

# Non-Suspiciousness: A Generalisation of Convexity in the Frame of Foundations of Numerical Analysis and Learning

M. Bianchini, S. Fanelli, M. Gori, *IEEE Member*, and M. Protasi

*Abstract*— The effectiveness of connectionist models in emulating intelligent behaviour is strictly related to the capability of the learning algorithms to find optimal or near-optimal solutions. In this paper, a canonical reduction of gradient descent dynamics is proposed, allowing the formulation of the neural network learning as a *finite* continuous optimisation problem, under some *non-suspiciousness conditions*. In the linear case, the *non-suspect* nature of the problem guarantees the implementation of an iterative method with  $\mathcal{O}(n^2)$  as computational complexity. Finally, since non-suspiciousness is a generalisation of the concept of convexity, it is possible to apply this theory to the resolution of non linear problems.

*Keywords*— Neural Networks, Computational Complexity, Non-Suspiciousness, Terminal Attractors, Advances in Numerical Analysis.

## I. INTRODUCTION

In the last few years impressive efforts have been made in using connectionist models either for modelling human behaviour or for solving practical problems. Unlike other symbolic approaches to machine learning, that are based on “*intelligent search*” [1], in connectionist models the learning is typically framed as an optimisation problem. After the seminal PDP’s books, Minsky published an extended edition of Perceptron [2] that contains an intriguing epilogue on PDP’s novel issues: “... (*Backpropagation*) is nothing more than a straightforward hill-climbing algorithm. We have the impression that many people in the connectionist community do not understand that this is merely a particular way to compute a gradient and have assumed instead that *Backpropagation* is a new learning scheme that somehow gets around the basic limitation of hill-climbing” (see [2], p. 286). Minsky’s issues call for the need to give optimal learning a theoretical foundation, as simple gradient descent algorithms have no guarantee to learn the assigned task. From a mathematical point of view, the convexity of the error function associated to the neural network is in general the only hypothesis guaranteeing the convergence to the optimal solution of gradient descent algorithms. Weak re-

M. Bianchini is with Dipartimento di Sistemi e Informatica, Università degli Studi di Firenze (Italy). E-mail: monica@mcculloch.ing.unifi.it

M. Gori is with Dipartimento di Ingegneria dell’Informazione, Università degli Studi di Siena (Italy). E-mail: marco@mcculloch.ing.unifi.it

S. Fanelli and M. Protasi are with Dipartimento di Matematica, Università di Roma “Tor Vergata” (Italy). E-mail {fanelli,protasi}@mat.utovrm.it

laxations of convexity assumptions (like f.i. quasi- or pseudo-convexity (see [3])) do not substantially enlarge the set of learning problems for which the optimal solution can be computed by an algorithm of Backpropagation type. On the other hand, surprisingly enough, 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, therefore, evident that there is a strong need to determine theoretical conditions ensuring the optimal convergence of “*at least*” some special form of gradient descent algorithms. This work is a contribution towards this aim.

## II. SMALE’S THEORY AND FOUNDATIONS OF NUMERICAL ANALYSIS

Turing machines have played a central role in the foundation of Computer Science and, particularly, in the generalisation of the concept of algorithm, allowing, without ambiguity, the investigation of complexity lower bounds for every type of algorithm.

The introduction of continuous machines, considered as an idealisation of Turing’s model, yields the “*translation in numerical terms*” of the main concepts of Complexity Theory, thereby extending the classical finite approach of computation, according to the practical requirements of Numerical Analysis [4], [5]. In this way, it is possible to perform the error analysis independently of the particular solving algorithm. More precisely, it can be proved (see [6]) that an algorithm defined by a continuous machine is able to solve an approximate problem  $P(\varepsilon)$  in polynomial time if and only if:

$$\delta(\varepsilon, y) \leq c_1(\dim(y) + \log(\varepsilon) + \log(\aleph(y)))^{q_1}, \quad (1)$$

$$T(\varepsilon, y) \leq c_2(\dim(y) + \log(\varepsilon) + \log(\aleph(y)))^{q_2}, \quad (2)$$

where:

- $\dim(y)$  is the number of bits requested to store the input  $y$ ;
- $\varepsilon$  indicates the desired precision to compute the solution associated to  $y$ ;
- $\aleph(y)$  is a function describing the “*level of difficulty*” of the problem;
- $\delta(\cdot, \cdot)$  represents the maximum acceptable propagation error (the stability);
- $T(\cdot, \cdot)$  indicates the number of steps used in the computation of the solution associated to the input  $y$  (the complexity);

- $c_1, c_2, q_1, q_2$  are constants depending solely on the machine.

