

Optimisation Strategies for Nonconvex Functions and Applications to Neural Networks

Carmin 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

In this paper the authors describe some useful strategies for nonconvex optimisation in order to determine the global minimum of the error function of a Multi-Layer Perceptron. The proposed approach is founded on a new concept, called “non-suspiciousness”, which can be seen as a generalisation of convexity.

Relations both with classical unconstrained optimisation results and with recent contributions in the field of supervised neural networks are examined. The preliminary numerical experiences show that the ideas behind the illustrated algorithm are interesting, although they require further investigations.

1. Introduction

The theoretical solvability of the optimal learning problem on Supervised Networks (SN) is still an open question. From a mathematical point of view, in fact, the convexity of the error function is, in general, the only hypothesis ensuring the convergence to the optimal solution of BackPropagation (BP) type algorithms. Apart from some weak relaxations of convexity assumptions (like f.i. quasi- or pseudo-convexity (see [14])), few classical general contributions are known in the literature and, unfortunately, with no practical interest in the field of SN. However, some interesting results specifically oriented to the solution of the minimisation problem of the error function were proved in the last ten years. It is important to mention in particular the general analysis for Multi-Layered Perceptrons (MLP) given in [12], in which some theoretical conditions in terms of learning environment and network architecture guarantee local minima free error surfaces and the important contribution of [13], showing that the classical XOR problem has no “real” local minima. Furthermore, the fundamental and innovative approach founded on the “natural gradient”, introduced by Amari (see f.i. [1], [2], [17]), was proved to be a powerful stochastic method to solve the critical problem of escaping from plateaus of the error surface, thereby ensuring an “effective steepest descent” in any situation. Since algorithms based on the natural gradient have locally a Quasi-

Newtonian (QN) behaviour, Amari’s theory represents an outstanding contribution towards the determination of the best algorithms for optimal learning, by utilising second order probabilistic computational tools (Fisher Information Matrix).

On the other hand, surprisingly enough, by a suitable use of classical deterministic gradient techniques, practitioners are able to perform optimal learning in a variety of problems, where the shape of the error function is far from satisfying any weak form of convexity. It is trivial to observe that, if the initialisation weights are assumed in the neighbourhood of the global minimum, the convergence to the optimal solution is simply due to the local convexity of the error surface. However, in some operational problems the latter condition does not hold. Consequently, there is apparently no mathematical evidence to justify the favourable result of the optimisation algorithm in these cases.

Is it possible to give a theoretical and rigorous explanation to this phenomenon ?

In [11],[5] it was introduced a new definition involving both the mathematical properties of the error function $E(\mathbf{w})$ and the behaviour of the learning algorithm. The corresponding hypotheses, named “non-suspiciousness conditions”, represent a generalisation of the concept of convexity. Roughly speaking, a “non-suspect” minimisation problem is characterised by the fact that, under reasonable regularity assumptions on the error function (much more general than classical convexity!) a “suitable” gradient descent algorithm is able to converge to the optimal solution, thereby avoiding any possible entrapment in local minima. By applying a particular second order gradient descent method, Powell (see [15]) showed that, in order to obtain a convergence result, the essential key is to fulfill a boundness condition involving the two current difference vectors $\mathbf{w}_{k+1} - \mathbf{w}_k$, $\nabla E(\mathbf{w}_{k+1}) - \nabla E(\mathbf{w}_k)$ in each iteration. The latter condition, which represents a sort of “weak form of discrete convexity”, implies the same type of regularity on the error function E if the non-suspiciousness hypotheses hold. Since in [10] it was proved that Powell’s approach can be generalised to a wide class of QN-methods, named \mathcal{LQN} , we have investigated further extensions of Powell’s results to classical BP algorithms, by as-

suming the non-suspiciousness conditions.

In [4] the concept of non-suspect minimisation problem was introduced to address those learning processes in which there exists a canonical gradient descent scheme that turns out to be the optimal algorithm. The essential characteristic of this “special” gradient method is founded on a suitable lower bound ϵ_s on the norm of $\nabla E(\mathbf{w}_k)$, which is satisfied for a finite number K_0 of iterations until all the function level sets become convex.

The present paper shows that the latter property can be extended under non-suspiciousness hypotheses to classical gradient descent algorithms. This extension is possible by applying the same procedure utilised for quadratic functions in [6], thereby deriving an interesting formula for the lower bound ϵ_s . The latter formula, which can be seen as a generalisation of classical Demmel’s distance to the nearest ill-posed problem (see [8]), shows a clear relationship between non-suspect problems and well-conditioned non linear systems of equations.

