

Solving Linear Systems by a Neural Network Canonical Form of Efficient Gradient Descent

M. Bianchini¹, S. Fanelli², M. Gori³, *Member IEEE*, and M. Protasi²

¹Dipartimento di Sistemi e Informatica, Università di Firenze (Italy)
monica@mcculloch.ing.unifi.it

² Dipartimento di Matematica, Università di Roma “Tor Vergata” (Italy)
fanelli@mat.utovrm.it

³ Dipartimento di Ingegneria dell’Informazione, Università di Siena (Italy)
marco@mcculloch.ing.unifi.it

Abstract

In this paper the authors describe a novel terminal attractor algorithm for solving linear systems, named ELISA. The method here presented is based on a special neural network continuous form of gradient descent, approaching the minimum of a quadratic function in a constant time, depending solely on the initial value of the residual function. The algorithm is founded on a new concept, called *non-suspiciousness*, which can be seen as a generalisation of convexity. Under general hypotheses it is proven that ELISA has $\mathcal{O}(n^2)$ as computational complexity and therefore is theoretically optimal. The preliminary numerical experiences clearly assess ELISA’s efficiency both by dominant operation counting and in terms of CPU-time.

1 Introduction

Function optimisation seems to be an ubiquitous formulation of an impressive number of different problems. Optimal learning in multilayer perceptrons or in recurrent networks may directly be framed in an optimisation scenery. In this paper we introduce the concept of *non suspiciousness* to address typical situations arising from continuous function optimisation.

The above concept is strictly related to the absence of local minima, i.e. to the classic case in which there is no suspect that properly designed numerical algorithms get stuck and fail reaching the optimal solution. From this point of view non-suspiciousness can be seen as a generalisation of convexity.

We prove that there are some intriguing links between suspiciousness and complexity, allowing us to identify a class of *non-suspect problems* for which we can provide a lower bound on its complexity. Such a lower bound leads to establish the computational burden of an *optimisation linear solver* which is extremely competitive with the classical iterative methods. As a particular case, we prove here that, under suitable hypotheses, the solution of linear systems can be obtained by solving a non-suspect minimisation problem with an $\mathcal{O}(n^2)$ algorithm.

The paper is organised as follows. In the next section we introduce a special continuous form of gradient descent for function optimisation, that can be properly discretised. In the same section we give conditions under which the discretisation error can be kept arbitrarily close to zero. Moreover, we introduce the concept of non-suspiciousness, establishing an upper bound for the number of steps required by the discrete iteration scheme. In Section 3 we deal with the resolution of linear systems in an optimisation frame, by describing a new algorithm ELISA (*Efficient Linear Systems Algorithm*) and discussing its computational requirements. In Section 4 are reported some numerical examples in order to compare ELISA’s performances with classical methods. We show, in particular, that ELISA is competitive with Gauss algorithm for systems involving hundreds of equations and outperforms it for problems exceeding one thousand variables. Finally, in Section 5, conclusions and work in progress are briefly discussed.

2 Non-suspect problems of optimisation

The theoretical background of this paper relies on a general property of gradient descent algorithms for function optimisation that will be described in this section.

We consider a parametrical system

$$Y = S(U, W), \tag{1}$$

where the input $U \in \mathbb{R}^l$ is mapped to output $Y \in \mathbb{R}$ by means of the weight vector $W \in \mathbb{R}^n$. Neural networks like multilayer perceptrons and radial basis functions are special interesting cases of such parametrical systems.

Let $\Omega \subset \mathbb{R}^n$ be and $E : \Omega \rightarrow \mathbb{R}^+$ be the error function associated to the parametrical system (1). Assuming $E \in \mathcal{C}^1$, the minimisation of E can be carried out by following the differential scheme

$$\frac{dW}{dt} = -\gamma \nabla_W E = f(t, W),$$

where $W \in \Omega$ is the weight vector.

Let us choose

$$\gamma \doteq \frac{\Psi(E)}{\|\nabla_W E\|^2},$$

being Ψ a non-negative continuous function of E . Based on this choice of γ , the dynamics of the error function becomes

$$\frac{dE}{dt} = (\nabla_W E)^T \frac{dW}{dt} = -\Psi(E), \quad (2)$$

which forces the cost function E to continuously decrease to zero and approaching the final solution, which is a *terminal attractor* [Wang and Hsu, 1991] in finite time. Nevertheless, when γ is undefined, the substitution of $\frac{dW}{dt}$ is not allowed in (2). Such a situation occurs whenever a local minimum is reached. Local minima are attractors for the differential equation (2), which describes the error evolution only in the basin of attraction of each minimum.