Although the formulas (1) and (2) are of theoretical type, they state that the complexity and the stability of a given algorithm are deeply related and, in general, bound its efficiency and accuracy “with the same intensity”.

More exactly in [7] it was proved that, if the solving algorithm of a problem  $P$  is constrained to satisfy some stability conditions, then the complexity lower bounds of  $P$  can be significantly modified and more easily evaluated. Roughly speaking, it can eventually happen that the “best algorithm” in terms of computational complexity for  $P$  cannot be actually “the optimal algorithm”.

In the field of linear problems and, particularly, when dealing with the solution of linear systems (or matrix inversion), classical Numerical Analysis is not effective to quantify exactly the difficulty of the given problem in terms of matrix conditioning.

It is often said that, if the matrix has a “low condition number”, then the system can be solved both with efficiency (f.i. with complexity  $\mathcal{O}(n^3)$  or  $\mathcal{O}(n^{\log_2(7)})$ ) and with accuracy (with a satisfactory propagation error). However, a precise relationship between the function  $T$  in (2) and the condition number of the matrix cannot be in general achieved with classical tools. The well known results by Wilkinson [8], [9] gave a fundamental contribution about the connection between stability and conditioning.

As an example, we mention the following inequality, satisfied by positive definite matrices of order  $n$ :

$$\delta_r(A^{-1}) \leq 14.24 \left( n^{5/2} 2^{-s} \right) k(A), \quad (3)$$

being

- $s$  the number of digits employed by the computer;
- $k(A)$  the condition number of  $A$ ;
- $\delta_r(A^{-1})$  the relative propagation error in the computation of  $A^{-1}$ .

The formula (3) shows that in any algorithm implemented for the inversion of a positive definite matrix, the theoretical function  $\aleph$  indicated in the general formula (1) can be substantially identified with  $k(A)$ .

However, the determination of a practical relationship, in terms of an operational inequality, between the computational complexity of a general class of algorithms for linear problems and conditioning is still an open question.

Remarkable contributions towards the latter aim, at least in a probabilistic form, were given in [10], [11].

It is important to cite, in particular, an interesting result proved in [10]. Let  $IP$  be the set of all non-invertible real matrices of order  $n$ . Then, given a matrix  $A$ , the difficulty of its inversion can be expressed in

terms of “the minimum distance  $\text{dist}(A, IP)$  between  $A$  and  $IP$ ” with the formula:

$$\text{dist}(A, IP) = \frac{\|A\|_F}{k(A)}, \quad (4)$$

where  $\|\cdot\|_F$  denotes the Frobenius norm. By (4), if  $A/\|A\|_F$  is uniformly distributed in the unitary hypersphere, it follows that:

$$\frac{c \left(1 - \frac{1}{x}\right)^{n^2-1}}{x} \leq P[k(A) \geq x] \leq 2 \left( \left(1 + \frac{2n}{x}\right)^{n^2} - 1 \right), \quad (5)$$

being  $c > 0$  a constant proportional to the volume  $(n^2 - 1)$ -dimensional of the  $IP$ -manifold contained in the unitary hypersphere.

The formula (5) determines the connection between the size  $n$  of a matrix and its condition number. Even if (5) is impractical in many cases, it points out an important probabilistic inequality involving the conditioning of the problem and its dimension and hence, indirectly, the computational complexity of a given algorithm.

We shall see in Section IV how the fulfillment of certain “non-suspiciousness conditions” allow to establish in the linear case a precise deterministic relationship between  $k(A)$  and the computational complexity of a canonical form of gradient descent, which is theoretically optimal.