Moreover, this work investigates the connections between the mathematical assumptions characterising the non-suspiciousness conditions and the computational features of the \mathcal{LQN} methods, recently applied to MLP-networks in [9],[7].

2. Some results on Nonconvex Optimisation

Let us consider the general optimisation problem:

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

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, let us now consider the following gradient descent scheme associated to (1):

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \lambda_k \nabla E_k \quad (2)$$

where : $\nabla E_k = \nabla E(\mathbf{w}_k)$

The following assumptions will be referred to in this paper as “non-suspiciousness conditions” (see [11],[5]).

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. $E \in \mathcal{C}^2$ and has a limited Hessian, $(\exists H > 0: \|\mathcal{H}(\mathbf{w})\| \leq H)$.

By assuming solely condition 3., one can easily prove the following result

Theorem 1. *If $E \in \mathcal{C}^2$, $\|\mathcal{H}(\mathbf{w})\| \leq H$ and $0 < \lambda^* < \frac{2}{H}$, the iterative scheme:*

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \lambda^* \nabla E_k \quad (3)$$

is convergent to a stationary point of $E(\mathbf{w})$.

In [4] it was shown that, if $E(\mathbf{w})$ represents the error function of an MLP-network or, equivalently, the error function of a dynamic process described by a differential equation, the concept of non-suspect minimisation problem can be described in the frame of the theory of Terminal Attractors (see [18], [3]). More precisely, a non-suspect problem can be associated to the existence of a particular “canonical gradient descent” scheme that turns out to be the optimal algorithm. The next theorem, proved 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 by this special gradient descent to reach the optimal solution of (1).

Theorem 2. *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 canonical gradient descent iteration scheme.

Since the condition:

$$\frac{\epsilon_a}{R_E} < 1 \quad (6)$$

can be trivially assumed, from Theorem 1 we may immediately derive the following result, which extends Theorem 2 to classical gradient descent algorithms.

Theorem 3. *Let the non-suspiciousness conditions hold for problem (1). Then, $\forall \epsilon_a \in \mathbb{R}^+$, the requested approximation is reached by applying the gradient descent scheme (2) with the stepsize λ^{**} given by (4).*

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.*

Let us now study expressions of operational interest for the lower bound ϵ_s in condition 1. of Definition 1. Under the non-suspiciousness conditions, instead of considering (1), we may equivalently solve the following minimisation problem

$$\min_{\mathbf{w} \in \mathbb{R}^n} \mathcal{E}(\mathbf{w}) = \min_{\mathbf{w} \in \mathbb{R}^n} \frac{1}{2} \|\nabla E(\mathbf{w})\|^2 \quad (7)$$

it follows

$$\nabla \mathcal{E}(\mathbf{w})' = \nabla E(\mathbf{w})' \mathcal{H}(\mathbf{w}) \quad (8)$$

where $'$ denote the transpose operator.

Let us choose $\epsilon_a > 0$ and suppose $\mathcal{E}_k - \mathcal{E}_{\min} \geq \epsilon_a$, where $\mathcal{E}_k = \mathcal{E}(\mathbf{w}_k)$. As $\mathcal{E}_{\min} = 0$ for the problem (7), then

$$\mathcal{E}_k \geq \epsilon_a \quad (9)$$

so, if $\forall k$, $\mathcal{H}(\mathbf{w}_k)$ is invertible

$$\|\nabla \mathcal{E}(\mathbf{w}_k)\| \|\mathcal{H}(\mathbf{w}_k)^{-1}\| \geq \|\nabla E_k\| \quad (10)$$

hence

$$\|\nabla \mathcal{E}(\mathbf{w}_k)\| \geq \frac{\sqrt{2\epsilon_a}}{\|\mathcal{H}(\mathbf{w}_k)^{-1}\|}. \quad (11)$$

By (8), the inequation (11) implies

$$\|\nabla E_k\| \geq \frac{\sqrt{2\epsilon_a}}{\|\mathcal{H}(\mathbf{w}_k)\| \|\mathcal{H}(\mathbf{w}_k)^{-1}\|} = \epsilon_{s,k} \quad (12)$$

By Theorem 2 and Theorem 3 there exists only a finite set K_0 of integers k :

$$E_k - E_{\min} \geq \epsilon_a \quad (13)$$

therefore, by setting

$$\epsilon_s = \min_{k \in K_0} \epsilon_{s,k} - \epsilon_0 \quad (14)$$

we obtain, being ϵ_0 arbitrarily small

$$\|\nabla E_k\| > \epsilon_s. \quad (15)$$

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.*

From inequality (11), which is an extension of a result proved in [6] (see Lemma 4.1), and by condition 3 of Definition 1 we have:

$$\|\nabla E_k\| \geq \frac{\sqrt{2\epsilon_a}}{H \|\mathcal{H}(\mathbf{w}_k)^{-1}\|} \quad (16)$$

