ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 255-260

Training a Sigmoidal Network is Difficult

Barbara Hammer,

University of Osnabrück, Dept. of Mathematics/Comp. Science,
Albrechtstraße 28, 49069 Osnabrück, Germany

Abstract. In this paper we show that the loading problem for a 3-node architecture with sigmoidal activation is NP-hard if the input dimension varies, if the classification is performed with a certain accuracy, and if the output weights are restricted.

1. Introduction

Feedforward networks are a common tool in machine learning. Only some representative data is needed to train a network automatically such that it reproduces a complex input-output association. There exist theoretical guarantees for the generalization capability [7]. But in praxis the training algorithms are sometimes very slow, especially for large input dimensions.

The loading problem is to decide if a training set can be stored by a fixed architecture correctly. Blum and Rivest have shown the NP-completeness for a network with 3 computation units, varying input dimension, and perceptron activation [1]. Dasgupta et.al. have generalized the result to the semilinear activation [2]. Usually, one deals with the sigmoidals $\mathrm{sgd}(x)$ or $\tanh(x)$. Hoeffgen has proved the NP-completeness for sgd, but binary weights [4]. Šíma has shown the NP-hardness for a sigmoidal architecture with an additional condition which is fulfilled e.g. if the output bias is 0 [5]. This last approach deals with a realistic setting, but can neither be transformed to tanh nor be expanded to a classification with reference $\neq 0$. In [8] Vu has presented a result which focuses on the complexity of finding solutions with minimal squared error.

Here, we will deal with networks as a classification tool. We will show the NP-hardness of the loading problem for the sigmoidal 3-node architecture with growing input dimension if the classification accuracy is at least ϵ and the output weights are bounded by a constant B. This result generalizes to functions which can be approximated by a scaled or shifted sgd.

2. The loading problem

The **3-node architecture** computes for an input dimension n the function $f : \mathbb{R}^n \to \mathbb{R}$, $f(\mathbf{x}) = \alpha\, N_1(\mathbf{x}) + \beta\, N_2(\mathbf{x}) + \gamma$, where α, β, $\gamma \in \mathbb{R}$ and the two hidden nodes N_1 and N_2 compute the functions

$$N_1 : \mathbb{R}^n \to \mathbb{R}, \quad N_1(\mathbf{x}) = \mathrm{sgd}(a_0 + \textstyle\sum_{i=1}^n a_i x_i) \quad \text{and}$$
$$N_2 : \mathbb{R}^n \to \mathbb{R}, \quad N_2(\mathbf{x}) = \mathrm{sgd}(b_0 + \textstyle\sum_{i=1}^n b_i x_i)$$

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 255-260

on the input vector $\mathbf{x} = (x_1, \ldots, x_n)$. Here, $\mathbf{a} = (a_1, \ldots, a_n)$, $\mathbf{b} = (b_1, \ldots, b_n) \in \mathbb{R}^n$, a_0, $b_0 \in \mathbb{R}$, and $\mathrm{sgd}(x) = 1/(1 + \mathrm{e}^{-x})$.

The **loading problem** is the following problem: Consider a pattern set $(\mathbf{x}_1, y_1)$, ..., $(\mathbf{x}_m, y_m) \in \mathbb{R}^n \times \{-1, 1\}$. Is it possible to find weights $\mathbf{a}$, $\mathbf{b}$, a_0, b_0, α, β, and γ such that for the corresponding network $f(\mathbf{x}_i) \geq 0 \Leftrightarrow y_i = 1 \, \forall i$?

The **loading problem with accuracy at least ϵ and weight restriction** B is the loading problem as defined above with the additional restrictions: $|f(\mathbf{x}_i)| > \epsilon$ for any $\mathbf{x}_i$, $|\alpha| < B$, $|\beta| < B$. That is, the classification is performed at least with a fixed distance ϵ from the classification bound 0.

