Training a Single Sigmoidal Neuron Is Hard

Jiří Šíma*
Department of Theoretical Computer Science,
Institute of Computer Science,
Academy of Sciences of the Czech Republic,
P. O. Box 5, 182 07 Prague 8, Czech Republic, *sima@cs.cas.cz*

Abstract

We first present a brief survey of hardness results for training feedforward neural networks. These results are then completed by the proof that the simplest architecture containing only a single neuron that applies a sigmoidal activation function $\sigma : \Re \longrightarrow [\alpha, \beta]$, satisfying certain natural axioms, e.g. the standard (logistic) sigmoid or saturated-linear function, to the weighted sum of n inputs is hard to train. In particular, the problem of finding the weights of such a unit that minimize the quadratic training error within $(\beta - \alpha)^2$ or its average (over a training set) within $5(\beta - \alpha)^2/(12n)$ of its infimum proves to be NP-hard. Hence, the well-known back-propagation learning algorithm appears to be not efficient even for one neuron which has negative consequences in constructive learning.

1 The Complexity of Neural Network Loading

Neural networks establish an important class of learning models that are widely applied in practical applications to solving artificial intelligence tasks (Haykin, 1999). The most prominent position among the successful neural learning heuristics is occupied by the *back-propagation* algorithm (Rumelhart, Hinton, & Williams, 1986) which is often used for training feedforward networks. This algorithm is based on the gradient descent method that minimizes the quadratic regression error of a network with respect to a given training data. For this purpose, each unit (neuron) in the network applies a differentiable activation function, e.g. the *standard (logistic) sigmoid*, to the weighted sum of its local inputs

*Research partially supported by project LN00A056 of The Ministry of Education of the Czech Republic.

(the so-called *sigmoidal* or *standard sigmoid unit*) rather than the discrete *Heaviside (threshold) function* with binary outputs (the so-called *Heaviside* or *threshold neuron*). However, the underlying optimization process appears to be very time-consuming even for small networks and training tasks. This was confirmed by an empirical study of the learning time required by the back-propagation algorithm to converge, which suggested its exponential scaling with the size of training sets (Tesauro, 1987) and networks (Tesauro & Janssens, 1988). Its slow convergence is probably caused by the inherent complexity of training the feedforward networks.

The first attempt to theoretically analyze the time complexity of learning by feedforward networks is due to Judd (1990) who introduced the so-called *loading problem* which is the problem of finding the weight parameters for a given fixed network architecture and a training task so that the network responses are perfectly consistent with all training data. For example, an efficient loading algorithm is needed for the proper PAC (Probably Approximately Correct) learnability (Blumer, Ehrenfeucht, Haussler, & Warmuth, 1989) besides the polynomial VC-dimension that the most common neural network models possess (Anthony & Bartlett, 1999; Roychowdhury, Siu, & Orlitsky, 1994; Vidyasagar, 1997). However, Judd (1990) proved the loading problem of feedforward networks to be NP-complete even if very strong restrictions are imposed on their architectures and training tasks. The drawback of Judd's proofs is in using quite unnatural network architectures with irregular interconnection patterns having a fixed input dimension and growing number of outputs, which do not appear in practice. On the other hand, his arguments are valid for practically all the common unit types including the standard sigmoid neurons. Eventually, Judd (1988) provided a polynomial-time loading algorithm for restricted shallow architectures whose practical applicability was probably ruled out by the hardness result for loading *deep* triangular networks (Šíma, 1994) which was further extended to pyramidal networks (de Souto & de Oliveira, 1999). Furthermore, Parberry (1992) proved a similar NP-completeness result for loading feedforward networks with irregular interconnections and only a small constant number of units. In addition, Wiklicky (1994) has shown that the loading problem for higher-order networks with integer weights is even algorithmically not solvable.

In order to achieve the hardness results for common layered architectures with complete connectivity between neighbor layers, Blum and Rivest in their seminal work (Blum & Rivest, 1992) considered the smallest conceivable two-layer network with only 3 binary neurons (two hidden and one output units) employing the Heaviside activation function. They proved the loading problem for such a *3-node network* with n inputs to be NP-complete and generalized the proof for a polynomial number of hidden units (in terms of n) when the output neuron computes logical AND (Blum & Rivest, 1992). Hammer (1998a) further replaced the output AND gate by the threshold unit while Kuhlmann (2000) achieved the proof for an output unit implementing any subclass of the Boolean functions depending on all the outputs from the hidden nodes. Lin and Vitter (1991) extended the NP-completeness result even for a 2-node cascade architecture with one hidden unit connected to the output neuron that also receives the inputs. Megiddo (1988), on the other hand, has shown that the loading problem for two-layer networks with a *fixed* number of real inputs and the Heaviside hidden nodes, and the output unit implementing an arbitrary Boolean function is solvable in polynomial time.

Much effort has been spent to generalize the hardness results also for continuous activation functions, especially for the standard sigmoid employed in the back-propagation heuristics for which the loading problem is probably at least algorithmically solvable (Macintyre & Sontag, 1993). DasGupta, Siegelmann, and Sontag (1995) proved that loading a 3-node network whose two hidden units apply the continuous *saturated-linear* activation function while the output threshold unit is used for dichotomic *classification*, is NP-complete. Further, Höffgen (1993) showed the NP-completeness of loading a 3-node network employing the standard sigmoid for *exact interpolation* but with the severe restriction to *binary weights*. A more realistic setting as concerns the back-propagation learning was first considered by Šíma (1996) who proved that loading a 3-node network with two standard sigmoid hidden neurons is NP-hard, although he assumed an additional constraint on the weights of the output threshold unit used for binary classification, which is satisfied e.g. when the output bias is zero. Hammer (1998b) replaced this constraint by requiring the output unit with bounded weights to respond with outputs that are in the absolute value greater than a given *accuracy*, which excludes a small output interval around zero from the binary classification. This approach also provides the hardness result for a more general class of sigmoidal activation functions than just the standard sigmoid. On the other hand, there exists a fixed sigmoidal activation function that, although artificial, has still appropriate mathematical properties and for which the feedforward networks are always loadable (Sontag, 1992).

The loading problem assumes a correct classification of all training data while in practice one is typically satisfied by the weights yielding a small training error. Therefore, the complexity of *approximately interpolating* a training set by feedforward neural networks with, in general, *real outputs* has been studied. Jones (1997) considered a 3-node network with n inputs, two hidden neurons employing any monotone Lipschitzian sigmoidal activation function (e.g. the standard sigmoid) and one linear output unit with bounded weights and zero bias. For such a 3-node network he proved that learning each pattern with real outputs from $[0,1]$ within a small absolute error $0 < \varepsilon < 1/10$ is NP-hard implying that the problem of finding the weights that minimize the *total* quadratic regression error within a fixed ε of its infimum (or absolutely) is also NP-hard. He further extended this NP-hardness proof to polynomial number k of sigmoidal hidden neurons (in terms of n) and a *convex* linear output unit (with nonnegative weights whose sum is 1) when the total quadratic error is required to be within $1/(16k^5)$ of its infimum (or within $1/(4k^3)$ of its infimum for the Heaviside hidden units).