### III. EXTENDING THE BORDER BETWEEN FINITE AND INFINITE PROBLEMS

The dichotomy between Computer Science and classical Numerical Analysis has been for many decades the main obstacle to the development of “eclectic computational tools”. With the latter term, we indicate the capability of implementing algorithms properly adapted to particular computational requirements. Typical examples can be found in the field of neural networks, where the simple application of classical gradient descent techniques (even in the most efficient form) fail discovering the optimal solution in many cases. Although it is well known that gradient methods guarantee in general the convergence to a stationary point only, one might expect that for particular problems (such as f.i. the minimisation of the error function) the determination of the optimal solution be within reach under suitable operational conditions.

The antithesis between Discrete and Continuous Mathematics is even stronger from a computational point of view.

Classical algorithms for problems on discrete sets (graphs, integer variables and so forth) are characterised by a procedure formalized in a finite number of steps, while the Numerical Calculus techniques are

based in the majority of cases on the convergence of a sequence to the optimal solution. Proper stopping rules on the truncation error reduce the latter computational scheme to a finite process, but, unfortunately, a precise forecast on the number of iterations requested to achieve the desired approximation cannot be in general obtained in advance. On the other hand, the real border between finite and infinite computational procedures is not theoretically established by the type of the variables associated to the problems.

Linear Programming, Convex Quadratic Programming or the simple minimisation of a symmetric positive semidefinite bilinear form are continuous problems that can be exactly solved with a finite number of steps. This proves that the distinction between algorithms and infinite numerical procedures is not always characterised respectively by the discrete or the continuous range of the variables of the problem.

More precisely, it is not clear if quadratic convex problems represent the real boundary separating the finite continuous optimisation problems from the infinite ones.

Most of Numerical Analysis is based upon the application of Fixed Point Theorem [12], which assures the convergence of the iterative procedures by means of a contraction of the distance between successive terms of the sequence approximating the optimal solution. Typically, the minimisation of the error function  $E(W)$  of a neural network can be carried out by following the differential scheme:

$$\begin{aligned} \frac{dW}{dt} &= -\gamma \nabla_W E = f(t, W), \\ W(0) &= W_0, \end{aligned} \quad (6)$$

whose discretisation, Euler's approximation of (6), is:

$$\begin{aligned} W_{k+1} &= W_k - \gamma \tau(k) \nabla E_k = \phi(W_k), \\ W(0) &= W_0. \end{aligned} \quad (7)$$

The formula (7) represent gradient descent's standard iterative method. The convergence of the sequence  $W_k$  to a stationary point for  $E(W)$  is guaranteed by classical Lipschitz's condition for the problem (6), assuring the existence of a fixed point for the operator  $\phi(\cdot)$  in (7).

In [13] it was pointed out by Zak that there exist singular solutions of the differential equation (6), named "terminal attractors", violating Lipschitz's condition. The main characteristic of these singular solutions lies upon the fact that they can be reached in finite time. So, at least from a theoretical point of view, the computation of the optimal solution could be performed with a finite number of steps. If we choose in (6):

$$\gamma \doteq \frac{\psi(E)}{\|\nabla_W E\|^2}, \quad (8)$$

being  $\psi$  a non-negative continuous function of  $E$ , the dynamics of the function  $E(W)$  in (6) becomes:

$$\frac{dE}{dt} = (\nabla_W E)^T \frac{dW}{dt} = -\psi(E). \quad (9)$$

In this way, the function  $E$  is forced to continuously decrease to zero and approaching the optimal solution, which is a terminal attractor.

Unfortunately, there are two main difficulties in the implementation of algorithms of this type:

1. By the singularity of the terminal attractor, one must be aware of numerical instability in the neighbourhood of the optimal solution.
2. The system trajectory is in general attracted to any local minimum (see [14]), in contrast to the results claimed in [15].

It is therefore natural to define a special class of functions for which the above difficulties can be overcome.

