

COMPUTATIONAL EXPERIENCES OF A NOVEL GLOBAL ALGORITHM FOR OPTIMAL LEARNING IN MLP-NETWORKS

Carmine Di Fiore, Stefano Fanelli, Paolo Zellini

Dipartimento di Matematica
Università di Roma “Tor Vergata”
Via della Ricerca Scientifica, 00133 Roma, Italy
E-mail: fanelli@mat.uniroma2.it

ABSTRACT

This paper presents some numerical experiments related to a new global “pseudo-backpropagation” algorithm for the optimal learning of feedforward neural networks. The proposed method is founded on a new concept, called “non-suspiciousness”, which can be seen as a generalisation of convexity. The algorithm described in this work follows several adaptive strategies in order to avoid possible entrapments into local minima. In many cases the global minimum of the error function can be successfully computed. The paper performs also a useful comparison between the proposed method and a global optimisation algorithm of deterministic type well known in the literature.

1. INTRODUCTION

The aim of evaluating the global minimum of the error function of a MLP-network or, more in general, of a supervised network is considered crucial in the literature. Several approaches are possible and precisely:

i) determining the conditions under which the error function is local minima-free. Suitable hypotheses regarding both the network structure and the learning environment can be established (see [1],[4],[16],[17],[18]).

ii) computing the matrices of optimal weights by solving in the least-squares sense a set of systems of linear equations, derived by the condition of perfect learning ([13]) with possible additional assumptions.

iii) solving the inverse mapping problem associated to the Lyapunov function, updating the input vector determined by the pseudo-inverse of the gradient of the latter function ([19])

iiii) superimposing global optimisation algorithms of deterministic type in the computational scheme of a gradient descent method (see [8],[9])

In [11] an approach of class iii) was utilised to derive a preliminary version of a “pseudo-backpropagation”

algorithm. This approach is founded on a new definition, involving both the mathematical properties of the error function and the behaviour of the learning algorithm. The corresponding hypotheses, called “non-suspiciousness conditions”, represent a sort of generalisation of the concept of convexity. Roughly speaking, a “non-suspect” minimisation problem is characterised by the fact that, under general regularity assumptions on the error function, a “suitable” pseudo-backpropagation algorithm is able to compute the optimal solution, thereby avoiding unfair entrapments into local minima.

In [3] it was shown that the concept of non-suspect minimisation problem can be described in the frame of the theory of terminal attractors (see [2],[4],[20]). However, in [11] we proved that with a proper choice of the stepsizes the optimal algorithm can be obtained by a classical gradient descent-type scheme.

This work has the aim of refining the global algorithm introduced in [11] in order to deal with multi-dimensional problems more efficiently. A useful comparison between the proposed approach and a global optimisation algorithm proved in [8], [9] is presented. The latter method is particularly interesting since it represents one of the few rigorous global algorithms of deterministic type known in the literature.

2. THEORETICAL RESULTS

Let us consider the optimisation problem:

$$\min_{\mathbf{w} \in \mathbb{R}^n} E(\mathbf{w}) \quad (1)$$

where $E(\mathbf{w})$ is the error function of an MLP-network, i.e. typically

$$E(\mathbf{w}) = \frac{1}{2} \sum_{p=1}^P \sum_{i=1}^m \left(d_{p_i} - \frac{1}{1 + e^{-\sum_j w_{ij} o_{pj}}} \right)^2 \quad (2)$$

being:

- P the number of patterns
- m the number of output units
- d_{p_i} the desired output of unit i for pattern p
- o_{p_i} the computed output of unit i for pattern p
- w_{ij} the weight of the arc (j, i)
- $\mathbf{w} = (w_{ij})$ the matrix of weights
- $o_{p_i} = \frac{1}{1+e^{-\sum_j w_{ij} o_{p_j}}}$.

Let us suppose that (1) has a solution \mathbf{w}^* (global minimum) and let E_{min} denote the corresponding value of the function E . Moreover, we will assume that $E_{min} \approx 0$. Although the latter condition is restrictive, it is usually satisfied for the majority of MLP-networks. Let us now consider a general gradient descent scheme associated to (1):

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \lambda_k \nabla E_k, \quad \nabla E_k = \nabla E(\mathbf{w}_k). \quad (3)$$

The following assumptions, named ‘‘non-suspiciousness conditions’’ were originally introduced in [5], [14] and redefined in a more suitable form in [11].

