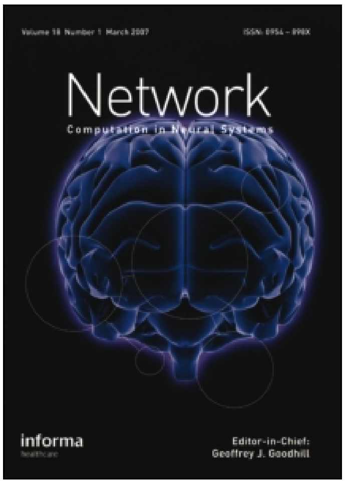

Network: Computation in Neural Systems
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t713663148

Complete solution of the local minima in the XOR problem

P. J. G. Lisboa [a]; S. J. Perantonis [b]
[a] Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool, UK [b] Department of Applied Mathematics and Theoretical Physics and Department of Computer Science, University of Liverpool, UK

Online publication date: 01 January 1999

To cite this Article Lisboa, P. J. G. and Perantonis, S. J.(1999) 'Complete solution of the local minima in the XOR problem', Network: Computation in Neural Systems, 2: 1, 119 — 124

To link to this Article: DOI: 10.1088/0954-898X/2/1/007

URL: http://dx.doi.org/10.1088/0954-898X/2/1/007

Network 2 (1991) 119–124. Printed in the UK

Complete solution of the local minima in the XOR problem

P J G Lisboa[†] and S J Perantonis[‡]

[†]Department of Electrical Engineering and Electronics, University of Liverpool, Liverpool L69 3BX, UK

[‡]Department of Applied Mathematics and Theoretical Physics and Department of Computer Science, University of Liverpool L69 3BX, UK

Received 13 July 1990

Abstract. A complete solution of the excitation values which may occur at the local minima of the XOR problem is obtained analytically for two-layered networks in the two most commonly quoted configurations, using the gradient backpropagation algorithm. The role of direct connections which bypass the two-layered system is discussed in connection to the XOR problem and other related training tasks.

1. Introduction

Although the XOR problem is widely quoted as a test for a variety of neural networks (Rumelhart *et al* 1986a), and its historical pedigree for illustrating difficulties in the learning abilities of neural networks is well established (Minsky and Papert 1969), a complete solution of the local minima for even the most common multilayered topologies trained by gradient backpropagation (GBP) (Rumelhart *et al* 1986a, b) appears not to have been published.

In this paper we show that this problem can be solved analytically. The solution is interesting, because it illustrates that local minima of the cost function for some training tasks in multilayered networks can be revealed by analysis. It is also instructive, not least because it sheds light on the effect of different topologies and particularly of the use of direct connections which bypass the multilayered perceptron. Taking into account the general features of our analytical solution we investigate ways of initializing the weights of the networks of minimal architecture in order to train them to solve the XOR problem more effectively. We also comment on the usefulness of the direct connections for accomplishing other training tasks, notably higher-order parity problems as well as autoassociation problems.

2. Derivation

In order to facilitate the analytical treatment of networks trained by GBP, the following variant of the usual quadratic cost function will be used (Wallace 1987):

$$L = -\sum_{\alpha}\sum_{k} \ln\left[(O_k^{\alpha})^{t_k^{\alpha}}(1 - O_k^{\alpha})^{1 - t_k^{\alpha}}\right]. \tag{1}$$

0954-898X/91/010119+06$03.50 © 1991 IOP Publishing Ltd

The indices α and k label the nominated patterns and output nodes, respectively. O_k^α are the excitations of the top layer of nodes and t_k^α are their corresponding target values.

The cost function in equation (1) has the advantage over the more frequently quoted

$$E = \frac{1}{2}\sum_{k\alpha}(O_k^\alpha - t_k^\alpha)^2 \tag{2}$$

that it exhibits no stationary points at values of the network outputs equal to zero or one, where weights to the top layer of nodes assume infinite values. It also has the effect of increasing the size of the error which is propagated back through the network during training. Apart from these points, the rest of the analysis in this section does not depend on which of the two cost functions is used.

Using the cost function L and keeping to networks with a single output node and an arbitrary number N of hidden nodes arranged in one layer, we can study the effect that changes in the network topology have on the occurrences of local minima of the cost function for the XOR problem. The target outputs of the four input patterns $I^1 = (1, 1)$, $I^2 = (1, 0)$, $I^3 = (0, 1)$ and $I^4 = (0, 0)$ are $t^1 = \delta$, $t^2 = 1 - \delta$, $t^3 = 1 - \delta$ and $t^4 = \delta$ respectively. We shall use a sigmoid response function $f(x) = 1/[1 + \exp(-x)]$ for the formal neurons.