The inequation (16) can be seen as a nonlinear generalisation of classical Demmel's distance for a given matrix A [8]:

$$d(A, IP) = \frac{\|A\|_F}{\text{cond}(A)} \geq \frac{1}{\|A^{-1}\|_s} \quad (17)$$

where IP denotes the set of all non-invertible real matrices and $\|\cdot\|_F$, $\|\cdot\|_s$ are the Frobenius and the spectral norm, respectively.

(16) clearly indicates that the level of difficulty of problem (1) can be expressed as a function of the norm of inverse of the Hessian, thereby showing a relationship between well-posed problems and non-suspect ones.

Particularly interesting is the case in which $E_{\min} = 0$. As a matter of fact, if $E_k \geq \epsilon_a$, then, by (12), the condition 1. of Definition 1 is satisfied whenever:

$$\frac{\sqrt{2\epsilon_a}}{\text{cond}(\mathcal{H}_k)} \geq \epsilon_a \quad (18)$$

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

3. Applications to MLP-networks

Let E in (1) be the error function of an MLP-network, i.e. typically

$$E(\mathbf{w}) = \frac{1}{2} \sum_{p=1}^P \sum_{i=1}^m \left(d_{p_i} - \frac{1}{1 + e^{-\sum_j w_{ij} o_{p_j}}} \right)^2 \quad (19)$$

where:

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

being: $o_{pi} = \frac{1}{1+e^{-\sum_j w_{ij} o_{pj}}}$.

It follows obviously that $E \geq 0$. Moreover, we will assume that $E_{min} = 0$. Although the latter condition is restrictive, it is usually satisfied for the majority of MLP-networks. From an historical point of view, it is interesting to observe that the original BP-method proposed in [16] suggested the use of a small and constant learning rate η^* during the implementation of the algorithm. Since in the BP computational scheme learning rates represent the neural interpretation of stepsizes λ_k , the latter suggestion was apparently in contrast to the classical gradient descent scheme (2). The favourable results obtained by practitioners can be rigorously explained by Theorem 1, which assures in a suitable compact set of the weight-space the fulfilment of the inequality $\eta^* < \frac{2}{H}$.

Given k , if $E_k \geq \epsilon_a$ and the inequality (18) is verified for a suitable $\epsilon_a > 0$, then a sufficient condition to implement the optimal BP-algorithm is to determine λ_k :

$$\frac{\sqrt{2\epsilon_a}}{H\|\mathcal{H}_k^{-1}\|} \leq \frac{\sqrt{2\epsilon_a}}{cond(\mathcal{H}_k)} \leq \|\nabla E_k\| \leq \sqrt{\frac{\epsilon_a}{\lambda_k}} \quad (20)$$

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 (20) guarantee that :

$$E_k \geq \epsilon_a \implies \|\nabla E_k\| \geq \epsilon_a \quad (21)$$

According to [7], [10] let $A.G.$ be the Armijo-Goldstein set of all $\lambda > 0$ such that:

$$\begin{cases} E(\mathbf{w}_k + \lambda \mathbf{d}_k) \leq E_k + c_1 \lambda \nabla E'_k \mathbf{d}_k \\ \nabla E(\mathbf{w}_k + \lambda \mathbf{d}_k)' \mathbf{d}_k \geq c_2 \nabla E'_k \mathbf{d}_k. \end{cases} \quad (22)$$

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

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 ϵ be a sufficiently small value.

By using the inequalities (12), (18) and (20), we can therefore state the following 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 $\epsilon_a \leq \frac{2}{[cond(\mathcal{H}_k)]^2}$

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

THEN

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

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

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

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

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

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

It is important to point out that, when $\frac{\epsilon_a}{\|\nabla E_k\|^2}$ is below a certain threshold δ , the values λ_k must be evaluated by simply setting $\lambda_k \in A.G.$. The described algorithm is essentially based on the following computational scheme:

α) if $E_k \geq \epsilon_a$, $E_k \geq \epsilon_c$, by (21) any possible entrapment in local minima is avoided if ϵ_a is not too small (non-suspect problem).