Definition 1. *The non-suspiciousness conditions hold if $\exists \lambda_k$:*

1. $\forall \epsilon_a \in \mathbb{R}^+, \exists \epsilon_s: \|\nabla E_k\| > \epsilon_s$ during the gradient descent, apart from $k: E_k - E_{min} < \epsilon_a$;
2. $\lambda_k \|\nabla E_k\|^2 \leq \epsilon_a$;
3. $E \in C^2$ and has a limited Hessian ($\exists H > 0: \|\mathcal{H}(\mathbf{w})\| \leq H$).

The next theorem, proved in [14] and extended in [11], gives an indication on the choice of the one-dimensional stepsize λ_k guaranteeing the desired approximation ϵ_a and, consequently, the number of steps required to reach the optimal solution of (1).

Theorem 1. *Let the non-suspiciousness conditions hold for problem (1). Then, $\forall \epsilon_a \in \mathbb{R}^+$, the requested approximation is reached by choosing one-dimensional stepsizes λ_k no higher than*

$$\lambda^{**} = 2 \lfloor \frac{\epsilon_a}{H R_E} \rfloor \quad (4)$$

where: $R_E > E(\mathbf{w}_0) - E_{min}$, being \mathbf{w}_0 the initialisation weights.

Moreover, $E_{k^{**}} - E_{min} < \epsilon_a$ holds after

$$k^{**} = \frac{1}{2} \lceil \frac{H R_E^2}{\epsilon_a} \rceil \quad (5)$$

steps of the gradient descent iteration scheme (3).

Remark 1. *It is important to emphasise that the stepsize λ^{**} has to be considered as a function of the value ϵ_a . So, from an operational point of view, λ^{**} can be suitably modified during the implementation of the algorithm.*

By assuming the non-suspiciousness conditions the following inequation holds (see [6],[11]):

$$\|\nabla E_k\| > \min_k \frac{\sqrt{2\epsilon_a}}{\text{cond}(\mathcal{H}_k)} = \epsilon_s > 0 \quad (6)$$

being $\mathcal{H}_k = \mathcal{H}(\mathbf{w}_k)$, $\text{cond}(\mathcal{H}_k) = \|\mathcal{H}_k\| \|\mathcal{H}_k^{-1}\|$.

Remark 2. *Once again, we underline that ϵ_s is actually a function of ϵ_a . It follows that also the lower bound on the norm of ∇E_k can be adaptively modified during the algorithm.*

An operational way to utilise the inequality (6) is to try to verify the latter inequation $\forall k$. As a matter of fact, if $E_k \geq \epsilon_a$, then, by (6), the condition 1. of Definition 1 is satisfied whenever:

$$\frac{\sqrt{2\epsilon_a}}{\text{cond}(\mathcal{H}_k)} \geq \epsilon_a \quad (7)$$

Moreover, given k , if $E_k \geq \epsilon_a$ and the inequality (7) is verified for a suitable $\epsilon_a > 0$, then a sufficient condition to implement the optimal BP-algorithm is to determine λ_k :

$$\frac{\sqrt{2\epsilon_a}}{H \|\mathcal{H}_k^{-1}\|} \leq \frac{\sqrt{2\epsilon_a}}{\text{cond}(\mathcal{H}_k)} \leq \|\nabla E_k\| \leq \sqrt{\frac{\epsilon_a}{\lambda_k}} \quad (8)$$

In this way, in fact, all the non-suspiciousness conditions are satisfied in the iteration k for the value $\epsilon_a > 0$ and the inequalities (8) guarantee that:

$$E_k \geq \epsilon_a \implies \|\nabla E_k\| \geq \epsilon_a \quad (9)$$

By (9) any possible entrapment in local minima is avoided if ϵ_a is not too small.

According to [11] let $A.G.$ be the Armijo-Goldstein set of all $\lambda > 0$ such that:

$$\begin{cases} E(\mathbf{w}_k + \lambda \mathbf{d}_k) \leq E_k + c_1 \lambda \nabla E'_k \mathbf{d}_k, \\ \nabla E(\mathbf{w}_k + \lambda \mathbf{d}_k)' \mathbf{d}_k \geq c_2 \nabla E'_k \mathbf{d}_k. \end{cases} \quad (10)$$

being $0 < c_1 < c_2 < 1$ proper constants. Clearly $A.G.$ is not empty whenever \mathbf{d}_k is a descent direction.