#### IV. NON-SUSPICIOUSNESS: DEFINITIONS AND MAIN RESULTS

Let us now consider equation (7), where the iteration index  $k$  is related to the continuous time  $t$  and to the quantisation step  $\tau$  by  $t = \tau k$ .

*Definition 1:* The Euler's approximation (7) is consistent to the continuous equation (6) 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.: [16], [17]) 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 (7), the inequality  $\|\nabla_t E(W)\| \leq \varepsilon_s$  holds only when  $W$  approaches global minima.

*Proof:* (see [18]).

This lemma guarantees that the updating equation (7) meets the chosen stopping criterion only when approaching a global minimum.

The following assumptions will be referred to as "non-suspiciousness conditions" (see [18]).

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

1.  $\forall \varepsilon_a \in \mathbb{R}^+, \|\nabla E_k\| > \varepsilon_s$  during the gradient descent, apart from  $k : |E_k - E_{min}| < \varepsilon_a$ ;
2. given  $\eta, \tau \leq \frac{\varepsilon_a}{\eta}$ ;
3.  $E \in \mathcal{C}^2$  and has a bounded Hessian  $\mathcal{H}$ , (i.e.  $\exists H > 0 : \|\mathcal{H}\| < H$ ).

The next theorem gives an indication on the choice of the quantisation step  $\tau$  that guarantees the desired

approximation  $\varepsilon_a$  of continuous equation (6) and, consequently, the number of steps required to reach the optimal solution.

*Theorem 1:* Let the non-suspiciousness conditions hold for the difference equation (7). Then,  $\forall \varepsilon_a \in \mathbf{R}^+$ , Euler's approximation is consistent to the continuous evolution of the error dynamics in the domain

$$\mathcal{D}_{\varepsilon_a} \doteq \{W \in \mathbf{R}^n : |E(W) - E_{min}| > \varepsilon_a \},$$

choosing quantization steps no higher than

$$\tau^* = 2 \frac{\varepsilon_s^2 \varepsilon_a}{H \eta R_E},$$

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

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

$$k^* = \left\lceil \frac{H R_E^2}{2 \varepsilon_s^2 \varepsilon_a} \right\rceil \quad (10)$$

steps of the discrete iteration scheme.

*Proof:* (see [18]).

*Remark 1:* The quantisation step  $\tau$  is independent of the special function involved and is 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.

Particular interesting is the case of a quadratic optimisation problem or, in other words, the study of a linear gradient descent dynamics [19]. In the latter situation, the problem of solving a linear system

$$AW = b, \quad A \in \mathbf{R}^{n,n}, \quad W, b \in \mathbf{R}^n \quad (11)$$

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

$$E(W) \doteq \frac{1}{2} \frac{\|AW - b\|^2}{\|b\|^2}, \quad \|b\| \neq 0. \quad (12)$$

As previously stated, the absence of local minima is an obvious case in which the reduction of the dynamics described by (9) takes place with no trouble and, for equation (12), 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 if  $A$  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

$$\min_{W \in \Omega} E(W), \quad (13)$$

with  $E(W)$  defined in (12). 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 (11) —  $\mathcal{O}(n^2)$  represents also the cost of the data-acquisition phase — and, therefore, the algorithm based on equation (6) would be a candidate optimal algorithm.

*Lemma 2:* Let  $E$  be defined as in (12). Then,  $\forall \varepsilon_a \in \mathbf{R}^+$ ,  $\varepsilon_s \doteq \frac{1}{k(A)} \frac{\sqrt{2\varepsilon_a} \|A\|}{\|b\|}$  exists, such that using the discrete updating equation (7), the inequality  $\|\nabla E_k\| > \varepsilon_s$  holds  $\forall W_k \in \mathcal{D}_{\varepsilon_a}$ . Therefore the non-suspiciousness condition 1. holds for (13).

*Proof:* (see [19]).

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

*Lemma 3:* Let  $E$  be defined as in (12). Then, if the matrix  $A$  has a bounded spectral norm,  $E$  has a bounded Hessian  $\mathcal{H}$ . Therefore the non-suspiciousness condition 3. holds for (13).

*Proof:* (see [19]).