Finally **the loading problem with accuracy ϵ, weight restriction B, and unbounded input dimension** consists of all pattern sets with arbitrary input dimension such that each pattern set can be classified with accuracy ϵ and weight restriction B by a 3-node network with appropriate input dimension.

3. A geometric view

Assume a 3-node architecture classifies a pattern set correctly with accuracy ϵ and weight restriction B. The set of parameters such that the patterns are mapped correctly is an open set in $\mathbb{R}^{5+2n}$; therefore after a slight shift of the parameters if necessary we can assume, that $\gamma \neq 0$, $\alpha + \gamma \neq 0$, $\beta + \gamma \neq 0$, and $\alpha + \beta + \gamma \neq 0$. Further, we can assume that $\mathbf{a}$ and $\mathbf{b}$ are linearly independent and $\alpha \neq 0$, $\beta \neq 0$. We are interested in the boundary that is defined by

$$(*) \qquad \alpha \, \mathrm{sgd}(\mathbf{a}^t\mathbf{x} + a_0) + \beta \, \mathrm{sgd}(\mathbf{b}^t\mathbf{x} + b_0) + \gamma = 0 \, .$$

This is empty or forms an $(n-1)$-dimensional manifold M with the following form: If $\mathbf{x} \in M$, then $\mathbf{x} + \mathbf{v} \in M$ for any $\mathbf{v}$ orthogonal to $\mathbf{a}$ and $\mathbf{b}$. Consequently, M is constant in the directions orthogonal to $\mathbf{a}$ and $\mathbf{b}$; to describe M it is sufficient to describe the curve that is obtained if M is intersected with a plane containing $\mathbf{a}$ and $\mathbf{b}$. After a rotation, translation, and scaling we can assume $\mathbf{a}^t\mathbf{x} + a_0 = x_1$ where x_1 is the first component of $\mathbf{x}$. Then the curve can be parametrized by x_1, a normal vector by $n(x_1) = \alpha \, \mathrm{sgd}'(x_1) \cdot \mathbf{a} + \beta \, \mathrm{sgd}'(\mathbf{b}^t\mathbf{x} + b_0) \cdot \mathbf{b}$ where the term $\mathbf{b}^t\mathbf{x} + b_0$ can be substituted using $(*)$. Define $\tilde{n}(x_1) = n(x_1)/|n(x_1)|$. Considering γ, $\gamma + \alpha$, $\gamma + \beta$, and $\gamma + \alpha + \beta$ several cases result:
Case 1: All values are positive or all values are negative: M is empty.
Case 2: One value is positive, the other three are negative:
Since $\mathrm{sgd}(-x) = 1 - \mathrm{sgd}(x)$ we can assume that $\gamma > 0$, $\alpha < -\gamma$, and $\beta < -\gamma$.

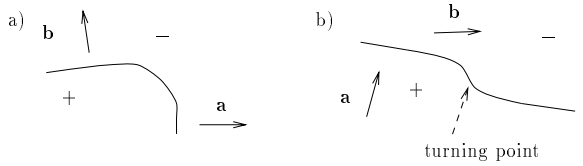

Figure 1: Classifications by the 3-node architecture

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 255-260

Dividing $(*)$ by γ we obtain $\gamma = 1$, $\alpha < -1$, and $\beta < -1$. The curve describing M looks like in Fig.1a, especially, it is convex, as can be seen as follows: For $\mathrm{sgd}(x_1) \approx -1/\alpha$ the normal vector is $\tilde{n}(x_1) \approx -\mathbf{a}/|\mathbf{a}|$. For $\mathrm{sgd}(x_1) \approx 0$ it is $\tilde{n}(x_1) \approx -\mathbf{b}/|\mathbf{b}|$. In general, $\tilde{n}(x_1) = \lambda_1(x_1)\,\mathbf{a} + \lambda_2(x_1)\,\mathbf{b}$ for appropriate functions λ_1 and λ_2. Assume, the curve is not convex. Then, there would exist at least two points at the curve with identical $\tilde{n}$, identical λ_1/λ_2 and, consequently, at least one point x_1 with $(\lambda_1/\lambda_2)'(x_1) = 0$. But one can compute $(\lambda_1/\lambda_2)'(x_1) = C(x_1) \cdot (-\beta - 1 + \mathrm{sgd}(x_1)(2\beta + 2) + \mathrm{sgd}^2(x_1)(2\alpha + \alpha^2 + \alpha\beta))$ with some factor $C(x_1) \neq 0$. If $(\lambda_1/\lambda_2)'(x_1)$ was 0, $\alpha = \beta = -1$ or

