, these points are local minima of E, the value of E being* $(t_1-t_2)^2/2$." Here we will show that all stationary points with $L_1 = w_1 = w_2 = 0$ laying on a line $w = L$ + constant in the $(w,L)$ plane, such that $\sigma(w-L) \neq t_1$ are saddle points, so they are *not* local minima. The proof is splitted into two parts: first we will show that all stationary points with $w = 0$ are saddle points, next, that the stationary points with $w \neq 0$ on the given line are saddle points. (Remark that $\sigma(w-L) = t_1$ for stationary points with $L_1 = w_1 = w_2 = 0$ implies that $t_1 = t_2$ and $E = 0$. These points are absolute minima.)

## Points with *w* = 0

The proof that these points are saddle points, is similar to the proof for the simplest XOR network, where points with the weight from the hidden unit to the output unit equal to zero, are saddle points [5]. It is based on the observation that all partial derivatives of the error with respect to the weights $w_1$ and $w_2$ have at least one factor $w$ and so:

$$\left.\frac{\partial^{i+j} E}{\partial w_1^i \partial w_2^j}\right|_{w=0} = 0 \text{ if } i+j > 0$$

Let us introduce the following abbreviations:

$$\begin{aligned} A_{00} &= -L + 2w\sigma(-L_1) \\ A_{01} &= -L + w(\sigma(-L_1 + w_2) + \sigma(-L_1 + w_1)) \\ A_{11} &= -L + 2w\sigma(-L_1 + w_1 + w_2) \end{aligned}$$

thus $z_{0,0} = \sigma(A_{00})$, $z_{0,1} = \sigma(A_{01})$ and $z_{1,1} = \sigma(A_{11})$. Then we find:

$$\begin{aligned} \frac{\partial E}{\partial w_1} &= 2(z_{0,1} - t_2)\sigma'(A_{01})w\sigma'(-L_1 + w_1) + \\ &\quad 2(z_{1,1} - t_1)\sigma'(A_{11})w\sigma'(-L_1 + w_1 + w_2) \end{aligned}$$

$$\begin{aligned} \frac{\partial^2 E}{\partial w_1 \partial w} &= 2\{\sigma'(A_{01})\}^2 w\sigma'(-L_1 + w_1)(\sigma(-L_1 + w_2) + \sigma(-L_1 + w_1)) + \\ &\quad 2(z_{0,1} - t_2)\sigma''(A_{01})w\sigma'(-L_1 + w_1)(\sigma(-L_1 + w_2) + \sigma(-L_1 + w_1)) + \\ &\quad 2(z_{0,1} - t_2)\sigma'(A_{01})\sigma'(-L_1 + w_1) + \\ &\quad 4\{\sigma'(A_{11})\}^2 w\sigma'(-L_1 + w_1 + w_2)\sigma(-L_1 + w_1 + w_2) + \\ &\quad 4(z_{1,1} - t_1)\sigma''(A_{11})w\sigma'(-L_1 + w_1 + w_2)\sigma(-L_1 + w_1 + w_2) + \\ &\quad 2(z_{1,1} - t_1)\sigma'(A_{11})\sigma'(-L_1 + w_1 + w_2) \\ &= 2(z_{0,1} - t_2)\sigma'(A_{01})\sigma'(-L_1 + w_1) + \\ &\quad 2(z_{1,1} - t_1)\sigma'(A_{11})\sigma'(-L_1 + w_1 + w_2) + O(w) \end{aligned}$$

Thus in stationary points with $L_1 = w_1 = w_2 = 0$ we obtain

$$\left.\frac{\partial^2 E}{\partial w_1 \partial w}\right|_{L_1 = w_1 = w_2 = w = 0} = 2\sigma'(w - L)\sigma(0)(2\sigma(w - L) - t_1 - t_2)$$

which is equal to zero in the considered stationary points, since $\sigma(w–L) = (t_1+t_2)/2$. Further differentiation leads to:

$$\frac{\partial^3 E}{\partial w_1 \partial w_2 \partial w} = 2(z_{1,1} - t_1)\sigma'(A_{11})\sigma''(-L_1 + w_1 + w_2) + O(w)$$