In [15] it was introduced a new algorithm of conjugate-gradient-type, based on the automatic adaptability of learning rate and momentum term. This algorithm incorporated the computation of a ‘‘first order term’’ λ_k having the following form:

$$\lambda_k = a(n) \frac{E_k}{\|\nabla E_k\|^2} \quad (11)$$

being $a(n)$ a scaling parameter, dimension dependent. In the neighbourhood of the global minimum or, more generally, if the error function value E_k is sufficiently small, (11) is practically identical to the condition 2.

of Definition 1. On the other hand, in the basin of attraction of a local minimum the term (11) plays the typical role of a repeller (see [8],[9]). More in general, when the value of $\epsilon_a/(\|\nabla E_k\|^2)$ is too small and hence by (7), when \mathcal{H}_k is ill-conditioned, the upper bound on λ_k derived by the right inequality in (8) can be “relaxed” by utilising (11).

The constants c_1 and c_2 in (10) can be also modified to strengthen the capability of escaping from local minima. When $\epsilon_a/(\|\nabla E_k\|^2)$ is below a certain threshold δ , the values λ_k must be evaluated by simply setting $\lambda_k \in A.G.$. Clearly, the latter criterion is a possible alternative to (11).

It is important to point out that a non-suspect problem is characterised by a proper structure of local minima. More precisely, it is sufficient to verify condition (9) for a suitable subsequence $\{k_i\}$ such that the corresponding subsequence $E(\mathbf{w}_{k_i})$ describes a quasi-convex function approximating the error function $E(\mathbf{w})$.

Figure 1 shows an example of a favourable situation. From one hand, the basins of attraction of all the local minima are so narrow that jumping out from their neighbourhood can be easily performed by proper choices of the values λ_k . On the other hand, the global minimum is located in a wide and steep basin guaranteeing the desired convergence. The function illustrated is:

$$E(w) = \frac{w + 0.2}{w^8 + 1} \sin \frac{1}{w} \sin \frac{1}{w - 1.5} + 0.82$$

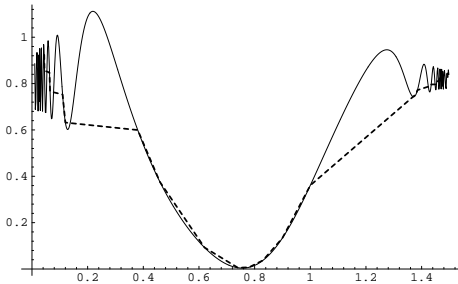


Figure 1: a non-suspect case

3. THE ALGORITHM

Let ϵ_c be an estimated value of quasi-convexity for the function $E(\mathbf{w})$, i.e. such that the level sets $\{\mathbf{w} : E(\mathbf{w}) \leq \epsilon_c\}$ are convex. Moreover, let a be a suitable constant dimension dependent and assume that ϵ be a sufficiently small value.

By using (7), (8), (10), (11), we can therefore define a novel heuristic optimal BP-algorithm.

Given \mathbf{w}_0 , compute $\nabla E(\mathbf{w}_0)$, $\mathcal{H}(\mathbf{w}_0)^{-1}$.

For $k = 0, \dots$:

choose $\epsilon_a \leq \frac{2}{[\text{cond}(\mathcal{H}_k)]^2}$

IF $E_k \geq \epsilon_a$ OR $\epsilon_c < E_k < \epsilon_a$

THEN

$(\mathbf{d}_k = -\nabla E_k$

$\tilde{\mathbf{w}}_{k+1} = \mathbf{w}_k + \frac{\epsilon_a}{\|\nabla E_k\|^2} \mathbf{d}_k$

$\bar{\mathbf{w}}_{k+1} = \mathbf{w}_k + \lambda_k \mathbf{d}_k, \lambda_k \in A.G., \lambda_k \leq \frac{\epsilon_a}{\|\nabla E_k\|^2}$

$\bar{\bar{\mathbf{w}}}_{k+1} = \mathbf{w}_k + \lambda_k \mathbf{d}_k, \lambda_k \in A.G.$