$$(**) \qquad \mathrm{sgd}(x_1) = \frac{-\beta - 1}{\alpha(\alpha + \beta + 2)} \pm \sqrt{\frac{(1+\beta)\left((\alpha+1)^2 + \beta(\alpha+1)\right)}{\alpha^2(\alpha+\beta+2)^2}}\,,$$

where the term of the root is negative except for $\alpha = -1$ or $\beta = -1$.
Case 3: Exactly two values are positive:
Arguing as before we can assume $\gamma = 1$, $\alpha < -1$, $\beta > -1$, and $\alpha + \beta < -1$. If $\mathrm{sgd}(x_1) \approx -\gamma/\alpha$ or $\mathrm{sgd}(x_1) \approx (-\gamma - \beta)/\alpha$ it is $\tilde{n}(x_1) \approx -\mathbf{a}/|\mathbf{a}|$. The curve describing M has an S-shaped form (see Fig.1b) because there exists at most one point of the curve where $(\lambda_1/\lambda_2)'(x_1)$ vanishes: This is $\mathrm{sgd}(x_1) = 0.5$ if $\alpha + \beta + 2 = 0$, it is the solution $(**)$ with positive sign if $\alpha + \beta < 0$, and the solution $(**)$ with negative sign if $\alpha + \beta > 0$.
Case 4: Exactly 3 values are positive: This is dual to case 2.

4. Main theorem

Theorem 1 *For fixed $\epsilon \in]0, 0.5[$ and $B \geq 2$ it is NP-hard to solve the loading problem with accuracy ϵ, weight restriction B, and unbounded input dimension for the 3-node sigmoidal architecture.*

Proof: The $(2,3)$**-set splitting problem** (SSP) is the following problem: Given a set $S = \{s_i \mid 1 \leq i \leq n\}$ and a set $C = \{c_j \mid 1 \leq j \leq m\}$ of subsets of S where each c_j contains exactly 3 elements, does two disjoint subsets S_1, $S_2 \subset S$ exist such that $S = S_1 \cup S_2$ and $c_j \not\subset S_1$, $c_j \not\subset S_2$ for $j \in \{1, \ldots, m\}$?

The SSP is NP-complete [3]. It will be reduced to the loading problem in polynomial time showing that the loading problem is NP-hard.
Reduction: For a SSP the following $m + n + 15$ patterns in $\mathbb{R}^{n+5}$ can be loaded exactly if the SSP is solvable:
Positive examples, i.e. the output shall be $> \epsilon$, are

- the points $(0, \ldots, 0, 1, 0 \ldots, 0, 1, 0 \ldots, 0, 1, 0, \ldots, 0)$ with an entry 1 at the place i, k, and l for any $c_j = \{s_i, s_k, s_l\}$ in C,
- the points $(0, \ldots, 0)$, $(0, \ldots, 0, 1, 1, 0, 0, 0)$, and $(0, \ldots, 0, 0, 1, 1, 0, 0)$,
- the points $(0, \ldots, 0, -0.5, 0.5)$, $(0, \ldots, 0, 0.5, 0.5)$,
- the points $(0, \ldots, 0, c, c)$, $(0, \ldots, 0, -c, c)$, where c is a constant such that $c > 1 + (4B)/\epsilon \cdot \left(\mathrm{sgd}^{-1}(1 - \epsilon/(2B)) - \mathrm{sgd}^{-1}(\epsilon/(2B))\right)$.

