Department: Head
Editor: Name, xxxx@email

The origin and development of Krylov subspace methods

Yousef Saad
University of Minnesota, Department of Computer Science and Engineering

***Abstract*—Krylov subspace methods have had an unparalleled success in solving real-life problems across disciplines ranging from computational fluid dynamics to statistics, machine learning, control theory, computational chemistry, among many others. This article provides a brief history of these methods, discussing their origin, their expansion, and the lives of the people behind them.**

INTRODUCTION

It is not an easy task to pin-point exactly when or by whom the term *'Krylov subspace methods'* (KSMs) was first used. The term emerged toward the late 1970s and early 1980s [33], [36], [34], and then it was popularized possibly by Parlett's book on eigenvalue problems [34]. What matters is that modern Krylov subspace methods appeared in the early 1950s, then they were all but abandoned for a while and subsequently reappeared in force in the 1970s, when they started showing a great success in solving various problems in scientific computing.

A Krylov subspace method can be defined as a process that extracts an approximate solution to a givem problem from a Krylov subspace, which is a subspace of the form

$$K_m = \mathbf{span}\{v, Av, \cdots, A^{m-1}v\} \qquad (1)$$

where $A \in \mathbb{R}^{n\times n}$ and $v \in \mathbb{R}^n$. What is remarkable about subspaces of this type is how frequently they are invoked to help solve various problems in science and engineering. Roughly speaking, this is to be attributed to the important property that K_m can be viewed as the subspace $\mathbb{R}^n$ that best captures the actions of A.

The root: Krylov's article, 1931

A subspace of the form (1) was introduced for the first time in 1931 in an article [24] by Aleksei Nikolaevich Krylov, the Russian mathematician whose name gave the eponym to this class of methods. This article can be viewed as the root of Krylov Subspace Methods. In it, the author described a new procedure for computing the characteristic polynomoial of an arbitrary square matrix. Given an $n \times n$ matrix A, whose characteristic polynomial is

$$p_n(t) = t^n - \mu_{n-1}t^{n-1} - \cdots - \mu_1 t - \mu_0,$$

and given an arbitrary nonzero vector v_1, then, assuming that v_1 is of grade [1] n, the n vec-

[1]In linear algebra, the minimal polynomoal of a nonzero vector $v \in \mathbb{R}^n$ with respect to an $n \times n$ matrix A is the monic polynomial p of smallest degree such that $p(A)v = 0$. The degree of p is called the grade of v, see e.g., [49] for details.

 Published by the IEEE Computer Society

tors $v_1, Av_1, \cdots, A^{n-1}v_1$ are lineary independent while the $n+1$ vectors $v_1, Av_1, \cdots, A^n v_1$ must be dependent and in fact they are known to satisfy the relation:

$$A^n v_1 - \mu_{n-1}A^{n-1}v_1 - \cdots - \mu_1 Av_1 - \mu_0 v_1 = 0.$$

This is clearly equivalent to the assumption that v_1 is of grade n. Note also that if the entries of v_1 are randomly selected then it is highly probable that the grade of v_1 is n. Defining the 'Krylov sequence':

$$v_{j+1} = Av_j, \ j = 1, \cdots, n \tag{2}$$

then this property shows that

$$\mu_0 v_1 + \mu_1 v_2 + \cdots \mu_{n-1} v_n = v_{n+1}. \tag{3}$$

Therefore, the scalars μ_i, for $i = 0, 1, \cdots, n-1$ can be obtained by solving a linear system with the coefficient matrix $[v_1, v_2, \cdots, v_n]$ and right-hand side v_{n+1}. Note that the column-vectors $v_i, i = 1, \cdots, n$ of the linear system (3) are known to be almost linearly dependent even for relatively small matrices and this means that the system (3) if generally *ill-conditioned*, even for a small n, see [49, Sec. 6-22].

An alternative viewpoint to the algorithm will unravel a key linear algebraic relation. Indeed, underlying Krylov's method is a technique for reducing A into a special form by a similarity transformation. Invoking the basis [2] $V = [v_1, \cdots, v_n]$, we immediatly obtain:

$$A[v_1|v_2|\cdots|v_n] = [v_1|v_2|\cdots|v_n]H \tag{4}$$

where:

$$H = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 & \mu_0 \\ 1 & 0 & 0 & \dots & 0 & \mu_1 \\ & 1 & 0 & \dots & 0 & \mu_2 \\ & & \cdot & \dots & \cdot & \vdots \\ & & & 1 & 0 & \mu_{n-2} \\ & & & & 1 & \mu_{n-1} \end{bmatrix}. \tag{5}$$

The above matrix is known as a *companion matrix* see, e.g., [49, Sec. 1-10]. Equation (5) is of the form $AV = VH$ and, recalling that V is invertible, it means that we have reduced A into the companion form by a similarity transformation since $V^{-1}AV = H$.

The method described above is rather simple but it was a breaktrhough when it made its appeareance in 1931, at a time where computing eigenvalues and eigenvectors was challenging, even for small matrices. Before it, the only practical method available for computing characteristic polynomials was one developed by Leverrier in 1840, see, Householder [20] for details and references. As is often the case in this situations, it is not the original discovery that matters but the alternatives generated after it.