Consider first networks without direct bottom-to-top synaptic connections (the minimal network of this architecture which can be used to solve the XOR problem is shown in figure 1(*a*). With the notation suggested in figure 1(*a*), the stationary point equations are given by:

$$\partial L/\partial\phi = 0 \Rightarrow \sum_\alpha R^\alpha = 0 \tag{3}$$

$$\partial L/\partial v_j = 0 \Rightarrow \sum_\alpha R^\alpha y_j^\alpha = 0 \tag{4}$$

$$\partial L/\partial\theta_j = 0 \Rightarrow v_j \sum_\alpha R^\alpha y_j^\alpha(1 - y_j^\alpha) = 0 \tag{5}$$

$$\partial L/\partial u_{ij} = 0 \Rightarrow v_j \sum_\alpha R^\alpha y_j^\alpha(1 - y_j^\alpha) I_i^\alpha = 0 \tag{6}$$

where $R^\alpha = t^\alpha - O^\alpha$ and y_j^α are the intermediate layer outputs. Equations (4)–(6), for each value of $j (j = 1, 2, \ldots, N)$, form a set of simultaneous equations for the R^α. These equations have non-zero solutions, which represent stationary points, only when their determinants are equal to zero:

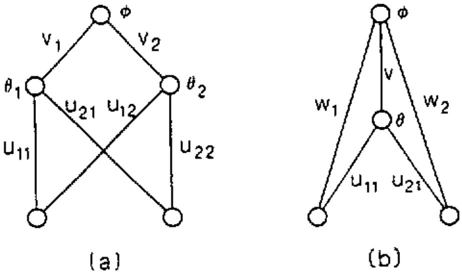

Figure 1. Minimal-size feedforward networks used to solve the XOR problem. (*a*) Minimal-size network with only consecutive layer interactions. (*b*) Minimal-size network with direct bottom-to-top connections.

$$\Delta_j = (v_j)^3 \prod_{\alpha=1}^{4} y_j^\alpha(1 - y_j^\alpha)\left(\frac{1}{1 - y_j^1} + \frac{1}{1 - y_j^4} - \frac{1}{1 - y_j^3} - \frac{1}{1 - y_j^2}\right). \tag{7}$$

For each hidden node $j = 1, 2, \ldots, N$ let

$$\eta_{0j} = \exp(\theta_j) \tag{8}$$

$$\eta_{1j} = \exp(u_{1j}) \tag{9}$$

$$\eta_{2j} = \exp(u_{2j}). \tag{10}$$

Equations (7) then become

$$(v_j)^3 \prod_{\alpha=1}^{4} y_j^\alpha(1 - y_j^\alpha)\eta_{0j}(1 - \eta_{1j})(1 - \eta_{2j}) = 0 \qquad j = 1, 2, \ldots, N. \tag{11}$$

From these equations we may readily read off values of the weights and thresholds for which undesirable local minima of L may occur. These are:

(i) $v_j = 0$

or (ii) $u_{ij} = 0$ corresponding to $\eta_{ij} = 1$, $i = 1$ or 2

or (iii) $u_{ij} = \pm\infty$ or $\theta_j = \pm\infty$ corresponding to $y_j^\alpha = 0$ or 1.

We next combine these values of the weights for each value of j and substitute them back into equations (3)–(6). Considering the cases where one, two, three or all four outputs differ from their targets separately, we can identify the output values for which stationary points can occur.

For example, suppose that only two outputs, say O^1 and O^2, differ from their target values, so that $R^1 \neq 0$, $R^2 \neq 0$ and $R^3 = R^4 = 0$. From equations (5) and (6) we then obtain for a certain j either that $v_j = 0$ or that y_j^1 and y_j^2 are equal to 0 or 1. In the second case, however, y_j^1 and y_j^2 must both be equal to 0 or both equal to 1, otherwise equation (4) gives $R^1 = 0$ or $R^2 = 0$, contrary to our assumption that only two network outputs differ from their targets. Hence for all values of j we have $v_j = 0$ or $y_j^1 = y_j^2$. Taking into account this result for all values of j we conclude that $O^1 = O^2$. Since $R^1 + R^2 = 0$ (by equation (3)), it follows that $O^1 = O^2 = \frac{1}{2}$. The other cases are treated similarly and the output values for which stationary points can occur are the following:

(a) $O^1 = O^2 = O^3 = O^4 = \frac{1}{2}$;

(b) $O^1 = O^2 = \frac{1}{2}, O^4 = \delta, O^3 = 1 - \delta$ and similar solutions with $O^1 \leftrightarrow O^4$ and $O^2 \leftrightarrow O^3$;

(c) $O^1 = O^2 = O^3 = (2 - \delta)/3, O^4 = \delta$ and the corresponding solution with O^1 and O^4 interchanged;

(d) $O^1 = O^3 = O^4 = (1 + \delta)/3, O^2 = 1 - \delta$ and the corresponding solution with O^2 and O^3 interchanged.