Negative examples, i.e. the output shall be $< -\epsilon$, are

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 255-260

the points $(0,\ldots,0,1,0,\ldots,0)$ with an entry 1 at the place $i \leq n+3$,
the points $(0,\ldots,0,1,1,1,0,0)$, $(0,\ldots,0,-1.5,0.5)$, $(0,\ldots,0,1.5,0.5)$,
the points $(0,\ldots,0,1+c,c)$, $(0,\ldots,0,-1-c,c)$ with c as above.

Assume, the SSP is solvable. Consider the weights $\alpha = \beta = -1$, $\gamma = 0.5$, $\mathbf{a} = k \cdot (a_1,\ldots,a_n,1,-1,1,1,-1)$, $\mathbf{b} = k \cdot (b_1,\ldots,b_n,-1,1,-1,-1,-1)$, $a_0 = -0.5 \cdot k$, $b_0 = -0.5 \cdot k$, where k is a constant,

$$a_i = \begin{cases} 1 & \text{if } s_i \in S_1 \\ -2 & \text{otherwise} \end{cases} \quad \text{and} \quad b_i = \begin{cases} 1 & \text{if } s_i \in S_2 \\ -2 & \text{otherwise} \end{cases} .$$

For appropriate k this solves the loading problem with accuracy $\epsilon < 0.5$.

Assume, the loading problem is solvable. First, the cases 1, 3, and 4 are excluded. Then a solution of the SSP is constructed using the convexity of the positive region in case 2. Obviously, case 1 can be excluded.

Assume the classification is of case 4: We will consider only the last two dimensions, where the following problem is included: (We drop the first $n+3$ coefficients which are 0.) $(-0.5,0.5)$, $(0.5,0.5)$, (c,c), $(-c,c) \mapsto 1$ and $(-1.5,0.5)$, $(1.5,0.5)$, $(1+c,c)$, $(-1-c,c) \mapsto -1$ (see Fig.2a). Define $p_0 := \mathrm{sgd}^{-1}(\epsilon/(2B))$ and $p_1 := \mathrm{sgd}^{-1}(1-\epsilon/(2B))$. $\{\mathbf{x} \,|\, p_0 \leq \mathbf{a}^t\mathbf{x} + a_0 \leq p_1\}$ and $\{\mathbf{x} \,|\, p_0 \leq \mathbf{b}^t\mathbf{x} + b_0 \leq p_1\}$ are called the **a**- resp. **b**-relevant region. Outside, $\mathrm{sgd}(\mathbf{a}^t\mathbf{x} + a_0)$ resp. $\mathrm{sgd}(\mathbf{b}^t\mathbf{x} + b_0)$ can be substituted by a constant, the difference is at most $\epsilon/2$.

Since the points with second component 0.5 cannot be separated by one hyperplane, one point $(x, 0.5)$ with $x \in [-1.5, 1.5]$ exists inside the **a**- resp. **b**-relevant region. If the points (c,c) and $(1+c,c)$ were both outside the **a**-relevant region then they would be separated by any hyperplane with normal vector **b** which intersects the separating manifold outside the **a**-relevant region (see Fig.2b). The normal vector of the manifold is approximately $-\mathbf{a}/|\mathbf{a}|$ for large resp. small $\mathbf{b}^t\mathbf{x} + b_0$. Therefore we can find a hyperplane where both points are located on the same side. Contradiction. The same argumentation holds for $(-c,c)$ and $(-1-c,c)$. Therefore the diameter of the **a**-relevant region restricted to the last two dimensions is at least $c-1$. Consequently $a \leq (p_1 - p_0)/(c-1) = \epsilon/(4B)$ where $a = |(a_{n+4}, a_{n+5})|$.