In addition, Vu (1998) found the *relative* error bounds (with respect to the error infimum) for hard approximate interpolation which are independent on the training set size p by considering the *average* quadratic error that is defined as the total error divided by p. In particular, he proved that it is NP-hard to find the weights of a two-layer network with n inputs, k hidden sigmoidal neurons (satisfying certain Lipschitzian conditions) and one linear output unit with zero bias and positive weights such that for a given training data the average quadratic error is within a fixed bound of order $O(1/(nk^5))$ of its infimum. Moreover, for two-layer networks with k hidden neurons employing the Heaviside activations and one sigmoidal (or threshold) output unit, Bartlett and Ben-David (1999) improved this bound to $O(1/k^3)$ which is even independent of the input dimension. In the case of the threshold output unit used for classification, DasGupta and Hammer (2000) proved the same relative error bound $O(1/k^3)$ on the fraction of correctly classified training patterns, which is NP-hard to achieve for training sets of size $k^{3.5} \leq p \leq k^4$ related to number k of hidden units. They also showed that it is NP-hard to approximate this *success ratio* within a relative error smaller than $1/2244$ for 3-node networks with n inputs, two hidden standard sigmoid neurons and one output threshold unit (with bounded weights) exploited for the classification with accuracy $0 < \varepsilon < 0.5$. On the other hand, minimizing the *failure ratio*, i.e. the number of misclassified training patterns divided by the minimum possible number of misclassifications (when the presence of errors is assumed), within every constant $c > 1$ is NP-hard for feedforward threshold networks with zero biases in the first hidden layer (DasGupta & Hammer, 2000).

The preceding results suggest that training feedforward networks with *fixed* architectures is hard indeed. However, a possible way out of this situation might be the *constructive learning algorithms* that adapt the network architecture to a particular training task. It is conjectured that for successful generalization the network size should be kept small, otherwise a training set can easily be wired into the network implementing a look-up table (Sontag, 1992). A constructive learning algorithm usually requires an efficient procedure for minimizing the training error by adapting the weights of only a *single* unit that is being added to the architecture while the weights of the remaining units in the network are already fixed, e.g. the cascade-correlation learning architecture (Fahlman & Lebiere, 1990). Clearly, for a single binary neuron employing the Heaviside activation function the weights that are consistent with given training data can be found in polynomial time by linear programming provided that they exist although this problem restricted to binary weights is NP-complete (Pitt & Valiant, 1988) and also to decide whether the Heaviside unit can implement a Boolean function given in a disjunctive or conjunctive normal form is co-NP-complete (Hegedüs & Megiddo, 1996). However, such weights do not often exist but a good approximate solution would be sufficient for constructive learning. Nevertheless, several authors provided NP-completeness proofs for the problem of finding the weights for a single Heaviside unit so that the number of misclassified training patterns is at most a given constant (Höffgen, Simon, & Van Horn, 1995; Roychowdhury, Siu, & Kailath, 1995) which remains NP-complete even if the bias is assumed to be zero (Amaldi, 1991; Johnson & Preparata, 1978). In this case also minimizing the failure ratio within every constant $c > 1$ is NP-hard (Arora, Babai, Stern, & Sweedyk, 1997).

Hush (1999) further generalized these results for a single *sigmoidal* neuron by showing that it is NP-hard to minimize the training error under the L_1 norm strictly within 1 of its infimum. He conjectured that a similar result holds for the quadratic error corresponding to the L_2 norm which is used in the back-propagation learning. In the present paper this conjecture is proved. In particular, it will be shown that the issue of deciding whether there exist weights of a single neuron employing a sigmoidal activation function $\sigma : \Re \longrightarrow [\alpha, \beta]$ satisfying certain natural axioms (e.g. the standard sigmoid or saturated-linear function) so that the total quadratic error with respect to the training data is at most a given constant, is NP-hard. The presented reduction also provides an argument that the problem of finding the weights that minimize the quadratic training error within $(\beta-\alpha)^2$ or its average within $5(\beta-\alpha)^2/(12n)$ of its infimum is NP-hard. This implies that the popular back-propagation learning algorithm may be not efficient even for a single neuron and thus has negative consequences in constructive learning.

A preliminary version of this paper (Šíma, 2001) focused only on the standard sigmoid while the results presented now are valid for a more general class of sigmoidal activation functions.

2 Training a Sigmoidal Neuron

In this section the basic definitions regarding a sigmoidal neuron and its training will be reviewed. A single (perceptron) *unit (neuron)* with n real inputs $x_1, \ldots, x_n \in \Re$ first computes its real *excitation*

$$\xi = w_0 + \sum_{i=1}^{n} w_i x_i \tag{1}$$

where $\mathbf{w} = (w_0, w_1, \ldots, w_n) \in \Re^{n+1}$ is a real *weight* vector including a *bias* w_0. The *output* y is then determined by applying a nonlinear *activation function* σ to its excitation:

$$y = \sigma(\xi)\,. \tag{2}$$

Here we assume any *sigmoidal activation function* $\sigma : \Re \longrightarrow [\alpha, \beta]$ $(\alpha < \beta)$ that can be *normalized* as

$$\sigma_1(\xi) = \sigma(\xi_0 + s\xi) \qquad \text{for all } \xi \in \Re \tag{3}$$

where $\xi_0 \in \Re$ is a shift, $s \in \{-1, 1\}$ is a sign, and function $\sigma_1 : \Re \longrightarrow [\alpha, \beta]$ satisfies the following four axioms:

1. σ_1 is nondecreasing,

2. $\sigma_1(-\xi) - \alpha = \beta - \sigma_1(\xi)$ for all $\xi \in \Re$,

3. $\lim_{\xi \to -\infty} \sigma_1(\xi) = \alpha$,

4. $\xi_d \in \Re$ is given such that $\sigma_1(\xi_d) > \alpha$ and for all $\xi_1, \xi_2 \leq \xi_d$ and $\Delta > 0$, $\xi_1 < \xi_1 + \Delta \leq \xi_2 - \Delta < \xi_2$ implies $(\sigma_1(\xi_1 + \Delta) - \alpha)^2 - (\sigma_1(\xi_1) - \alpha)^2 \leq (\sigma_1(\xi_2) - \alpha)^2 - (\sigma_1(\xi_2 - \Delta) - \alpha)^2$.