The analysis leading to the identification of stationary points of types (a)–(d) is independent of the number N of hidden nodes in the intermediate layer. This is basically a consequence of the fact that equations (11) have a similar structure for all values of N. Consequently, the stationary points of the XOR problem for all networks of a 2–N–1 architecture with N arbitrary can be classified into the categories (a)–(d). What changes as N varies is the relative incidence of the stationary points in the weight space.

Stationary points of types (b), (c) and (d) occur only if at least one of the weights takes on a value equal to $\pm\infty$ (case (iii) above), whereas a stationary point of type (a) can involve finite values of the weights (cases (i) and (ii) above).

Equations (7) are only necessary and not sufficient conditions for local minima to exist. They serve to characterize the stationary points of the network in general.

Numerical simulations presented in the next section show that case (a) occurs frequently as a saddle point. However, examination of the second derivatives of L with respect to the weights shows that for certain values of the weights all types (a)–(d) of stationary points can be true local minima. Our numerical simulations have verified their stability. A sample of characteristic weights for points in the vicinity of local minima, the stability of which was also verified numerically, is shown in table 1. Some of these points were reached through GBP starting from random weights, while others were constructed by solving equations (a)–(d) and taking into account the second derivatives of L.

We next consider adding direct bottom-to-top connections to the three-layer network of a 2–N–1 architecture with N arbitrary (the minimal configuration of this architecture used to solve the XOR problem is shown in figure 1(*b*). The effect of adding direct connections is to impose additional constraints, namely

$$\sum_{\alpha} R^{\alpha} I_i^{\alpha} = 0 \tag{13}$$

with the coefficients of the R^{α} independent of the intermediate node outputs. From equations (3) and (13) we obtain

$$R^1 = -R^2 = -R^3 = R^4. \tag{14}$$

Substituting this result into the equations for the stationary points we conclude that there remain only local minima of type (a), where all outputs are equal. Again, this result holds true irrespectively of the number of intermediate nodes for networks with direct connections.

3. Training efficiency and higher-order parity problems

In the light of the results of the previous section, it is interesting to study the frequency of occurrence of the stationary points when the minimal networks of figures 1(*a*) (network A) and 1(*b*) (network B) are trained through GBP, whereby the weights and biases, here generically called Q, are updated according to the relation

$$\Delta Q(t+1) = -\eta \frac{\partial L}{\partial Q} + e\Delta Q(t). \tag{15}$$

Table 1. A sample of weights at typical local minima of the cost function for the XOR problem. The value $\delta = 0.1$ has been used.

O^1	0.5	0.1	0.633	0.366	0.5
O^2	0.9	0.9	0.633	0.366	0.5
O^3	0.5	0.5	0.633	0.9	0.5
O^4	0.1	0.5	0.1	0.366	0.5
u_{11}	−5.520 58	−11.521 44	−13.708 96	−10.424 64	0.0
u_{12}	−4.508 67	−11.899 91	6.014 91	−12.714 14	−0.572 21
u_{21}	−13.690 16	−1.105 68	−13.708 96	8.557 86	0.483 49
u_{22}	12.274 68	4.010 44	6.014 91	11.477 85	0.0
v_1	−2.783 35	4.592 62	−3.421 22	0.663 31	0.0
v_2	−5.056 70	−6.141 46	0.437 58	4.237 84	0.0
θ_1	1.419 13	12.598 65	1.394 27	= 10.280 05	1.509 31
θ_2	4.735 79	11.707 33	5.217 02	−10.972 65	−0.896 11
ϕ	5.056 70	1.548 84	0.108 97	−0.546 56	0.0

With network A, a stationary point of type (a) can occur if at least two of the weights are equal to zero. Thus the algorithm is often trapped in the vicinity of a saddle point of type (a) for a number of iterations if the initial values of the weights are small. On the other hand, if the algorithm is initialized using relatively large weights, the danger increases that the network be trapped in the vicinity of a stationary point of type (b), (c) or (d), for which at least one of the bottom-to-hidden layer weights assume infinite values. In table 2 we show the results for the relative occurrence of the various types of stationary points for different ranges of random initializing weights and biases. Although results are shown for the values $\eta = 0.25$ and $e = 0.5$, the frequency of minima of types (b), (c) and (d) is largely independent of these values. Evidently, there is an intermediate region (random weights between $-a$ and a with $a = 0.3$–0.5) for which learning is fast and the probability of occurrence of local minima is small (3–4%).

For network B, no stationary points appear at infinite values of the weights. Relatively large initializing weights can now be used to avoid a saddle point or minimum of type (a) thus speeding up training without fear of encountering minima at infinite values of the weights, as illustrated in table 2.