From Lemmas 2 and 3 and Remark 2 we can immediately derive, by Definition 2, that problem (13) is non-suspect.

Hence, we can establish the following important:

*Theorem 2:* Given a linear system

$$AW = b, \quad A \in \mathbf{R}^{n,n}, \quad W, b \in \mathbf{R}^n$$

and the quadratic function

$$E(W) \doteq \frac{1}{2} \frac{\|AW - b\|^2}{\|b\|^2}, \quad \|b\| \neq 0,$$

then,  $\forall \varepsilon_a \in \mathbf{R}^+$ , the inequality

$$\|\nabla E_k\| > \frac{1}{k(A)} \frac{\sqrt{2\varepsilon_a} \|A\|}{\|b\|}$$

is satisfied  $\forall W_k \in \mathcal{D}_{\varepsilon_a}$ .

*Remark 3:* Theorem 2 states, in particular, that if a sequence of matrices  $A_n$  satisfies the condition:

$$k(A_n) \leq x, \quad \forall n,$$

then there exists a linear gradient descent dynamics with  $\mathcal{O}(n^2)$  as computational complexity.

The iterative linear solver shown in the present paper belongs to the class of parameter-free methods, i.e. the iterates are computed by using information

obtained solely by the iteration steps. Such methods are f.i. the well known Jacobi and Gauss–Seidel procedures, several generalized Conjugate Gradients [20], [21], [22], [23], error minimising and conjugate Krylov subspace [24], [25] and quasi-minimal residual algorithms [26], [27]. Attempting to apply iterative methods to general indefinite non-symmetric matrices, during the last two decades a considerable part of the research has been devoted to the study and implementation of the Generalised Minimum RESidual (GMRES) algorithm [28]. Unfortunately, although GMRES and related schemes generate at each iteration optimal approximate solutions of the linear system, storage requirements per iteration grow linearly and, therefore, are too expensive for large size full matrices. For this reason, more recently, research in non-symmetric matrix iterations has focused mainly on possible implementations with low and roughly constant storage per iteration [29]. *The main advantage of the present iterative linear solver is that it can be applied to any well-conditioned matrix independently of its structure and with the minimum effort in terms of non-arithmetic operations and storage.*

## V. CONCLUSIONS

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. If  $E(W)$  represents the error function of a neural network, non-suspiciousness implies the optimality of the Backpropagation algorithm based on the updating rule derived by equation (7). On the other hand, in the frame of foundations of Numerical Analysis, the theoretical results gained can be seen as a new approach to the numerical solution of particular classes of non linear systems. Consequently, our future aim will be the investigation of possible improvements and generalisations of the classical Newton–Raphson method. More precisely, we will try to discover if there exist particular classes of non convex functions which satisfy the non-suspiciousness conditions, thereby allowing an efficient resolution of the corresponding non linear system of the stationary equations.