If one of the points (c,c) and $(1+c,c)$ and one of the points $(-c,c)$ and $(-1-c,c)$ is contained in the **b**-relevant region, it follows $b \leq \epsilon/(4B)$ for $b = |(b_{n+4}, b_{n+5})|$. This leads to a contradiction: For $\mathbf{x}_1 = (c,c)$ and $\mathbf{x}_2 = (1+c,c)$ it is $|f(\mathbf{x}_1) - f(\mathbf{x}_2)| \leq 2|\alpha|\,|\mathbf{a}^t\mathbf{x}_1 - \mathbf{a}^t\mathbf{x}_2| + 2|\beta|\,|\mathbf{b}^t\mathbf{x}_1 - \mathbf{b}^t\mathbf{x}_2| \leq \epsilon$.

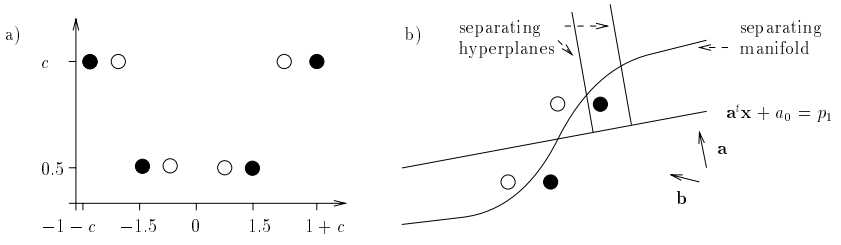

Figure 2: a) Classification problem; b) Outside the **b**-relevant region

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 255-260

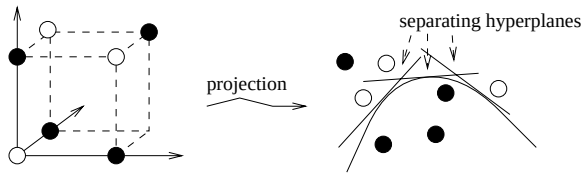

Figure 3: Classification problem; projection of the classification to the $\mathbf{a}/\mathbf{b}$-plane, at least one negative point is not classified correctly

If both points (c, c) and $(1 + c, c)$ or both points $(-c, c)$ and $(-1 - c, c)$ are outside the $\mathbf{b}$-relevant region, the difference of the values $\mathrm{sgd}(\mathbf{b}^t\mathbf{x} - b_0)$ with corresponding $\mathbf{x}$ is at most $\epsilon/(2B)$. The same contradiction results.

Assume the classification is of case 4: The classification includes in the dimensions $n + 1$ to $n + 3$ the problem depicted in Fig.3. The negative points are contained in a convex region, each positive point is separated by at least one tangential hyperplane of the separating manifold M. Consider the projection to a plane parallel to $\mathbf{a}$ and $\mathbf{b}$. Following the convex curve which describes M the signs of the coefficients of a normal vector can change at most once. But a normal vector oriented towards the positive region and separating a positive point has necessarily the signs $(+, +, -)$ for $(1, 1, 0)$, $(-, +, +)$ for $(0, 1, 1)$, and $(-, -, -)$ for $(0, 0, 0)$ in the dimensions $n + 1$ to $n + 3$. Contradiction.

Solution of the SSP: The classification is of case 2. We can assume $\gamma = -1$, $\alpha > 1$, and $\beta > 1$. Define $S_1 = \{s_i \mid a_i \text{ is positive }\}$, $S_2 = S - S_1$. It is

$$\begin{array}{lll} (i) & \alpha\,\mathrm{sgd}(a_0) + \beta\,\mathrm{sgd}(b_0) < 1 & \text{origin is } + \\ (ii) & \alpha\,\mathrm{sgd}(a_0 + a_i) + \beta\,\mathrm{sgd}(b_0 + b_i) > 1 & \text{point } s_i \text{ is } - \\ (iii) & \alpha\,\mathrm{sgd}(a_0 + a_i + a_j + a_k) + \beta\,\mathrm{sgd}(b_0 + b_i + b_j + b_k) < 1 & \end{array}$$

(iii) is valid for all points such that c exists with $c = \{s_i, s_j, s_k\}$.