Correspondingly, we call such a neuron the *sigmoidal unit*. Note that conditions 2, 3 imply also $\lim_{\xi \to +\infty} \sigma_1(\xi) = \beta$. These axioms are satisfied for most commonly used activation functions, especially for the *standard (logistic) sigmoid* $\sigma_s : \Re \longrightarrow [0,1]$ ($\alpha = 0$, $\beta = 1$):

$$\sigma_s(\xi) = \frac{1}{1 + e^{-\xi}} \tag{4}$$

which is employed in the widely applied *back-propagation* learning heuristics. In this case $\sigma_s \equiv \sigma_1$ is normalized and we can choose $\xi_d = \ln 2$ for which

$$(\sigma_s^2)''(\xi) = \frac{2e^{-\xi}\left(2e^{-\xi} - 1\right)}{\left(1 + e^{-\xi}\right)^4} \geq 0 \tag{5}$$

for all $\xi \leq \xi_d$ conforming to axiom 4. Another well-known example is the *saturated-linear* activation function $\sigma_\ell : \Re \longrightarrow [0,1]$ ($\alpha = 0$, $\beta = 1$):

$$\sigma_\ell(\xi) = \begin{cases} 0 & \text{for } \xi \leq 0 \\ \xi & \text{for } 0 < \xi < 1 \\ 1 & \text{for } \xi \geq 1 \end{cases} \tag{6}$$

which can be normalized as $\sigma_1(\xi) = \sigma_\ell(\frac{1}{2} + \xi)$ to satisfy axiom 2, where $\xi_d = \frac{1}{2}$ then meets axiom 4.

Furthermore, a *training set*

$$T = \{(\mathbf{x}_k, d_k);\, \mathbf{x}_k = (x_{k1}, \ldots, x_{kn}) \in \Re^n,\, d_k \in [\alpha, \beta],\, k = 1, \ldots, p\} \tag{7}$$

is introduced containing p pairs—*training patterns*, each composed of an n-dimensional real input $\mathbf{x}_k$ and the corresponding desired scalar output value d_k from $[\alpha, \beta]$ to be consistent with the range of activation function σ. Given a weight vector $\mathbf{w}$, the *quadratic training error*

$$E_T(\mathbf{w}) = \sum_{k=1}^{p} \left(y(\mathbf{w}, \mathbf{x}_k) - d_k\right)^2 = \sum_{k=1}^{p} \left(\sigma\left(w_0 + \sum_{i=1}^{n} w_i x_{ki}\right) - d_k\right)^2 \tag{8}$$

of the sigmoidal neuron with respect to training set T is defined as the difference between the actual outputs $y(\mathbf{w}, \mathbf{x}_k)$ depending on current weights $\mathbf{w}$ and the desired outputs d_k over all training patterns $k = 1, \ldots, p$ measured by the L_2 regression norm. The main goal of learning is to minimize the training error (8) in the weight space. The decision version for the problem of minimizing the error of the sigmoidal neuron with respect to a given training set is formulated as follows:

Minimum Sigmoidal-Unit Error (MSUE)
Instance: A training set T and a positive real number $\varepsilon > 0$.
Question: Is there a weight vector $\mathbf{w} \in \Re^{n+1}$ such that $E_T(\mathbf{w}) \leq \varepsilon$?

3 Minimizing the Training Error Is Hard

In this section the main result that training even a single sigmoidal neuron is hard will be proved:

Theorem 1 *MSUE is NP-hard.*

Proof: In order to achieve the NP-hardness result, a known NP-complete problem will be reduced to the MSUE problem in polynomial time (Garey & Johnson, 1979). In particular, the following *Feedback Arc Set* problem is employed which is known to be NP-complete (Karp, 1972):

Feedback Arc Set (FAS)
Instance: A directed graph $G = (V, A)$ and a positive integer $a \leq |A|$.
Question: Is there a subset $A' \subseteq A$ containing at most $a \geq |A'|$ directed edges such that the graph $G' = (V, A \setminus A')$ is acyclic?

The FAS problem was also exploited for a similar result concerning the Heaviside unit (Roychowdhury, Siu, & Kailath, 1995). However, the reduction is adapted here for general sigmoidal activation functions and its verification substantially differs.

3.1 The Construction of Graph G_r

Given a FAS instance $G = (V, A)$, a, a corresponding graph $G_r = (V_r, A_r)$ is first constructed so that every directed edge $(u, v) \in A$ in G is replaced by c_σ parallel oriented paths

$$P_{(u,v),h} = \{(u, u_v), (u_v, u_{vh1}), (u_{vh1}, u_{vh2}), \ldots, (u_{vh,r-1}, v)\} \tag{9}$$

for $h = 1, \ldots, c_\sigma$ in G_r sharing *only* their first edge (u, u_v) and vertices u, u_v, v as is depicted in Figure 1. The *constant* number c_σ of these parallel paths depends

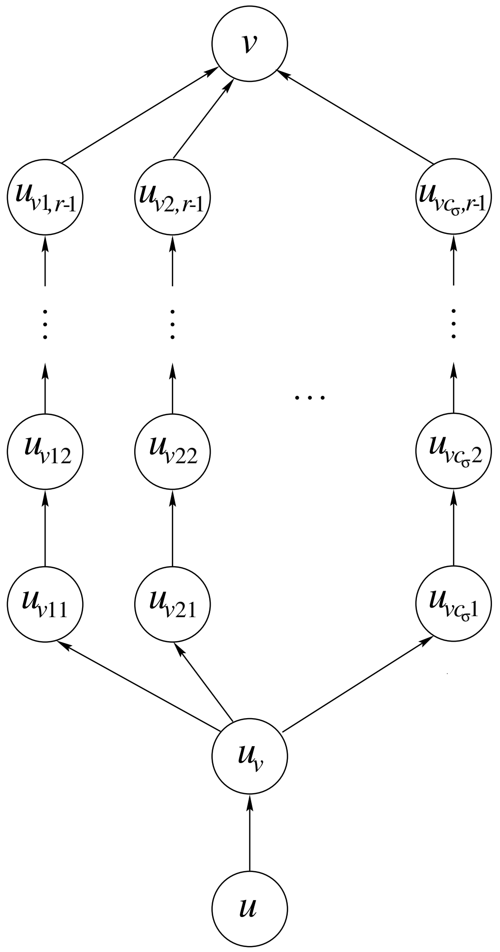

Figure 1: A subgraph of G_r with edges $A_{(u,v)}$ corresponding to (u,v) in G.

on the choice of activation function $\sigma : \Re \longrightarrow [\alpha, \beta]$:

$$c_\sigma = \left\lceil 2 \left(\frac{\beta - \alpha}{\sigma_1(\xi_d) - \alpha} \right)^2 \right\rceil \tag{10}$$

where ξ_d is given by axiom 4 for a corresponding normalized function σ_1. For example, $c_{\sigma_s} = 5$ for the standard sigmoid σ_s, or $c_{\sigma_\ell} = 2$ for the saturated-linear function σ_ℓ. In addition, each path $P_{(u,v),h}$ includes

$$r = 8a + 6 \tag{11}$$

additional vertices $u_v, u_{vh1}, u_{vh2}, \ldots, u_{vh,r-1}$ unique to $(u,v) \in A$, i.e. the subsets of edges

$$A_{(u,v)} = \bigcup_{h=1}^{c_\sigma} P_{(u,v),h} \tag{12}$$

corresponding to different $(u,v) \in A$ are pairwise disjoint. Thus,

$$\begin{aligned} V_r &= V \cup \{u_v;\, (u,v) \in A\} \\ &\quad \cup \{u_{vh1}, u_{vh2}, \ldots, u_{vh,r-1};\, (u,v) \in A,\, h = 1, \ldots, c_\sigma\} && (13) \\ A_r &= \bigcup_{(u,v)\in A} A_{(u,v)}\,. && (14) \end{aligned}$$

It follows that

$$\begin{aligned} |V_r| &= (c_\sigma(r-1)+1)|A| + |V| = (c_\sigma(8a+5)+1)|A| + |V| \\ &= O(|A|^2 + |V|) && (15) \\ |A_r| &= (c_\sigma r + 1)|A| = (c_\sigma(8a+6)+1)|A| = O(|A|^2) && (16) \end{aligned}$$

according to (11) (recall c_σ is a constant and $a \le |A|$). Obviously, the FAS instance G, a has a solution iff the FAS problem is solvable for G_r, a.