Although it is not possible to classify the stationary points of the cost function for higher-order parity problems in a closed form, numerical results show that our conclusions about the treatment of the XOR problem are useful for training multilayered networks to solve some higher-order parity problems as well. In particular, an investigation of the 4-parity problem using a 4–4–1 architecture without direct connections, showed that all local minima reached had a number of outputs equal to the targets and all wrong outputs equal to each other, a situation very similar to the solution for the local minima of the XOR problem. Many of these minima involve infinite bottom-to-middle layer weights and most of them (although not all) can be eliminated by using bottom-to-top synaptic connections. For both architectures (with or without direct bottom-to-top connections) training with small initial random weights (maximum magnitude $\simeq 0.5$) encounters a stationary point where all outputs are equal to 0.5 and the task cannot be solved in almost all trials. With relatively large weights (maximum magnitude 1.5) the system without direct connections has a relatively low success rate

Table 2. Relative occurrence of stationary points of the cost function for the XOR problem for the networks of figures 1(*a*) and 1(*b*) in 100 training sessions using GBP. The values $\eta = 0.25$ and $e = 0.5$ have been used for the gain and acceleration respectively.

	Network A						Network B	
Range of weights	1000 its			10000 its			1000 its	10000 its
	(a)	(b)	(c), (d)	(a)	(b)	(c), (d)	(a)	(a)
[−0.2, 0.2]	95	0	0	8	1	0	89	14
[−0.3, 0.3]	56	0	0	3	1	0	69	5
[−0.4, 0.4]	32	1	0	1	2	0	45	3
[−0.5, 0.5]	16	3	0	0	3	0	36	3
[−0.6, 0.6]	12	6	0	0	6	0	31	1
[−0.8, 0.8]	5	10	0	0	10	0	16	1
[−1, 1]	2	12	1	0	12	1	6	0
[−2, 2]	0	21	4	0	20	1	0	0
[−3, 3]	0	32	8	0	29	4	0	0

(32%) after 30000 iterations of the GBP algorithm, because it often encounters minima where at least one of the bottom-to-middle layer weights is infinite. By contrast, a success rate of 97% is reached when direct connections are included with a mean-square error of less than 10^{-2} in less than 2000 iterations for a 4–3–1 configuration. In the 6-parity problem, the structure of the local minima appears to be more complex. Nevertheless, using a network with direct connections and large initial weights still facilitates learning and a success rate of 88% in less than 30000 iterations can be reached for a 6–5–1 configuration with initial weights of maximum magnitude equal to 3.

Adding direct connections probably does not help solve higher-order parity problems, where the total number of weights is significantly larger than the number of direct weights (we found the 8-parity problem extremely difficult to solve for the minimal networks either with or without direct connections). It is nevertheless clear that the use of direct connections can facilitate learning in a number of tasks. Apart from the XOR and parity problems discussed here, there is evidence that adding a set of direct weights to the multilayered perceptron can increase substantially its rate of convergence in autoassociation problems, without significant detriment to its performance (Lisboa and Perantonis 1990). In these tasks, the use of the direct connections helps the system learn without forfeiting the benefits arising from its multilayered architecture.

3. Conclusion

The possible excitation values at local minima of the XOR problem have been identified and the associated weights and bias terms characterized.

The use of external synaptic connections which bypass the ordinary multilayered perceptron has the consequence of reducing the number of local minima available to the network, thus facilitating learning. This is not just due to the additional number of equations that must be satisfied at the extrema, but also to the nature of the new equations. Therefore the effect of this network configuration is different from that of simply adding new nodes, say, to the intermediate layer.

Finally, using direct connections has been found helpful in eliminating local minima and speeding up learning for a number of training tasks, including some higher-order parity and autoassociation problems.

References

Lisboa P J G and Perantonis S J 1990 Convergence of recursive associative memories obtained using the multilayered perceptron *J. Phys. A: Math. Gen.* **23** 4039–53

Minsky M and Papert S 1969 *Perceptrons: An Introduction to Computational Geometry* (Cambridge, MA: MIT Press)

Rumelhart D E, Hinton G E and Williams R J 1986a Learning internal representations by error propagation Parallel Distributed Processing: Explorations in the Microstructure of Cognition ed D E Rumelhart and J L McLelland (Cambridge, MA: MIT Press) vol 1, pp 318–62

—— 1986b Learning representations by backpropagating errors *Nature* **323** 533–6

Wallace D J 1987 Neural network models: a physicist's primer *Computational Physics* (*SUSSP 32*) ed R D Kenway and G S Pawley (Edinburgh: Scottish Universities Summer Schools in Physics) pp 168–211