$\mathbf{w}_{k+1}^* = \mathbf{w}_k + a \frac{E_k}{\|\nabla E_k\|^2} \mathbf{d}_k$

IF $E(\tilde{\mathbf{w}}_{k+1}) < E_k$

THEN

$\mathbf{w}_{k+1} = \tilde{\mathbf{w}}_{k+1},$

IF $E(\tilde{\mathbf{w}}_{k+1}) \geq E_k$ AND $\|\tilde{\mathbf{w}}_{k+1} - \mathbf{w}_k\| > \epsilon$

THEN

$\mathbf{w}_{k+1} = \bar{\mathbf{w}}_{k+1}$

IF $E(\tilde{\mathbf{w}}_{k+1}) \geq E_k$ AND $\|\tilde{\mathbf{w}}_{k+1} - \mathbf{w}_k\| \leq \epsilon$

THEN

IF $\frac{\epsilon_a}{\|\nabla E_k\|^2} < \delta$

THEN

IF $\|\mathbf{w}_{k+1}^* - \mathbf{w}_k\| > \|\bar{\bar{\mathbf{w}}}_{k+1} - \mathbf{w}_k\|$ THEN

$\mathbf{w}_{k+1} = \mathbf{w}_{k+1}^*$

ELSE

$\mathbf{w}_{k+1} = \bar{\bar{\mathbf{w}}}_{k+1}$

IF $\frac{\epsilon_a}{\|\nabla E_k\|^2} \geq \delta$

$\mathbf{w}_{k+1} = \tilde{\mathbf{w}}_{k+1}$

$k=k+1)$

ELSE

$(\mathbf{d}_k = -\mathcal{H}_k^{-1} \nabla E_k$

$\mathbf{w}_{k+1} = \mathbf{w}_k + \lambda_k \mathbf{d}_k, \lambda_k \in A.G.$

$k=k+1$

UNTIL $\|\nabla E_{k+1}\| < TOL)$

The described algorithm is essentially based on the following computational scheme:

$\alpha)$ if $E_k \geq \epsilon_a$, $E_k \geq \epsilon_c$, by (9) any possible entrapment in local minima is avoided whenever ϵ_a is not too small (non-suspect problem).

$\beta)$ if $\frac{\epsilon_a}{\|\nabla E_k\|^2}$ is less than a suitable (small) acceptable threshold, then, by (11), the algorithm is able to escape from a possible entrapment. Typically, the latter problem takes place whenever \mathcal{H}_k is ill-conditioned and can be effectively overcome if $E_k \gg E_{min}$.

$\gamma)$ if $E_k < \epsilon_c$ the convergence to the global minimum is guaranteed and can be accelerated by an efficient Quasi-Newtonian (QN) method, if ϵ_a is not satisfactorily small.

$\delta)$ if $\forall k \geq k_0$, $\epsilon_c \leq E_k < \epsilon_a$, in a finite number of iterations the case $\gamma)$ can be applied.

The exact computation of \mathcal{H}_k^{-1} and the eventual evaluation of \mathcal{H}_k (when H is not a satisfactory upper bound) can be effectively approximated by utilising a particular QN-method, belonging to a class named \mathcal{LQN} , recently introduced in [12] and applied to MLP-networks in [7], [10]. We underline that the \mathcal{LQN} methods define inverse Hessian approximations in $O(n \log n)$

flops per step, by utilising matrices of suitable spaces \mathcal{L} , diagonalized by fast unitary transforms (see [12]).

4. NUMERICAL EXPERIMENTS

Let us consider firstly the well known 6-Hump Camelback Two-Dimensional Function (6HCTD)

$$E(w_1, w_2) = (4 - 2.1w_1^2 + w_1^4/3)w_1^2 + w_1w_2 + (4w_2^2 - 4)w_2^2 + 1.1.$$

Such function, originally studied in [8], is particularly interesting for the location of its critical points. In particular, it has six different local minima and the global minimum is achieved both in $(-.0898, .7126)$ and in the symmetrical point $(.0898, -.7126)$. Moreover, $E_{min} \approx 0$, so $E(w_1, w_2)$ can be used in the benchmark study. We have implemented our algorithm by using the same initial weights indicated in [8]. The corresponding results are illustrated in Tables 1-2.