3.2 The Reduction to the MSUE Problem

Graph G_r is then exploited for constructing the corresponding MSUE instance which has $p = 2\,|A_r|$ training patterns, two for each edge in A_r, for a sigmoidal unit with $n = |V_r|$ inputs, one for each vertex in V_r. In particular, training set

$$\begin{aligned} T(G) \;=\; \Big\{ &\left(-s \cdot \mathbf{x}_{(i,j)}, \alpha\right), \left(s \cdot \mathbf{x}_{(i,j)}, \beta\right);\, (i,j) \in A_r, \\ &\mathbf{x}_{(i,j)} = \left(x_{(i,j),1}, \ldots, x_{(i,j),n}\right) \in \{-1,0,1\}^n \Big\} \qquad (17) \end{aligned}$$

contains one pair $(-s \cdot \mathbf{x}_{(i,j)}, \alpha), (s \cdot \mathbf{x}_{(i,j)}, \beta)$ for each edge $(i,j) \in A_r$ where $s \in \{-1, 1\}$ comes from (3) and

$$x_{(i,j),\ell} = \begin{cases} -1 & \text{for } \ell = i \\ 1 & \text{for } \ell = j \\ 0 & \text{for } \ell \neq i, j \end{cases} \qquad \ell = 1, \ldots, n\,, \qquad (i,j) \in A_r\,, \qquad (18)$$

i.e. input vector

$$\mathbf{x}_{(i,j)} = (0, \ldots, 0, \underset{\substack{\uparrow \\ i}}{-1}, 0, \ldots, 0, \underset{\substack{\uparrow \\ j}}{1}, 0, \ldots, 0) \in \{-1,0,1\}^n \qquad (19)$$

actually encodes oriented edge $(i,j) \in A_r$. In addition, the error in the MSUE instance is required to be at most

$$\varepsilon = (2a+1)(\beta - \alpha)^2\,. \qquad (20)$$

Clearly, the present construction of the corresponding MSUE instance can be achieved in a polynomial time in terms of the size of the original FAS instance according to (15) (16).

Now, the correctness of the reduction will be verified, i.e. it will be shown that the FAS instance has a solution iff the corresponding MSUE instance is solvable.

3.3 The MSUE Instance Is Solvable

So first assume that there exists a solution $A' \subseteq A$ of the FAS instance containing at most $a \geq |A'|$ directed edges making graph $G' = (V, A \setminus A')$ acyclic. Define a subset

$$A'_r = \{(u, u_v);\ (u, v) \in A'\} \tag{21}$$

containing $|A'_r| = |A'| \leq a$ edges from A_r. Clearly, graph $G'_r = (V_r, A_r \setminus A'_r)$ is also acyclic and hence its vertices $i \in V_r$ can be evaluated by *integers* w'_i so that any directed edge $(i, j) \in A_r \setminus A'_r$ satisfies $w'_i < w'_j$. The corresponding weight vector $\mathbf{w}$ is defined as

$$w_0 = \xi_0\,, \qquad w_i = K \cdot w'_i \quad \text{for } \ i \in V_r \tag{22}$$

where ξ_0 is introduced in (3) and $K > 0$ is a sufficiently large positive constant satisfying inequality

$$(\sigma_1(-K) - \alpha)^2 \leq \frac{(\beta - \alpha)^2}{2|A_r|} \tag{23}$$

(the existence of such K follows from axiom 3 for σ_1), which will be proved to be a solution for the MSUE instance. The error $E_{T(G)}(\mathbf{w})$ is expressed by using (8), (17), (18), and (22):

$$\begin{aligned} E_{T(G)}(\mathbf{w}) \ &= \sum_{(i,j)\in A_r} \left(\sigma\left(\xi_0 + s \cdot (w_i - w_j)\right) - \alpha\right)^2 \\ &\quad + \sum_{(i,j)\in A_r} \left(\sigma\left(\xi_0 + s \cdot (w_j - w_i)\right) - \beta\right)^2 \end{aligned} \tag{24}$$

which can further be rewritten as

$$\begin{aligned} E_{T(G)}(\mathbf{w}) \ &= \sum_{(i,j)\in A_r} \left(\sigma_1\left(w_i - w_j\right) - \alpha\right)^2 + \sum_{(i,j)\in A_r} \left(\sigma_1\left(w_j - w_i\right) - \beta\right)^2 \\ &= 2 \sum_{(i,j)\in A_r} \left(\sigma_1\left(w_i - w_j\right) - \alpha\right)^2 \end{aligned} \tag{25}$$

by using (3) and axiom 2 for σ_1. For $(i, j) \in A_r \setminus A'_r$ it holds

$$w_i - w_j = K(w'_i - w'_j) \leq -K < 0 \tag{26}$$

according to (22) where $w'_i - w'_j \leq -1$ due to w'_i, w'_j are integers. This implies

$$(\sigma_1(w_i - w_j) - \alpha)^2 \leq (\sigma_1(-K) - \alpha)^2 \leq \frac{(\beta - \alpha)^2}{2|A_r|} \quad \text{for } \ (i, j) \in A_r \setminus A'_r \tag{27}$$

by inequality (23) due to σ_1 is nondecreasing (axiom 1). Based on whether $(i, j) \in A'_r$ or $(i, j) \in A_r \setminus A'_r$, error (25) can be split into two parts that are upper bounded separately as follows:

$$\begin{aligned} E_{T(G)}(\mathbf{w}) &= 2 \sum_{(i,j)\in A'_r} \left(\sigma_1\left(w_i - w_j\right) - \alpha\right)^2 + 2 \sum_{(i,j)\in A_r \setminus A'_r} \left(\sigma_1\left(w_i - w_j\right) - \alpha\right)^2 \\ &\leq 2\,|A'_r|\,(\beta - \alpha)^2 + 2\,|A_r|\,\left(\sigma_1(-K) - \alpha\right)^2 \leq (2a + 1)(\beta - \alpha)^2 = \varepsilon \end{aligned} \tag{28}$$

by using $|A'_r| \leq a$, $(\sigma_1(\xi) - \alpha)^2 \leq (\beta - \alpha)^2$, $|A_r \setminus A'_r| \leq |A_r|$, and inequality (27). Hence, $\mathbf{w}$ is a solution of the MSUE problem.