Since $\sigma''(0) = 0$, also

$$\left.\frac{\partial^3 E}{\partial w_1 \partial w_2 \partial w}\right|_{L_1 = w_1 = w_2 = w = 0} = 0$$

but

$$\frac{\partial^4 E}{\partial w_1^2 \partial w_2 \partial w} = 2(z_{1,1} - t_1)\sigma'(A_{11})\sigma'''(-L_1 + w_1 + w_2) + O(w)$$

and thus

$$\left.\frac{\partial^4 E}{\partial w_1^2 \partial w_2 \partial w}\right|_{L_1 = w_1 = w_2 = w = 0} = 2(\sigma(w - L) - t_1)\sigma'(w - L)\sigma'''(0)$$

is unequal to zero if $t_1 \neq \sigma(w\text{–}L)$. Using theorem A.4 from [5] proofs that a stationary point with $\partial^{i+j}E/\partial w_1{}^i \partial w_2{}^j = 0$ for $0 < i+j < 8$ and $\partial^4 E/\partial w_1{}^2 \partial w_2 \partial w \neq 0$ is a saddle point. This saddle point with $t_1 = 0.1$ and $t_2 = 0.9$ is visualized in figure 4.

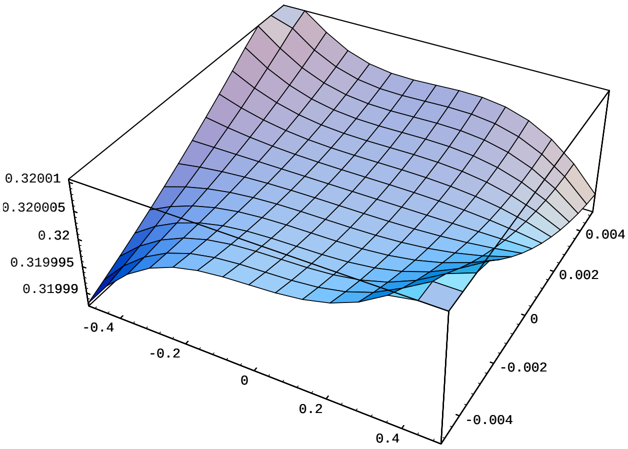

Figure 3. The error surface around the point with $L = w = L_1 = w_1 = w_2 = 0$, $t_1 = 0.1$ and $t_2 = 0.9$. The values of weights are varied such that $\Delta w_1 = \Delta w_2$ varies from -0.5 to 0.5, while $\Delta w$ varies from -0.005 to 0.005.

## Points with $w \neq 0$

In order to prove that these stationary points are saddle points we use similar techniques as in [6], where the more general case with all 9 weights independent is treated.

First we calculate the second order part of the Taylor series expansion around these points, resulting in:

$$\begin{aligned}\Delta E \approx\ & (\sigma'(w-L))^2\left[\Delta L - \Delta w + \frac{1}{2}w\Delta L_1\right]^2 + \\ & (\sigma'(w-L))^2\left[\Delta L - \Delta w + \frac{1}{2}w\Delta L_1 - \frac{1}{4}w\Delta w_1 - \frac{1}{4}w\Delta w_2\right]^2 + \\ & (\sigma'(w-L))^2\left[\Delta L - \Delta w + \frac{1}{2}w\Delta L_1 - \frac{1}{2}w\Delta w_1 - \frac{1}{2}w\Delta w_2\right]^2 + \\ & \frac{1}{2}\sigma''(w-L)\,w^2(\sigma(w-L) - t_1)\,[\Delta w_1 + \Delta w_2]^2\end{aligned}$$