## REFERENCES

- [1] R. Michalsky, J. Carbonell, and T. Mitchell, *Machine Learning, an Artificial Intelligence Approach*, vol. 1/2. San Mateo: Morgan Kaufmann, 1983.
- [2] M. Minsky and S. Papert, *Perceptrons — Expanded Edition*. Cambridge: MIT Press, 1988.
- [3] M. Minoux, *Mathematical Programming: Theory and Algorithms*. Wiley and Sons, 1986.
- [4] L. Blum, M. Shub, and S. Smale, “On the theory of computation and complexity over the real numbers: NP-completeness, recursive functions and universal machines,” *Bull. Amer. Math. Soc.*, vol. 21, pp. 1–46, 1989.
- [5] S. Smale, L. Blum, F. Cucker, and M. Shub, “Algebraic settings for the problem P-NP,” in *Lect. in Appl. Maths.* (M. S. . S. S. J. Renegar, ed.), 1996.
- [6] S. Smale, “Some remarks on the foundations of Numerical Analysis,” *SIAM Review*, vol. 32, no. 2, pp. 211–220, 1990.
- [7] W. Miller, “Computational complexity and numerical stability,” *SIAM J. Comp.*, vol. 2, no. 2, pp. 97–107, 1975.
- [8] J. Wilkinson, “Modern error analysis,” *SIAM Review*, vol. 13, pp. 548–568, 1971.
- [9] J. Wilkinson, “Note on matrices with a very ill-conditioned eigenproblem,” *Num. Math.*, vol. 19, pp. 176–178, 1972.
- [10] J. Demmel, “On condition numbers and the distance to the nearest ill-posed problem,” *Num. Math.*, no. 51, pp. 251–289, 1987.
- [11] J. Demmel, “The probability that a Numerical Analysis problem is difficult,” *Math. Comput.*, vol. 50, no. 182, pp. 449–481, 1988.
- [12] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore and London: The Johns Hopkins University Press, 1989.
- [13] M. Zak, “Terminal attractors in neural networks,” *Neural Networks*, vol. 2, pp. 259–274, 1989.
- [14] M. Bianchini, S. Fanelli, M. Gori, and M. Maggini, “Terminal attractor algorithms: A critical analysis,” *Neurocomputing*, vol. 15, pp. 3–13, 1997.
- [15] S. Wang and C. H. Hsu, “Terminal attractor learning algorithms for backpropagation neural networks,” in *International Joint Conference on Neural Networks*, (Singapore), pp. 183–189, IEEE Press, November 1991.
- [16] G. H. Golub and J. M. Ortega, *Scientific Computing and Differential Equations*. Academic Press, Inc., 1992.
- [17] J. Lambert, *Numerical Methods for Ordinary Differential Systems*. John Wiley & Sons, 1991.
- [18] P. Frasconi, S. Fanelli, M. Gori, and M. Protasi, “Suspiciousness of loading problems,” in *IEEE International Conference on Neural Networks*, vol. 2, (Houston), pp. 1240–1245, IEEE Press, 1997.
- [19] M. Bianchini, S. Fanelli, M. Gori, and M. Protasi, “Solving linear systems by a neural network canonical form of efficient gradient descent,” in *Progress in Connectionist-Based Information Systems, ICONIP-ANZIIS-ANNES’97*, vol. 1, (Dunedin), pp. 531–534, Springer Verlag, 1997.
- [20] O. Axelsson and G. Lindskog, “On the rate of convergence of the preconditioned conjugate gradient method,” *Numer. Math.*, vol. 48, pp. 499–523, 1986.
- [21] R. W. Freund, “On conjugate gradient type methods and polynomial preconditioners for a class of non-Hermitian matrices,” *Numer. Math.*, vol. 57, pp. 285–312, 1990.
- [22] S. F. Ashby, T. A. Manteuffel, and P. E. Saylo, “A taxonomy for conjugate gradient methods,” *SIAM J. Numer. Anal.*, vol. 27, pp. 1542–1568, 1990.
- [23] M. H. Gutknecht, “A completed theory of the unsymmetric Lanczos process and related algorithms, part 1,” *SIAM J. Matrix Anal.*, vol. 13, pp. 594–639, 1992.
- [24] R. Weiss, “Error-minimizing Krylov subspace methods,” *SIAM J. Sci. Comput.*, vol. 15, pp. 511–527, 1994.
- [25] R. Weiss, “Minimization properties and short recurrences for Krylov subspace methods,” *Electron. Trans. Numer. Anal.*, vol. 2, pp. 57–75, 1994.
- [26] R. W. Freund and N. M. Nachtigal, “QMR: a Quasi-Minimal Residual method for non-Hermitian linear systems,” *Numer. Math.*, vol. 60, pp. 315–339, 1991.
- [27] R. W. Freund, “A transpose-free quasi-minimal residual algorithm for non-Hermitian linear systems,” *SIAM J. Sci. Comput.*, vol. 14, pp. 470–482, 1993.
- [28] Y. Saad and M. H. Schultz *SIAM J. Sci. Statist. Comput.*, vol. 7, no. 856, 1988.
- [29] R. W. Freund, G. H. Golub, and N. M. Nachtigal, “Iterative solutions of linear systems,” *Acta Numerica*, pp. 57–100, 1992.