k	E_k	$(\mathbf{w}_k)_1$	$(\mathbf{w}_k)_2$	$\ \nabla E_k\ $
0	6.873	-2	-1	12.353
1	5.525	-1.93	-.948	12.055
..
8	3.208	-1.58	-.534	.395
9	3.202	-1.63	-.58	.279
10	1.332	-.26	.052	1.985
..
17	.068	-.089	.716	.053
18	.068	-.0898	.7126	.017

Table 1: 6HCTD, $\mathbf{w}_0 = (-2, -1)$

It is important to point out that at iteration 9 the algorithm escapes from the local minimum $(-1.61, .56)$ by simply using the value λ_k derived by the condition 2 of Definition 1 (Non-suspiciousness) and is able to converge to the global minimum in the successive descent.

k	E_k	$(\mathbf{w}_k)_1$	$(\mathbf{w}_k)_2$	$\ \nabla E_k\ $
0	163.9	3.	2.	304.46
1	162.3	2.99	1.995	300.57
..
15	3.2	-1.62	-.574	.15
16	2.96	.93	.46	2.8
..
22	1.88	.63	.571	3.83
23	1.86	.63	.572	3.83
24	.35	.18	.688	1.06
..
29	.068	-.090	.715	.053

Table 2: 6HCTD, $\mathbf{w}_0 = (3, 2)$

At iteration 24 the value λ_k is computed by utilising (11). It appears, in fact, that in this case the entrapment around the point $(.63, .57)$ is due to the low value of ϵ_a (ill-conditioned Hessian). Notice that the function value E_k is sufficiently larger than E_{min} , thereby allowing an efficient escape. Similar results are obtained by utilising $\mathbf{w}_0 = (-3, -2)$ (see [8] for a useful comparison).

Let us consider now the classical N -dimensional Test Function (N-dT) [8], or, more precisely, a slightly modified version given by :

$$E(w_1, \dots, w_n) = \frac{1}{2} \sum_{j=1}^n (w_j^4 - 16w_j^2 + 5w_j) - c_0 n$$

being c_0 a fixed constant value.

This function has 2^n local minima and a global minimum in $\mathbf{w} = (-2.90354., -2.90354., \dots, -2.90354.)$. In Figure 2 we have depicted the 2-dimensional case.

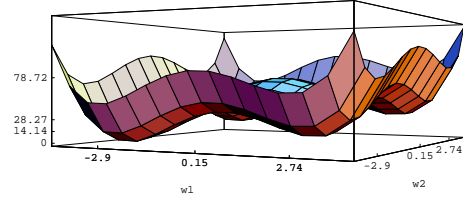


Figure 2: 2-d Test Function

This is a typical example of a suspect problem. The location and the structure of the local minima is in fact such that the algorithm is not always able to compute the global minimum even for low values of n . In the 2-d case, by using the initial weights $(1., 1.)$ the algorithm succeeds in escaping from two different entrapments before converging to the global minimum. The next Table 3 shows instead an unfavourable case, corresponding to the initial weights $(.5, -.5)$.

k	E_k	$(\mathbf{w}_k)_1$	$(\mathbf{w}_k)_2$	$\ \nabla E_k\ $
0	74.41	.5	-.5	13.36
1	72.79	.55	-.65	13.36
2	72.28	.58	-.68	13.87
3	71.77	.59	-.71	14.36
..
12	85.97	.72	-4.27	17.81
13	23.37	2.48	-3.53	35.33
..
18	14.16	2.75	-2.91	.29
19	14.15	2.75	-2.9	.11

Table 3: 2-dT, $\mathbf{w}_0 = (.5, -.5)$

It can be easily seen that the sequence generated by the algorithm is in this case trapped in the neighbourhood of the point (2.75,-2.90).

Final important remark For high-dimensional non-suspect problems ($n > 1000$) we suggest the following procedure:

Step 1. Determine a suitable initial vector \mathbf{w}_0 (f.i. near a "good" local minimum) by using an algorithm of LQN-type ([7],[12]), which is particularly efficient for large values of n .

Step 2. Apply the present algorithm to evaluate the global minimum.