This way of forcing the system dynamics has been suggested in the field of Automatic Control for Sliding Modes [Slotine and Sastry, 1983] and, more recently, in Neural Networks for learning the weights of multilayer perceptrons (see e.g.: [White, 1990, Wang and Hsu, 1991, Bianchini *et al.*, 1994, Parlos *et al.*, 1994]).

The latter contributions, exploiting the idea of forcing a gradient descent dynamics for numerical issues, are starting points for a theoretical investigation on the relationship between numerical efficiency and computational complexity.

2.1 Linear descent

Let $\Psi(E) = \eta$ be. If $E = 0$ is the optimal solution, $\eta = E_0/\sigma$ allows approaching the terminal attractor solution for $t_e = \sigma$ independently of the function at hand, being $E(t) = E_0 - \frac{E_0}{\sigma}t$, $E(0) = E_0$. The corresponding differential equation becomes

$$\frac{dW}{dt} = -\eta \frac{\nabla_W E}{\|\nabla_W E\|^2}. \quad (3)$$

We underline that, by the singularity of the terminal attractor, one must be aware of numerical instability in the neighbourhood of the optimal solution. Therefore, the gradient descent must be carefully investigated whenever the magnitude of the gradient goes below a given threshold. Basically, one has to stop the dynamics of differential equation (3) before instability takes place, but this can be done in such a way to get arbitrarily close to the optimal solution. As pointed out in [Bianchini *et al.*, 1994], however, the system trajectory (3) is attracted to any local minima and, consequently, the nice gradient descent described by equation (2) takes place in the basin of attraction of any local minimum. Hence, the reduction of the dynamics incorporated in equation (2) holds, provided that the trajectory will never be trapped into singular points, in which $\|\nabla_W E\| = 0$.

The absence of local minima is an obvious case in which the reduction of the dynamics described by equation (2) takes place with no trouble. In the general case of multimodal functions, the error evolution according to equation (2) can not be guaranteed.

2.2 From continuous to discrete models

In this section we address the problem of providing an arbitrarily close approximation of the continuous dynamics of equation (3). Although the analysis we present is of numerical nature, it is not intended to address numerical issues, but mainly *to discover bounds on the number of steps required to reach the solution*. The basic idea comes from the analysis of the previous section, where it has been pointed out that gradient descent equation (3) for local minima free functions *converges in a time independent of the function at hand*, thus suggesting that the optimisation process has the same computational requirements for any function of the latter type.

Let us now consider the following discrete version of equation (3) which represents the Euler's approximation to the weight dynamics

$$W_{k+1} = W_k - \tau \frac{\eta \nabla E_k}{\|\nabla E_k\|^2}, \quad (4)$$

where the iteration index k is related to the continuous time t and to the quantization step τ by $t = \tau k$.

Definition 1 *The Euler's approximation is consistent to the continuous equation (3) provided that*

- $\forall \varepsilon_a > 0, \exists \tau > 0 : \forall t = k\tau, |E(t) - E_k| < \varepsilon_a$.

Observe that, while in the typical convergence and consistence definition on differential equations (see e.g.: [Golub and Ortega, 1992, Lambert, 1991]) the discretisation is performed for function $W(t)$, we require that the approximation holds for the error function $E(W)$.

Lemma 1 *Suppose E is local minima free. Then $\varepsilon_s > 0$ exists such that, using the discrete updating equation (4), the inequality $\|\nabla_t E(W)\| \leq \varepsilon_s$ holds only when W approaches global minima.*

Proof

The proof follows straightforwardly by contradiction.

This lemma guarantees that the updating equation (4) meets the chosen stopping criterion only when approaching a global minimum.

The following discretisation analysis is based on some assumptions that will be referred to as *non-suspiciousness conditions* (see [Frasconi *et al.*, 1997]).

Definition 2 *The non-suspiciousness conditions hold if:*

1. $\forall \varepsilon_a \in R^+, \|\nabla E_k\| > \varepsilon_s$ during the gradient descent, apart from $k : |E_k - E_{min}| < \varepsilon_a$;

2. given η , $\tau \leq \frac{\varepsilon_a}{\eta}$;

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

The following theorem gives an indication on the choice of the quantisation step τ that guarantees the desired approximation ε_a of continuous equation (3) and, consequently, the number of steps required to reach the optimal solutions.