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

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

If the problem is non-suspect, then $\exists C > 0$ such that $cond(\mathcal{H}_k) \leq C$ for a sufficient number of iterations and the corresponding values $\{\epsilon_{s,k}\}$, defined in (12), are bounded below by a "satisfactory" ϵ_s .

Moreover, a non-suspect problem is characterized by a favourable structure of local minima. More precisely, the cardinality and the location of local minima must be such that with a proper choice of ϵ_a the correct values of λ_k can be effectively determined. So:

Remark 3. The values of ϵ_a associated to the sequence $\{\epsilon_{s,k}\}$, defined in (12), represent the crucial point for a successful implementation of the described algorithm.

On the other hand:

Remark 4. The estimation of a suitable value ϵ_c is useful to improve the efficiency of the algorithm but it is not essential.

The exact computation of \mathcal{H}_k and \mathcal{H}_k^{-1} , which is required in each iteration, can be efficiently approximated by utilising a particular QN-method, belonging to a class named \mathcal{LQN} , recently introduced in [10] and applied to MLP-networks in [9],[7]. Notice that all the \mathcal{LQN} methods have an $O(n \log n)$ computational complexity per step (see[10]).

In particular, the sequence of Hessian inverses $\{\mathcal{H}_{k+1}^{-1}\}$ in the above algorithm can be replaced with the sequence $\{B_{k+1}\}$ defined by the iterative formula

$$B_{k+1} = \mathcal{L}_{B_k} - \frac{1}{\mathbf{y}'_k \mathbf{s}_k} (\mathcal{L}_{B_k} \mathbf{y}_k \mathbf{s}'_k + \mathbf{s}_k \mathbf{y}'_k \mathcal{L}_{B_k}) + \left(1 + \frac{\mathbf{y}'_k \mathcal{L}_{B_k} \mathbf{y}_k}{\mathbf{y}'_k \mathbf{s}_k} \right) \frac{\mathbf{s}_k \mathbf{s}'_k}{\mathbf{y}'_k \mathbf{s}_k} \quad (23)$$

being $\mathbf{s}_k = \mathbf{w}_{k+1} - \mathbf{w}_k$, $\mathbf{y}_k = \nabla E_{k+1} - \nabla E_k$ and \mathcal{L}_{B_k} the best least squares fit to B_k from a space \mathcal{L} of matrices simultaneously diagonalised by a fast unitary transform. A similar formula can be utilised for the (eventual) computation of the sequence of Hessian $\{\mathcal{H}_k\}$.

By assuming that

$$\frac{\|\mathbf{y}_k\|^2}{\mathbf{y}'_k \mathbf{s}_k} \leq H \quad (24)$$

and by adding few simple additional hypotheses, $\liminf \|\nabla E_k\| = 0$ for the “nonsecant” \mathcal{LQN} algorithms (see [10] Lemma 5.2). We underline that if $E \in \mathcal{C}^2$ the inequality (24) implies the condition 3. of Definition 1.

4. Preliminary numerical results

Let us consider the same test-problem studied in [3], i.e. the one-dimensional error function

$$E(w) = w^6 - 2w^4 + \frac{28}{27}w^2 \quad (25)$$

which has a global minimum in $w = 0$ and two symmetrical local minima in $w = \pm \frac{\sqrt{6+2\sqrt{2}}}{3} \approx \pm 1$. By perturbing the coefficients, the shape of $E(w)$ can be suitably modified. Here, as a preliminary investigation, we apply the algorithm to the function $30E(w)$ with two different initialisation weights.

Since $w \in \mathbb{R}^1$, $\forall k$, by (18) ϵ_a must satisfy the inequality

$$\epsilon_a \leq 2 \quad (26)$$

Moreover, we assume $\epsilon = 0.1$ and we set $\epsilon_c = 1$, even if other choices are of course allowed.

Table 1 shows the performances of our algorithm for $w_0 = -1.048$.

k	w_k	E_k	∇E_k
0	-1.048	1.537	-16.478
1	-.926	1.469	10.336
2	-1.041	1.429	-14.005
3	-.937	1.37	9.166
4	-1.03	1.303	-10.69
5	-.843	2.567	14.69
6	-.979	1.113	2.269
7	-.999	1.11	-2.104
8	-.049	0.075	-3.042
9	-4.6e-4	6.7e-6	-.028
10	-3.8e-10	.0	-2.3e-8

The convergence to the global minimum is achieved in 10 iterations. At iteration 8 the algorithm escapes from the neighbourhood of the local minimum $w = -\frac{\sqrt{6+2\sqrt{2}}}{3}$. In the iterations 9 and 10 the \mathcal{LQN} -method is utilised to speed up the convergence; since $w \in \mathbb{R}^1$, obviously the \mathcal{LQN} -method is equivalent to classical Newton procedure.