5. REFERENCES

- [1] M.Bianchini, M.Gori, Optimal learning in artificial neural networks: A review of theoretical results, *Neurocomputing*, Vol. 13, pp. 313–346, 1996.
- [2] M.Bianchini, S.Fanelli, M.Gori, M.Maggini, Terminal attractor algorithms: a critical analysis, *Neurocomputing*, Vol. 15, pp. 3–13, 1997.
- [3] M.Bianchini, S.Fanelli, M. Gori, M.Protasi, Solving linear systems by a neural network canonical form of efficient gradient descent, *ICONIP-ANZIIS-ANNES'97*, Vol.1, Dunedin, pp. 531–534, 1997.
- [4] M.Bianchini, S.Fanelli, M.Gori, M.Protasi, Terminal Attractor Algorithms and the Class of Unimodal Loading Problems, *IMACS World Congress*, Vol.4, Berlin, pp. 277–282, 1997.
- [5] M. Bianchini, S. Fanelli, M.Gori, M.Protasi, Non-suspiciousness: a generalisation of convexity in the frame of foundations of Numerical Analysis and Learning, *IJCNN'98*, Vol.II, Anchorage, pp. 1619–1623, 1998.
- [6] M.Bianchini, S.Fanelli, M.Gori, Optimal algorithms for well-conditioned nonlinear systems of equations, *IEEE Transactions on Computers*, Vol. 50, pp. 689-698, 2001.
- [7] A.Bortoletti, C.Di Fiore, S.Fanelli, P.Zellini, A new class of quasi-newtonian methods for optimal learning in MLP-networks, *IEEE Transactions on Neural Networks*, submitted.
- [8] B.C. Cetin, J. Barhen, J.W. Burdick, Terminal Repeller Unconstrained Subenergy Tunneling(TRUST) for Fast Global Optimization, *Journal Optimization Theory and Applications*, Vol. 77, pp. 97–126, 1993.
- [9] B.C. Cetin, J. Barhen, J.W. Burdick, Global Descent Replaces Gradient Descent to Avoid Local Minima Problem in Learning with Artificial Neural Networks, *IEEE International Conf. on Neural Networks*, Vol. 2, San Francisco, pp. 836-842, 1993.
- [10] C.Di Fiore, S.Fanelli, P.Zellini, Matrix algebras in quasi-newtonian algorithms for optimal learning in multi-layer perceptrons, *ICONIP Workshop and Expo*, Dunedin, pp. 27–32, 1999.
- [11] C.Di Fiore, S.Fanelli, P.Zellini, Optimisation Strategies for Nonconvex Functions and Applications to Neural Networks, *ICONIP 2001*, Vol. 1, Shanghai, pp. 453–458, 2001.
- [12] C. Di Fiore, S. Fanelli, F. Lepore, P. Zellini, Matrix algebras in Quasi-Newton methods for unconstrained optimization, *Numerische Mathematik*, to appear.
- [13] M. Di Martino, S. Fanelli, M. Protasi, Exploring and Comparing the Best Direct Methods for the Efficient Training of MLP-Networks, *IEEE Transactions on Neural Networks*, Vol. 7, pp. 1497–1502, 1996.
- [14] P. Frasconi, S. Fanelli, M. Gori, M. Protasi, Suspiciousness of loading problems, *IEEE International Conf. on Neural Networks*, Vol. 2, Houston, pp. 1240–1245, 1997.
- [15] P. Geczy, S. Usui, Superlinear First Order Conjugate Gradient Learning Algorithm, *ICONIP 2001*, Vol.2, Shanghai, pp. 53–57, 2001.
- [16] M.Gori, A.Tesi, On the problem of local minima in backpropagation, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 14, pp. 76–85, 1992.
- [17] M. Gori, A.C. Tsoi, Comments on Local Minima Free Conditions in Multilayer Perceptrons, *IEEE Transactions on Neural Networks*, Vol. 9, 1998.
- [18] L.G.C.Hamey, XOR has no local minima: A case study in neural network error surface analysis, *Neural Networks*, Vol. 11, pp. 669–681, 1998.
- [19] S.Lee, R.M.Kil, Inverse mapping of continuous functions using local and global information, *IEEE Transactions on Neural Networks*, Vol. 5, pp. 409–423, 1994.
- [20] M.Zak, Terminal attractors in neural networks, *Neural Networks*, Vol. 2, pp. 259-274, 1989.