Theorem 1 *Let the non-suspiciousness conditions hold for the difference equation (4). Then, $\forall \varepsilon_a \in \mathbb{R}^+$, the Euler's approximation is consistent to the continuous evolution of the error dynamics in the domain $\mathcal{D}_{\varepsilon_a} \doteq \{W \in I \mathbb{R}^n : |E(W) - E_{min}| > \varepsilon_a\}$, choosing quantization steps no higher than*

$$\tau^* = 2 \left\lfloor \frac{\varepsilon_s^2 \varepsilon_a}{H \eta R_E} \right\rfloor, \quad (5)$$

where $R_E \geq E_0 - E_{min}$.

Moreover, $|E_{k^*} - E_{min}| \leq \varepsilon_a$ holds after at least

$$k^* = \frac{1}{2} \left\lceil \frac{H R_E^2}{\varepsilon_s^2 \varepsilon_a} \right\rceil \quad (6)$$

steps of the discrete iteration scheme.

Proof: (see [Frasconi et al., 1997]).

Remark 1 The theorem gives a suggestion on the choice of the quantisation step τ that guarantees the ε_a -approximation to the continuous equation. τ is independent of the special function involved and related to few parameters only. Thus, a class of functions E exists whose optimisation takes a number of steps that can be bounded by the same value.

3 Solving linear systems by an $\mathcal{O}(n^2)$ algorithm

The problem of solving a linear system

$$Ax = b, \quad A \in \mathbb{R}^{n,n}, \quad x, b \in \mathbb{R}^n \quad (7)$$

can be easily reformulated as an optimisation problem. As a matter of fact, if the linear system admits a solution, it can be discovered by minimising the **error** — residual — function

$$E(x) \doteq \frac{1}{2} \frac{\|Ax - b\|^2}{\|b\|^2}, \quad \|b\| \neq 0. \quad (8)$$

As previously stated, the absence of local minima is an obvious case in which the reduction of the dynamics described by (2) takes place with no trouble and, for equation (8), this is just the case, being the resulting residual error a convex function. Moreover, note that a solution for the optimisation problem can be computed even in

the case in which A is a singular matrix, whereas most numerical methods fail in this circumstance.

Therefore, the fundamental result that must be proved is that the non-suspiciousness conditions hold for the optimisation problem

$$\min_{x \in \Omega} E(x), \quad (9)$$

with $E(x)$ defined in (8).

Note that *establishing the non-suspect nature of the linear system resolution would lead to an iterative linear solver with $\mathcal{O}(n^2)$ as computational complexity*, since this is indeed the computational burden due to the gradient evaluation. Obviously, this is also the lower bound for the problem (7) and, therefore, the algorithm based on equation (3) would be a candidate optimal algorithm.

Lemma 2 *Let E be defined as in (8). Then, $\forall \varepsilon_a \in \mathbb{R}^+$, ε_s exists, $\varepsilon_s \doteq \frac{\sqrt{2\varepsilon_a}}{\|A^{-1}\| \|b\|}$, such that using the discrete updating equation (4), the inequality $\|\nabla E_k\| > \varepsilon_s$ holds $\forall x_k \in \mathcal{D}_{\varepsilon_a}$. Therefore the non-suspiciousness condition 1. holds for (9).*

Proof

Let us choose $\varepsilon_a > 0$ and suppose $|E_k - E_{min}| \geq \varepsilon_a$. As $E_{min} = 0$ for the problem (8)-(9), then

$$|E_k - E_{min}| = E_k = \frac{1}{2} \frac{\|Ax_k - b\|^2}{\|b\|^2} \geq \varepsilon_a. \quad (10)$$

Starting from (10), an upper bound for the norm of the gradient may be evaluated as follows:

$$\begin{aligned} \|\nabla E_k\| &= \frac{\|A^T(Ax_k - b)\|}{\|b\|^2} = \frac{\|A^{-T}\| \|A^T(Ax_k - b)\|}{\|A^{-1}\| \|b\|^2} \\ &\geq \frac{\|Ax_k - b\|}{\|A^{-1}\| \|b\|^2} \geq \frac{\sqrt{2\varepsilon_a} \|b\|}{\|A^{-1}\| \|b\|^2} = \frac{\sqrt{2\varepsilon_a}}{\|A^{-1}\| \|b\|}, \end{aligned}$$

which represents an admissible value for ε_s in order to verify the first non-suspiciousness condition in Definition 2.

Remark 2 The non-suspiciousness condition 2. is obviously verified.

Lemma 3 *Let E be defined as in (8). Then, if the matrix A has a limited spectral norm and the right hand side b is different from the null vector, E has a limited Hessian \mathcal{H} . Therefore the non-suspiciousness condition 3. holds for (9).*

Proof

For the optimisation problem (8)-(9)

$$\mathcal{H} = \frac{A^T A}{\|b\|^2}$$

holds, and thus the above statement directly follows.