By applying the following threshold criterion

$$\text{IF } \left(\frac{\epsilon_a}{\|\nabla E_k\|^2} \right) < \delta \text{ THEN } \lambda_k \in A.G. \quad (27)$$

being $\delta = 2e - 2$, the convergence can be furtherly accelerated, as shown in Table 2.

k	w_k	E_k	∇E_k
0	-1.048	1.537	-16.478
1	-1.002	1.115	-2.641
2	-.244	1.652	-11.86
3	.233	1.514	11.579
4	-.215	1.313	-11.077
5	.144	.62	8.258
6	.011	.004	0.706
7	5.6e-6	9.9e-10	3.5e-4

Iterations 6 and 7 are performed with the \mathcal{LQN} -method. Finally, in Table 3 are illustrated the results of our algorithm for $w_0 = -2$.

k	w_k	E_k	∇E_k
0	-2.	1084.4	-3964.4
1	1.786	462.54	2015.3
2	-1.473	91.638	-573.6
3	1.187	8.665	96.888
4	-.995	1.103	-1.137
5	.158	.742	8.913
6	.014	.007	.932
7	1.3e-5	5.3e-9	8.1e-4

In the above table the criterion (27) is used in the first three iterations.

In the general case of $\mathbf{w} \in \mathbb{R}^n$ the behaviour of the values

$$\epsilon_a \leq \frac{2}{[\text{cond}(\mathcal{H}_k)]^2} \quad (28)$$

is crucial to distinguish non-suspect problems from “suspect” (i.e. difficult) ones. Once again, we stress that the values of the sequence $\frac{\epsilon_a}{\|\nabla E_k\|^2}$ in the neighbourhood of local minima play an important role.

Numerical experiences regarding the multi-dimensional case are the object of our current research and will be described in a future work.

REFERENCES

[1] S.Amari, *Backpropagation and stochastic gradient descent method*, Neurocomputing **5** (1993), 185-196.

[2] S.Amari, *Natural gradient works efficiently in learning*, Neural Computation **10** (1998), 251-276.

[3] M.Bianchini, S. Fanelli, M.Gori, M.Maggini, *Terminal attractor algorithms: a critical analysis*, Neurocomputing **15** (1997), 3-13.

[4] 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, Dunedin **1** (1997), 531-534.

[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, Anchorage **II** (1998), 1619-1623.

[6] M.Bianchini, S.Fanelli, M.Gori, *Optimal algorithms for well-conditioned nonlinear systems of equations*, IEEE Transactions on Computers, to appear.

[7] A.Bortoletti, C.Di Fiore, S.Fanelli, P.Zellini, *A new class of quasi-newtonian methods for optimal learning in MLP-networks*, submitted.

[8] J.Demmel, *On condition numbers and the distance to the nearest ill-posed problems*, Numerische Mathematik **51** (1987), 251-287.

[9] 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 (1999), 27-32.

[10] C.Di Fiore, S.Fanelli, F.Lepore, P.Zellini, *Matrix algebras in quasi-Newton methods for unconstrained minimization*, submitted.

[11] P.Frasconi, S.Fanelli, M.Gori, M.Protasi, *Suspiciousness of loading problems*, IEEE International Conf. on Neural Networks, Houston **2** (1997), 1240-1245.

[12] M.Gori, A.Tesi, *On the problem of local minima in backpropagation*, IEEE Transactions on Pattern Analysis and Machine Intelligence **14** (1992), 76-85.

[13] L.G.C.Hamey, *XOR has no local minima: A case study in neural network error surface analysis*, Neural Networks **11** (1998), 669-681.

[14] M.Minoux, *Mathematical Programming: Theory and Algorithms*, Wiley and Sons (1986).

[15] M.J.D.Powell, *Some global convergence properties of a variable metric algorithm for minimization without exact line searches*, Nonlinear Programming, SIAM-AMS (R.W.Cottle, C.E.Lemke, ed) **9** (1976), 53-72.

[16] D.E.Rumelhart, J.L.McClelland, *Parallel distributed processing: exploration in the microstructure of cognition*, Mitpress **1** (1986).

[17] H.Yang, S.Amari, *Natural gradient descent for training Multi-Layer Perceptrons*, IEEE Transactions on Neural Networks, submitted.

[18] M.Zak, *Terminal attractors in neural networks*, Neural Networks **2** (1989), 259-274.