3.4 The FAS Instance Is Solvable

On the other hand, assume that there exists a weight vector $\mathbf{w} \in \Re^{n+1}$ such that

$$E_{T(G)}(\mathbf{w}) \leq \varepsilon \, . \tag{29}$$

Define a subset of edges

$$A' = \{(u, v) \in A; \, w_u \geq w_v\} \subseteq A \tag{30}$$

in G. First observe that graph $G' = (V, A \setminus A')$ is acyclic since each vertex $u \in V \subseteq V_r$ is evaluated by a real weight $w_u \in \Re$ so that any directed edge $(u, v) \in A \setminus A'$ in G' satisfies $w_u < w_v$.

Moreover, it must be checked that $|A'| \leq a$. This is accomplished by proving that the error $E_{T(G)}(\mathbf{w})$ on training set $T(G)$ is bounded below:

$$E_{T(G)}(\mathbf{w}) > |A'| \cdot \frac{\varepsilon}{a+1} \tag{31}$$

which combined with (29) implies that $|A'| < a + 1$ or $|A'| \leq a$ equivalently. For this purpose, error $E_{T(G)}(\mathbf{w})$ is again expressed:

$$\begin{aligned} E_{T(G)}(\mathbf{w}) \;\; = & \sum_{(i,j) \in A_r} \left(\sigma \left(w_0 + s \cdot (w_i - w_j)\right) - \alpha\right)^2 \\ & + \sum_{(i,j) \in A_r} \left(\sigma \left(w_0 + s \cdot (w_j - w_i)\right) - \beta\right)^2 \end{aligned} \tag{32}$$

which can be rewritten as follows:

$$\begin{aligned} E_{T(G)}(\mathbf{w}) \;\; = & \sum_{(i,j) \in A_r} \left(\sigma_1 \left(w'_0 + w_i - w_j\right) - \alpha\right)^2 \\ & + \sum_{(i,j) \in A_r} \left(\sigma_1 \left(-w'_0 + w_i - w_j\right) - \alpha\right)^2 \end{aligned} \tag{33}$$

where substitution $w'_0 = s \cdot (w_0 - \xi_0)$ (i.e. $w_0 = \xi_0 + s \cdot w'_0$), equation (3), and axiom 2 for σ_1 are applied. According to disjoint union (14) the error is split among the contributions from the subsets of edges $A_{(u,v)}$ corresponding to edges $(u, v) \in A$ in G and further lower bounded by considering only the edges (u, v) from $A' \subseteq A$:

$$E_{T(G)}(\mathbf{w}) = \sum_{(u,v) \in A} EA_{(u,v)} \geq \sum_{(u,v) \in A'} EA_{(u,v)} \tag{34}$$

where

$$EA_{(u,v)} = \sum_{(i,j)\in A_{(u,v)}} \Big(\left(\sigma_1\left(w_0' + w_i - w_j\right) - \alpha\right)^2 + \left(\sigma_1\left(-w_0' + w_i - w_j\right) - \alpha\right)^2 \Big) \,. \tag{35}$$

In the following, inequality

$$EA_{(u,v)} > \frac{\varepsilon}{a+1} \tag{36}$$

will be proved for any $(u,v) \in A'$, which is sufficient for proving (31) according to (34). Clearly, e.g. $w_0' \geq 0$ can be assumed here without loss of generality because definition (35) with w_0' replaced by $-w_0'$ remains the same. For $(u,v) \in A'$ let $P_{(u,v)} \subseteq A_{(u,v)}$ be a path associated with the minimum error

$$EP_{(u,v)} = \sum_{(i,j)\in P_{(u,v)}} \Big(\left(\sigma_1\left(w_0' + w_i - w_j\right) - \alpha\right)^2 + \left(\sigma_1\left(-w_0' + w_i - w_j\right) - \alpha\right)^2 \Big) \tag{37}$$

among paths $P_{(u,v),h}$ for $h = 1, \ldots, c_\sigma$. Furthermore, denote by $(c,d),(e,f) \in P_{(u,v)}$ the edges on this path with the first two maximal so-called *decrements* $w_c - w_d \geq w_e - w_f \geq w_i - w_j$ for all $(i,j) \in P_{(u,v)} \setminus \{(c,d),(e,f)\}$. Note that at least one decrement along path $P_{(u,v)}$ must be nonnegative since $w_u - w_v \geq 0$ for $(u,v) \in A'$, i.e. $w_c - w_d \geq 0$.

3.4.1 Case $w_0' + w_e - w_f \geq \xi_d$:

First consider the case when $w_0' + w_e - w_f \geq \xi_d$, i.e.

$$\left(\sigma_1\left(w_0' + w_e - w_f\right) - \alpha\right)^2 \geq \left(\sigma_1\left(\xi_d\right) - \alpha\right)^2 \tag{38}$$

since σ_1 is nondecreasing (axiom 1). Recall that there are c_σ parallel paths for $(u,v) \in A'$ sharing only their first edge $(u,u_v) \in A_{(u,v)}$ (see Figure 1). Therefore, error $EA_{(u,v)}$ can be lower bounded by c_σ times partial contribution $(\sigma_1(w_0' + w_e - w_f) - \alpha)^2$ from single edge (e,f) that has the second maximal decrement $w_e - w_f$ on path $P_{(u,v)}$ with minimal error $EP_{(u,v)}$. Note that in this lower bound, edge (e,f) with the *second* maximal decrement on path $P_{(u,v)}$ is used since edge (c,d) with the maximal decrement might coincide with (u,u_v) that is shared by all parallel paths $P_{(u,v),h}$ for $h = 1, \ldots, c_\sigma$. On the other hand, if (e,f) coincides with (u,u_v) then edge $(c,d) \in P_{(u,v)} \setminus \{(u,u_v)\}$ could be used here whose decrement $w_c - w_d$ is still lower bounded by $w_e - w_f$. Hence,

$$\begin{aligned} EA_{(u,v)} &\geq c_\sigma \cdot \left(\sigma_1\left(w_0' + w_e - w_f\right) - \alpha\right)^2 \geq c_\sigma \cdot \left(\sigma_1\left(\xi_d\right) - \alpha\right)^2 \\ &\geq 2(\beta - \alpha)^2 > \frac{\varepsilon}{a+1} \end{aligned} \tag{39}$$

according to (38), (10), and (20).