The second order part of the Taylor series expansion is zero if we are considering directions such that $2\Delta L\text{–}2\Delta w + w\Delta L_1 = 0$ and $\Delta w_1 + \Delta w_2 = 0$. So we will investigate points in the neighbourhood of the stationary points with $\Delta w = x$, $\Delta L_1 = y$, $\Delta w_1 = z$, $\Delta L = x - \frac{1}{2}wy$ and $\Delta w_2 = \text{–}z$. The variation in the $x$-direction results in other points on the line $L\text{–}w$ = constant. So we will investigate the results of variations of $y$ and $z$. We obtain the error $E$ as function of $y$ and $z$:

$$E = \frac{1}{2}(\sigma(A_{00}) - t_1)^2 + (\sigma(A_{01}) - t_2)^2 + \frac{1}{2}(\sigma(A_{11}) - t_1)^2$$

with

$$A_{00} = -L + \frac{1}{2}wy + 2w\sigma(-y)$$

$$A_{01} = -L + \frac{1}{2}wy + w(\sigma(-y-z) + \sigma(-y+z))$$

$$A_{11} = -L + \frac{1}{2}wy + 2w\sigma(-y)$$

Considering the partial derivatives of E with respect to *y* and *z*, it is clear that all first and second order partial derivatives are equal to zero if *y* = *z* = 0. Calculation of third order partial derivatives yields:

$$\left.\frac{\partial^3 E}{\partial y \partial z^2}\right|_{y = z = 0} = -4(\sigma(-L+w) - t_2)\,\sigma'(-L+w)\,w\sigma'''(0)$$

which is unequal to zero, and thus also these stationary points are saddle points. One of these saddle points is visualized in figure 5.

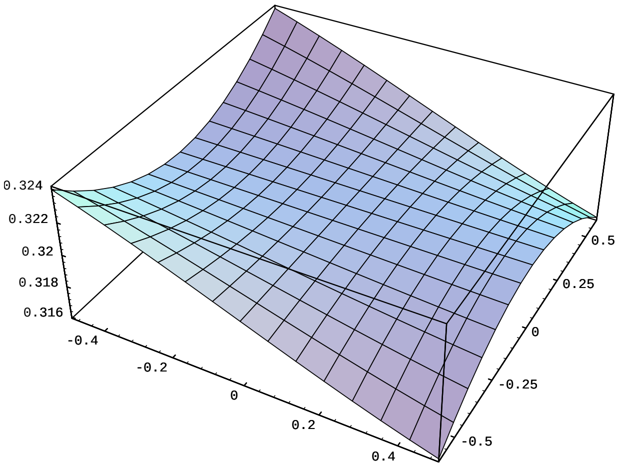

Figure 4. The error surface around the point with $L = w = 1$, $L_1 = w_1 = w_2 = 0$, $t_1 = 0.1$ and $t_2 = 0.9$. The values of weights are varied such that $\Delta L_1 = -2\Delta L$ varies from –0.5 to 0.5, while $\Delta w_1 = -\Delta w_2$ varies from -0.6 to 0.6.

## Conclusion

All stationary points of the network of figure 3 for the XOR function as specified in table 1 with $L_1 = w_1 = w_2 = 0$, satisfying the equation $\sigma(w-L) = (t_1 + t_2)/2$, $0 < t_1, t_2 < 1$, $t_1 \neq t_2$, are saddle points.

## References

[1] E.K. Blum; “Approximation of Boolean Functions by Sigmoidal Networks: Part I: XOR and other Two-variable functions”. In: *Neural Computation 1*, 532-540, 1989.

[2] P. Frasconi, M. Gori and A. Tesi; “Success and Failures of Backpropagation: a Theoretical Investigation”. In: O. Omidvar (ed.); *Progress in Neural Networks (5)*, Ablex Publishing, 1993.

[3] M. Gori and A. Tesi; “On the Problem of Local Minima in Backpropagation”. In: *IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 14, No. 1*, 76-86, 1992.

[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, 1994.

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