Assume $c = \{s_i, s_j, s_k\}$ exists such that all three coefficients are positive. Necessarily, b_i, b_j, $b_k < 0$. In the components i, j, k the classification $(1, 0, 0)$, $(0, 1, 0)$, $(0, 0, 1) \mapsto -1$ and $(0, 0, 0)$, $(1, 1, 1) \mapsto 1$ is contained. The positive points are contained in a convex region, each negative point is separated by at least one tangential hyperplane of the separating manifold M. We project to a plane parallel to $\mathbf{a}$ and $\mathbf{b}$. Following the curve which describes M, the normal vector, oriented towards the positive region, is $\approx -\mathbf{a}/|\mathbf{a}|$, then the signs of each component of the normal vector change one time, finally it is $\approx -\mathbf{b}/|\mathbf{b}|$. But a vector where the three signs in dimension i, j, and k are equal cannot separate a negative point, further the sign in dimension i has to be negative if s_i is separated, the same is valid for j and k. Contradiction.

The same argumentation shows that at least one of b_i, b_j, and b_k is negative, i.e. at least one of a_i, a_j, and a_k is positive because of (i) and (ii). □

Note, that it is not obvious if the loading problem is contained in NP. This is due to the fact that the weights in the first layer and the precision that is necessary for the computation is not limited a priori.

ESANN'1998 proceedings - European Symposium on Artificial Neural Networks
Bruges (Belgium), 22-23-24 April 1998, D-Facto public., ISBN 2-9600049-8-1, pp. 255-260

Corollary 2 *The loading problem with accuracy* $\epsilon \in]0, 0.33[$ *and weight restriction* B *is NP-hard for any activation function* σ *which can be approximated by the* sgd *activation as follows:* $a, b, c, d \in \mathbb{R}$ *exist with* $|a| \leq B/2$ *and* $|a\sigma(bx+c)+d-\mathrm{sgd}(x)| < \epsilon/(4B)\ \forall x$. *This is valid for* $\tanh(x) = 2\,\mathrm{sgd}(x) - 1$.

5. Discussion

As a consequence training of a sigmoidal network can be very expensive for large input dimensions under the reasonable assumptions that the output weights are restricted and the classification is performed with a certain accuracy.

The restrictions are necessary because otherwise the geometric configuration that leads to a solution of the SSP (case 2) and a configuration which should be excluded (case 3) cannot be distinguished - they differ only slightly on a bounded region. Further, the restrictions offer the possibility to generalize the result to functions that can be approximated by the sigmoidal.

Unfortunately, it is not obvious how the proof of Theorem 1 can be expanded to other, even very simple activations. The main argument has been that the manifold limits a convex set. This is not true for such simple functions like a monotonous, piecewise linear function. Further, there exist activation functions with some nice properties where any consistent input set can be implemented and the loading problem is trivially solvable [6]. It is not obvious if the same holds with the additional condition concerning the accuracy and weights.

Finally, it remains unsolved if an NP-hardness result holds for architectures containing more layers than the 3-node network as formulated in [2].

References

[1] A. Blum and R. Rivest. Training a 3-node neural network is NP-complete. In *First Workshop on Comp. Learning Theory*. Morgan-Kaufmann, 1988.

[2] B. DasGupta, H. T. Siegelmann, and E. D. Sontag. On the complexity of training neural networks with continuous activation. *IEEE Transactions on Neural Networks*, 6, 1995.

[3] M. Garey and D. Johnson. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. W.H.Freeman and Company, 1979.

[4] K.-U. Höffgen. Computational limitations on training sigmoidal neural networks. *Information Processing Letters*, 46, 1993.

[5] J. Šíma. Back-propagation is not Efficient. *Neural Networks*, 6, 1996.

[6] E. D. Sontag. Feedforward nets for interpolation and classification. *Journal of Computer and System Sciences*, 45, 1992.

[7] M. Vidyasagar. *A Theory of Learning and Generalization*. Springer, 1997.

[8] V.H. Vu. On the infeasibility of training neural networks with small squared error. In *Neural Information Processing Systems*, 1997.