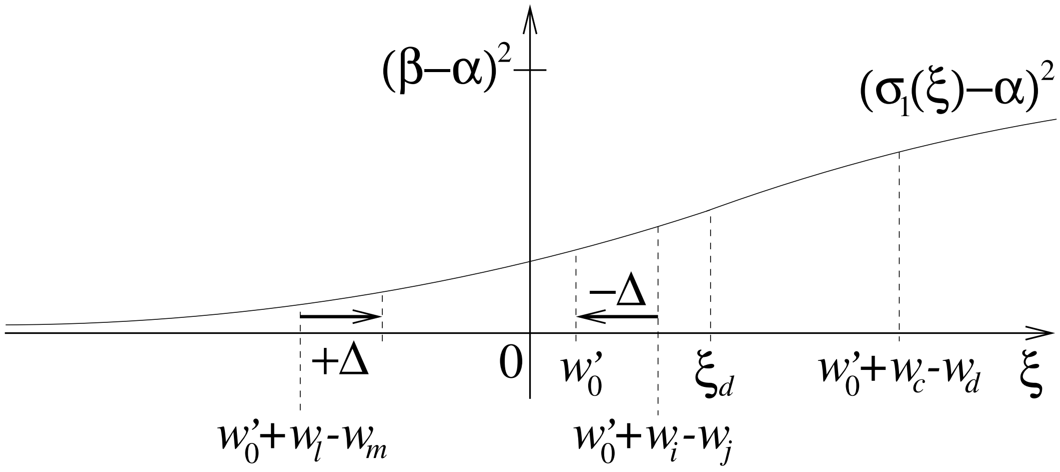

Figure 2: Weight re-labeling procedure.

3.4.2 Case $w_0' + w_e - w_f < \xi_d$:

On the other hand, suppose that $w_0' + w_e - w_f < \xi_d$. In this case vertices $i \in V_r \setminus V$ on path $P_{(u,v)}$ for $(u, v) \in A'$ will possibly be re-labeled with new weights $w_i' \in \Re$ so that the following three conditions are met:

1. $w_u \geq w_v$ are fixed,
2. $EP_{(u,v)}$ does not increase (possibly decreases),
3. (a) at most one edge $(c, d) \in P_{(u,v)}$ has positive decrement $w_c' - w_d' > 0$
   (b) *or* all edges $(i, j) \in P_{(u,v)}$ are associated with nonnegative decrements $w_i' - w_j' \geq 0$.

Note that cases 3a and 3b can happen simultaneously when all the decrements $w_i' - w_j' = 0$ are zero except for $w_c' - w_d' > 0$. Furthermore, notice that error $EP_{(u,v)}$ actually depends only on decrements $w_i - w_j$ rather than on the actual weights w_i, w_j. For example, these decrements can arbitrarily be permuted along path $P_{(u,v)}$ producing new weights whereas $EP_{(u,v)}$ and w_u, w_v do not change. Now, suppose that there exists an edge $(i, j) \in P_{(u,v)} \setminus \{(c, d)\}$ with positive decrement $0 < w_i - w_j \leq w_c - w_d$ together with an edge $(\ell, m) \in P_{(u,v)}$ associated with negative decrement $w_\ell - w_m < 0$ as depicted in Figure 2. Such a situation breaks condition 3 above and hence the underlying decrements are updated as follows:

$$w_i' - w_j' = w_i - w_j - \Delta \tag{40}$$
$$w_\ell' - w_m' = w_\ell - w_m + \Delta \tag{41}$$

where

$$\Delta = \min(w_i - w_j, w_m - w_\ell) > 0\,, \tag{42}$$

which leads to at least one new zero decrement. This can be achieved e.g. by permuting the decrements along path $P_{(u,v)}$ so that $w_i - w_j$ follows immediately after $w_\ell - w_m$ (this produces new weights but preserves $EP_{(u,v)}$) and by decreasing the weight of the middle vertex common to both decrements by Δ, which may influence error $EP_{(u,v)}$. However, axiom 4 of σ_1 can be employed for $\xi_1 = w'_0 + w_\ell - w_m < \xi_2 = w'_0 + w_i - w_j \leq w'_0 + w_e - w_f < \xi_d$, and for Δ defined in (42), which implies

$$\begin{aligned}&(\sigma_1(w'_0 + w_i - w_j) - \alpha)^2 + (\sigma_1(w'_0 + w_\ell - w_m) - \alpha)^2 \\ &\geq \left(\sigma_1(w'_0 + w'_i - w'_j) - \alpha\right)^2 + (\sigma_1(w'_0 + w'_\ell - w'_m) - \alpha)^2 \,. \qquad (43)\end{aligned}$$

Similarly, inequality (43) with $w'_0 \geq 0$ replaced by $-w'_0$ holds, which ensures that error $EP_{(u,v)}$ may only decrease while $w'_i - w'_j = 0$ or $w'_\ell - w'_m = 0$. By repeating this re-labeling procedure eventually at most one positive decrement $w'_c - w'_d > 0$ remains (case 3a) or all the negative decrements are eliminated (case 3b).

Furthermore,

$$w'_c - w'_d + \sum_{(i,j)\in P_{(u,v)}\setminus\{(c,d)\}} (w'_i - w'_j) = w_u - w_v \geq 0 \qquad (44)$$

due to $(u,v) \in A'$ which implies

$$w'_d - w'_c \leq \sum_{(i,j)\in P_{(u,v)}\setminus\{(c,d)\}} (w'_i - w'_j) \,. \qquad (45)$$

Hence,

$$w'_d - w'_c \leq w'_i - w'_j \qquad (46)$$

for all $(i,j) \in P_{(u,v)}$ since the decrements $w'_i - w'_j$ for $(i,j) \in P_{(u,v)} \setminus \{(c,d)\}$ in sum (45) are all nonpositive (case 3a) or all nonnegative (case 3b). According to (46), inequality (36) would follow from

$$\begin{aligned}EA_{(u,v)} \geq EP_{(u,v)} &\geq (\sigma_1(w'_0 + w'_c - w'_d) - \alpha)^2 + r \cdot (\sigma_1(w'_0 + w'_d - w'_c) - \alpha)^2 \\ &+ (\sigma_1(-w'_0 + w'_c - w'_d) - \alpha)^2 + r \cdot (\sigma_1(-w'_0 + w'_d - w'_c) - \alpha)^2 > \frac{\varepsilon}{a+1} \qquad (47)\end{aligned}$$

because there are r edges (i,j) on path $P_{(u,v)}$ except for (c,d), and σ_1 is nondecreasing (axiom 1). Particular terms of addition (47) can suitably be coupled so that it suffices to show

$$(\sigma_1(\xi) - \alpha)^2 + r \cdot (\sigma_1(-\xi) - \alpha)^2 = (\sigma_1(\xi) - \alpha)^2 + r \cdot (\beta - \sigma_1(\xi))^2 > \frac{\varepsilon}{2(a+1)} \qquad (48)$$

for any excitation $\xi \in \Re$ where axiom 2 of σ_1 is used. For this purpose, domain $\Re$ of nondecreasing σ_1 is split into two disjoint intervals X and $\Re \setminus X$ where

$$X = \left\{\xi \in \Re\,;\; (\sigma_1(\xi) - \alpha)^2 \leq \frac{\varepsilon}{2(a+1)}\right\} \,. \qquad (49)$$

Obviously, for $\xi \notin X$, desired inequality (48) follows from definition (49).