From Lemmas 2 and 3 and Remark 2 we can immediately derive, by Definition 2, that problem (9) is non suspect. We are now interested in showing that the above estimations for ε_s and H guarantee the existence of an iterative algorithm such that in a finite, dimension independent, number of steps k^* , the solution of the linear system (7) can be ε_a -approximated. This result is assured by Theorem 1. However, in order to build up an algorithm which is not only theoretically optimal, but also numerically robust, the problem of reaching sufficient precision, simultaneously avoiding numerical instability, must be carefully addressed. The nature of the terminal attractor algorithm suggests the choice of a suitable “smooth” stopping rule (i.e. in the range of values 0.05–0.1 of the euclidean norm of the relative error).

4 Numerical examples

The test has the aim of establishing the range of ELISA’s applicability, in terms of matrix dimension, by showing its performances w.r.t. Gauss Direct Algorithm. The choice of the latter method as a comparison is justified by the fact that other well known iterative algorithms (f.i. Jacobi or Gauss–Seidel) are competitive only if the matrix has special structures and do not converge for matrices of general type.

Table 1: Symmetric matrices with $k(A) = 6$. Relative error norm tolerance 7%

<i>System dimension</i>	<i>Complexity ratio E/G</i>	<i>ELISA run-time</i>	<i>Gauss run-time</i>
500	3.43	1’ 30’’	1’
1000	1.75	6’	8’ 50’’
1500	1.16	13’ 26’’	31’ 30’’
1800	0.96	15’ 48’’	56’ 30’’

Numerical results clearly assess ELISA’s quadratic computational cost, which directly follows from dominant operation counting, while bearing in evidence, via run-time measurements, its appealing “global” performance behaviour vs. Gauss classical method. We emphasise that, although the precision obtained by Gauss algorithm is outstanding and certainly the strongest in general, there is no way of relaxing the accuracy requirements and, consequently, reducing its computational cost *by the very nature of direct methods*. Therefore, *ELISA is strongly recommended if one desires a fast efficient resolution of linear systems of large size*.

5 Conclusions

The canonical discretised form of gradient descent illustrated in the present paper has the dynamics of a terminal attractor and, in particular, if the problem is non suspect, allows determining, in a simple mathematical form, the

time required for approaching the optimal solution. It is important to underline that the non-suspiciousness conditions are a generalisation of the concept of convexity and therefore it is possible to apply this theory to the resolution of non linear problems too. If the parametrical system (1) represents a neural network (f.i. a multilayer perceptron) non-suspiciousness implies the optimality of the back propagation algorithm based on the updating rule derived by equation (4). On the other hand, in a context of Numerical Analysis, the method here presented can be seen as a new approach to the resolution of “special” non linear systems of equations. For this reason, future research will be devoted to the investigation of possible generalisations of the classical Newton–Raphson method.

References

- [Bianchini *et al.*, 1994] Bianchini, M., Gori, M., and Maggini, M. (1994). Does terminal attractor backpropagation guarantee global optimization? In *International Conference on Artificial Neural Networks*, volume 1, pages 377–380, Sorrento, Italy. Springer-Verlag.
- [Frasconi *et al.*, 1997] Frasconi, P., Fanelli, S., Gori, M., and Protasi, M. (1997). Suspiciousness of loading problems. In *IEEE International Conference on Neural Networks*, Houston, 1997. To appear.
- [Golub and Ortega, 1992] Golub, G. and Ortega, J. (1992). *Scientific Computing and Differential Equations*. Academic Press, Inc.
- [Lambert, 1991] Lambert, J. (1991). *Numerical Methods for Ordinary Differential Systems*. John Wiley & Sons.
- [Parlos *et al.*, 1994] Parlos, A., Fernandez, B., Atyla, A., Muthusami, J., and Tsai, W. (1994). An accelerated algorithm for multilayer perceptron networks. *IEEE Transactions on Neural Networks*, 5(3):493–497.
- [Slotine and Sastry, 1983] Slotine, J. and Sastry, S. S. (1983). Tracking control of non-linear systems using sliding surfaces, with applications to robot manipulators. *International Journal of Control*, 38(2):465–492.
- [Wang and Hsu, 1991] Wang, S. and Hsu, C. H. (1991). Terminal attractor learning algorithms for backpropagation neural networks. In *International Joint Conference on Neural Networks*, pages 183–189, Singapore. IEEE Press.
- [White, 1990] White, H. (1990). The learning rate in backpropagation systems: an application of Newton’s method. In *International Joint Conference on Neural Networks*, volume 1, pages 679–684, San Diego, (CA).