For $\xi \in X$, on the other hand, inequality

$$(\sigma_1(\xi) - \alpha)^2 \leq \frac{\varepsilon}{2(a+1)} = \frac{(2a+1)}{2(a+1)}(\beta - \alpha)^2 < (\beta - \alpha)^2 \tag{50}$$

holds according to (49) and (20), which implies

$$\sigma_1(\xi) \leq \alpha + (\beta - \alpha)\sqrt{\frac{2a+1}{2(a+1)}}\,, \tag{51}$$

and $\sigma_1(\xi) < \beta$, i.e.

$$(\beta - \sigma_1(\xi))^2 > 0\,. \tag{52}$$

Thus, if r is sufficiently large then inequality (48) is satisfied and the proof is complete. It will be proved that the choice (11) of r is sufficient. In this case, it will even be shown that

$$(\sigma_1(\xi) - \alpha)^2 + r \cdot (\beta - \sigma_1(\xi))^2 \geq (\beta - \alpha)^2 > \frac{\varepsilon}{2(a+1)}\,, \tag{53}$$

which reduces to

$$r \geq \frac{(\beta - \alpha)^2 - (\sigma_1(\xi) - \alpha)^2}{(\beta - \sigma_1(\xi))^2} = \frac{\beta - 2\alpha + \sigma_1(\xi)}{\beta - \sigma_1(\xi)} \tag{54}$$

for (52). For every $y \in Y$ where

$$Y = \sigma_1(X) = \{\sigma_1(\xi)\,;\, \xi \in X\} \subseteq \left[\alpha, \alpha + (\beta - \alpha)\sqrt{\frac{2a+1}{2(a+1)}}\right] \tag{55}$$

follows from (51), define function $F : Y \longrightarrow \Re$ by formula from (54):

$$F(y) = \frac{\beta - 2\alpha + y}{\beta - y}\,. \tag{56}$$

Function F proves to be increasing; namely, for any $y_1, y_2 \in Y$ inequality $F(y_1) < F(y_2)$ can be shown equivalent with $y_1 < y_2$ by using its explicit formula (56). Hence, according to (55) it suffices to show inequality (54) with $\sigma_1(\xi)$ replaced by

$$\alpha + (\beta - \alpha)\sqrt{\frac{2a+1}{2(a+1)}}\,, \tag{57}$$

i.e.

$$r \geq \frac{1 + \sqrt{\frac{2a+1}{2(a+1)}}}{1 - \sqrt{\frac{2a+1}{2(a+1)}}}\,. \tag{58}$$

Inequality (58) can further be rewritten as

$$\left(\frac{r-1}{r+1}\right)^2 \geq \frac{2a+1}{2(a+1)} \tag{59}$$

which can be verified by using (11). This completes the argument for (48) and consequently for (36) providing the proof that A' is a solution of the FAS problem. $\square$

3.5 The Relative Error

The proof of Theorem 1 also gives the NP-hardness result regarding the relative (average) error bounds:

Corollary 2 *Given a training set T containing $p = |T|$ training patterns, it is NP-hard to find a weight vector $\mathbf{w} \in \Re^{n+1}$ of the sigmoidal neuron with n inputs and activation function $\sigma : \Re \longrightarrow [\alpha, \beta]$ satisfying the underlying axioms (see Section 2), for which the quadratic error $E_T(\mathbf{w})$ with respect to T is within $(\beta-\alpha)^2$ of its infimum, or the average quadratic error $E_T(\mathbf{w})/p$ is within $5(\beta-\alpha)^2/(12n)$ of its infimum.*

Proof: Given a FAS instance $G = (V, A)$, a, a corresponding MSUE instance $T(G)$, ε is constructed according to (17), (20) in polynomial time. Assume that a weight vector $\mathbf{w}^* \in \Re^{n+1}$ could be found such that

$$E_{T(G)}(\mathbf{w}^*) \leq \inf_{\mathbf{w}\in\Re^{n+1}} E_{T(G)}(\mathbf{w}) + (\beta-\alpha)^2 \,. \tag{60}$$

The corresponding subset of edges $A^* \subseteq A$ making graph $G^* = (V, A \setminus A^*)$ acyclic can be then read from $\mathbf{w}^*$ according to (30). It will be proved in the following that $|A^*| \leq a$ iff the original FAS instance has a solution. This means that finding the weight vector $\mathbf{w}^*$ that satisfies (60) is NP-hard.

It suffices to show that for $|A^*| \geq a+1$ there is no subset $A' \subseteq A$ such that $|A'| \leq a$ and $G' = (V, A \setminus A')$ is acyclic since the opposite implication is trivial. On the contrary, suppose that such a subset A' exists. Thus a weight vector $\mathbf{w}' \in \Re^{n+1}$ corresponding to $A' \subseteq A$ could be defined by (22) that would lead to an error

$$E_{T(G)}(\mathbf{w}') \leq 2a(\beta-\alpha)^2 + 2\,|A_r|\,(\sigma_1(-K)-\alpha)^2 \tag{61}$$

according to (28). On the other hand, for $|A^*| \geq a+1$ inequality

$$\begin{aligned} \inf_{\mathbf{w}\in\Re^{n+1}} E_{T(G)}(\mathbf{w}) &\geq E_{T(G)}(\mathbf{w}^*) - (\beta-\alpha)^2 \\ &> |A^*| \cdot \frac{(2a+1)(\beta-\alpha)^2}{a+1} - (\beta-\alpha)^2 \geq 2a(\beta-\alpha)^2 \end{aligned} \tag{62}$$

follows from (60), (31), and (20). Hence, by axiom 3 of σ_1 there exists $K > 0$ such that

$$2\,|A_r|\,(\sigma_1(-K) - \alpha)^2 < -2a(\beta - \alpha)^2 + \inf_{\mathbf{w} \in \Re^{n+1}} E_{T(G)}(\mathbf{w}) \tag{63}$$

which provides contradiction $E_{T(G)}(\mathbf{w}') < \inf_{\mathbf{w} \in \Re^{n+1}} E_{T(G)}(\mathbf{w})$ by using (61).

Finally, it follows from the previous reduction that approximating the average quadratic error $E_T(\mathbf{w})/p$ within $5(\beta - \alpha)^2/(12n)$ of its infimum is also NP-hard due to $p < 12n/5$ that can be shown by using (15) and (16). □

4 Conclusions

The hardness results for loading feedforward networks are completed by the proof that approximate training of only a single sigmoidal neuron, e.g. by using the popular back-propagation heuristics, is hard. This suggests that the constructive learning algorithms that minimize the training error gradually by adapting unit by unit may also be not efficient. These results are valid for a general class of sigmoidal activation functions characterized by axioms including the most important standard (logistic) sigmoid.

References

Amaldi, E. (1991). On the complexity of training perceptrons. In T. Kohonen, K. Mäkisara, O. Simula, & J. Kangas (Eds.), *Proceedings of the First ICANN'91 International Conference on Artificial Neural Networks* (pp. 55–60). North-Holland, Amsterdam: Elsevier Science Publisher.

Anthony, M., & Bartlett, P. L. (1999). *Neural Network Learning: Theoretical Foundations.* Cambridge, UK: Cambridge University Press.

Arora, S., Babai, L., Stern, J., & Sweedyk, Z. (1997). The hardness of approximate optima in lattices, codes, and systems of linear equations. *Journal of Computer and System Sciences, 54*(2), 317–331.

Bartlett, P. L., & Ben-David, S. (1999). Hardness results for neural network approximation problems. In P. Fischer, & H.-U. Simon (Eds.), *Proceedings of the Fourth EuroCOLT'99 European Conference on Computational Learning Theory, LNAI 1572* (pp. 50–62). Berlin: Springer-Verlag.

Blum, A. L., & Rivest, R. L. (1992). Training a 3-node neural network is NP-complete. *Neural Networks, 5*(1), 117–127.

Blumer, A., Ehrenfeucht, A., Haussler, D., & Warmuth, M. K. (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM, 36*(4), 929–965.

DasGupta, B., & Hammer, B. (2000). On approximate learning by multi-layered feedforward circuits. In H. Arimura, S. Jain, & A. Sharma (Eds.), *Proceedings of the Eleventh ALT'2000 International Conference on Algorithmic Learning Theory, LNAI 1968* (pp. 264–278). Berlin: Springer-Verlag.

DasGupta, B., Siegelmann, H. T., & Sontag, E. D. (1995). On the complexity of training neural networks with continuous activation functions. *IEEE Transactions on Neural Networks, 6*(6), 1490–1504.

de Souto, M. C. P., & de Oliveira, W. R. (1999). The loading problem for pyramidal neural networks. *Electronic Journal on Mathematics of Computation.* Retrieved from `http://gmc.ucpel.tche.br/ejmc/`.

Fahlman, S. E., & Lebiere, C. (1990). The cascade-correlation learning architecture. In D. S. Touretzky (Ed.), *Advances in Neural Information Processing Systems (NIPS'89), 2* (pp. 524–532). San Mateo: Morgan Kaufmann.

Garey, M. R., & Johnson, D. S. (1979). *Computers and Intractability: A Guide to the Theory of NP-Completeness.* San Francisco: W. H. Freeman.

Hammer, B. (1998a). Some complexity results for perceptron networks. In L. Niklasson, M. Boden, & T. Ziemke (Eds.), *Proceedings of the Eight ICANN'98 International Conference on Artificial Neural Networks* (pp. 639–644). Berlin: Springer-Verlag.

Hammer, B. (1998b). Training a sigmoidal network is difficult. In M. Verleysen (Ed.), *Proceedings of the Sixth ESANN'98 European Symposium on Artificial Neural Networks* (pp. 255–260). Brussels: D-Facto Publications.

Haykin, S. (1999). *Neural Networks: A Comprehensive Foundation* (2nd ed.). Upper Saddle River, NJ: Prentice-Hall.

Hegedüs, T., & Megiddo, N. (1996). On the geometric separability of Boolean functions. *Discrete Applied Mathematics, 66*(3), 205–218.

Höffgen, K.-U. (1993). Computational limitations on training sigmoid neural networks. *Information Processing Letters, 46*(6), 269–274.

Höffgen, K.-U., Simon, H.-U., & Van Horn, K. S. (1995). Robust trainability of single neurons. *Journal of Computer and System Sciences, 50*(1), 114–125.

Hush, D. R. (1999). Training a sigmoidal node is hard. *Neural Computation, 11*(5), 1249–1260.

Johnson, D. S., & Preparata, F. P. (1978). The densest hemisphere problem. *Theoretical Computer Science, 6*(1), 93–107.

Jones, L. K. (1997). The computational intractability of training sigmoidal neural networks. *IEEE Transactions on Information Theory, 43*(1), 167–173.

Judd, J. S. (1988). On the complexity of loading shallow networks. *Journal of Complexity, 4*(3), 177–192.

Judd, J. S. (1990). *Neural Network Design and the Complexity of Learning.* Cambridge, MA: The MIT Press.

Karp, R. M. (1972). Reducibility among combinatorial problems. In R. E. Miller, & J. W. Thatcher (Eds.), *Complexity of Computer Computations* (pp. 85–103). New York: Plenum Press.

Kuhlmann, Ch. (2000). Hardness results for general two-layer neural networks. In N. Cesa-Bianchi, & S. A. Goldman (Eds.), *Proceedings of the Thirteenth COLT 2000 Annual Conference on Computational Learning Theory* (pp. 275–285). San Fransisco, CA: Morgan Kaufmann.

Lin, J.-H., & Vitter, J. S. (1991). Complexity results on learning by neural nets. *Machine Learning, 6*, 211–230.

Macintyre, A., & Sontag, E. D. (1993). Finiteness results for sigmoidal "neural" networks. In *Proceedings of the Twenty-Fifth STOC'93 Annual ACM Symposium on the Theory of Computing* (pp. 325–334). New York: ACM Press.

Megiddo, N. (1988). On the complexity of polyhedral separability. *Discrete Computational Geometry, 3*, 325–337.

Parberry, I. (1992). On the complexity of learning with a small number of nodes. In *Proceedings of the International Joint Conference on Neural Networks* (Vol. 3, pp. 893–898). Piscataway, NJ: IEEE Press.

Pitt, L., & Valiant, L. G. (1988). Computational limitations on learning from examples. *Journal of the ACM, 35*(4), 965–984.

Roychowdhury, V. P., Siu, K.-Y., & Kailath, T. (1995). Classification of linearly non-separable patterns by linear threshold elements. *IEEE Transactions on Neural Networks, 6*(2), 318–331.

Roychowdhury, V. P., Siu, K.-Y., & Orlitsky, A. (Eds.) (1994). *Theoretical Advances in Neural Computation and Learning.* Boston: Kluwer Academic Publishers.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533–536.

Šíma, J. (1994). Loading deep networks is hard. *Neural Computation, 6*(5), 842–850.

Šíma, J. (1996). Back-propagation is not efficient. *Neural Networks, 9*(6), 1017–1023.

Šíma, J. (2001). Minimizing the quadratic training error of a sigmoid neuron is hard. In N. Abe, R. Khardon, & T. Zeugmann (Eds.), *Proceedings of the Twelfth ALT'2001 International Conference on Algorithmic Learning Theory, LNAI 2225* (pp. 92–105). Berlin: Springer-Verlag.

Sontag, E. D. (1992). Feedforward networks for interpolation and classification. *Journal of Computer and System Sciences, 45*(1), 20–48.

Tesauro, G. (1987). Scaling relationships in back-propagation learning: dependence on training set size. *Complex Systems, 1*(2), 367–372.

Tesauro, G., & Janssens, B. (1988). Scaling relationships in back-propagation learning. *Complex Systems, 2*(1), 39–44.

Vu, V. H. (1998). On the infeasibility of training neural networks with small squared errors. In M. I. Jordan, M. J. Kearns, & S. A. Solla (Eds.), *Advances in Neural Information Processing Systems (NIPS'97), 10* (pp. 371–377). The MIT Press.

Vidyasagar, M. (1997). *A Theory of Learning and Generalization.* London: Springer-Verlag.

Wiklicky, H. (1994). The neural network loading problem is undecidable. In J. Shawe-Taylor, & M. Anthony (Eds.), *Proceedings of the First EuroCOLT'93 European Conference on Computational Learning Theory* (pp. 183–192). Oxford: Clarendon Press.