About Krylov

Aleksei Nikolaevich Krylov was born on August 15, 1863 in Visyaga, Simbirskoy (renamed Ulyanovskaya) in Russia and died on October 26, 1945 in Leningrad (Now St Petersbourg). The son of an artillery officer, he joined the Maritime Acamedy as a student and then a teacher and remained at the academy for 50 years. His work was rather broad: shipbuilding, magnetism, artillery, mathematics, astronomy, geodesy. He became increasingly influential as a leader in science policy in Russia and later in the USSR, and in particular made efforts to maintain contacts with western researchers.

Hessenberg's contribution, 1942

The near linear dependence of the v_i's in Krylov's method will cause numerical difficulties in most practical circumstances. More than a decade after the publication of Krylov's seminal article, an attempt was made to remedy this particilar issue. In 1942, K. Hessenberg in a doctoral dissertation [17] developed a method, in the same

[2] Following common usage, $V = [v_1, \cdots, v_n]$ denotes a matrix whose columns are the 'column-vectors' $v_1, \cdots, v_n$. The term 'basis' or 'system' V refers to the *basis* or *generating system* consisting of the columns $v_1, v_2, \cdots, v_n$ of V.

spirit as Krylov's, to reduce A into what we now call the 'Hessenberg form' or 'quasi-triangular form', i.e., the form of a matrix H that satisfies $h_{ij} = 0$ for $i > j + 1$ [3]. In other words, a relation like (4) will be satisfied but the matrix H is no longer in companion form but in Hessenberg form. Hessenberg's basic idea was to redefine each vector v_j in (2) by subtracting from it a linear combination of the previous v_i's, i.e.,

$$h_{j+1,j}v_{j+1} = Av_j - \sum_{i=1}^{j} h_{ij}v_i. \qquad (6)$$

where $h_{j+1,j}$ is a scaling factor to normalize v_{j+1}. Here, the scalars h_{ij} for $i = 1, \cdots, j + 1$ are selected in such a way that

$$v_{j+1} \perp g_1, g_2, \cdots, g_j, \text{ and } g_{j+1}^T v_{j+1} = 1, \quad (7)$$

where the g_i's are a set of preselected vectors. It follows easily from (6) that the relation (4) holds where $V = [v_1, v_2, \cdots, v_n]$ and H is an upper Hessenberg matrix of size $n \times n$ whose nonzero entries are the h_{ij}'s that appear in (6).

The most common and natural choice for the g_i's is to take $g_i = e_i$ the i-th column of the identity matrix. From an algorithmic point of view, the simplest form of the algorithm can be written as follows.

Hessenberg procedure

1: **for** $j = 1, 2, \cdots, m$ **do**
2: $h_{ij} = e_i^T(Av_j),\ i = 1 : j$
3: $v = Av_j - \sum_{i=1}^{j} h_{ij}v_i$
4: $h_{j,j+1} = e_{j+1}^T v$ and $v_{j+1} = v/h_{j+1,j}$
5: **end for**

About Hessenberg

The 1996 publication [12] provides some information about Karl Hessenberg. We learn that he was born on September 8, 1904 in Frankfurt (Main). His father was a lawyer and his grandfather was a physician, also well-known in Germany as the author of the fairy tale 'Struwwelpeter'. From 1925 to 1930 he studied electronic engineering at the Darmstadt Technological Institute. Starting in 1931 he spent two years at the Faculty of Mathematics, where he studied under Professor A. Walther after which he spent another two years working at the Worms power plant and then one year with the company A.E.G. We learn that he published a total of 4 articles in electromechanical journals. He received a doctorate from Darmstadt University of Technology in 1942. In fact, the article [12] investigates a widespread mistake in the litterature regarding the common citation to his thesis [16] – which we used at first as obtained from many other sources. The title and year of this citation are both erroneous. The actual citation as provided in [12], with a scan of the degree certificate as a proof, is [17]. It is worth noting that the original work [17] was in a doctoral thesis produced at the height of the second World War at the University of Darmstadt in Germany.

Arnoldi's contribution, 1951

Wilkinson's treatise on eigenvalue problems includes a detailed discussion on the reduction to Hessenberg form [49, pp. 357-382]. In particular, Hessenberg's process can be viewed as a form of Gaussian elimination and partial pivoting can be added. In his discussion Wilkinson warns that while the relation $AV - VH \approx 0$ holds in a 'most satisfactory manner', the resulting similarity transformation V itself may be ill-conditioned, commenting that "we may contrast this with the situation which would have existed if V were unitary [4] " [49, p. 364].

It took another ten years after Hessenberg's work, before a method is developed that achieved just this. In an article that appeared in 1951 Walter E. Arnoldi developed a method to transform A into Hessenberg form by building a reduction matrix V that is orthogonal[4]. The Arnoldi process can actually be viewed as a form of the Gram-Schmidt process: we begin with a vector v_1 of norm 1, then at the j step, v_{j+1} is produced by orthonormalizing Av_j against $v_1, v_2, \cdots, v_j$. This is repeated for $j = 1, \cdots, m$. Thus, linear independence is enforced via an orthogonality requirement. Compared with the Hessenberg process, Arnoldi's procedure can be written in the

[3] This is the 'upper' Hessenberg form. A lower Hessenberg form is defined similarly.

[4] A unitary matrix is a square matrix whose columns are orthonormal. A nice feature of unitary matrices is that they preserve lengths and as such they tend not to amplify numerical errors. An $m \times n$ matrix (with $m > n$) is said to be orthogonal if its columns are orthonormal.

same form as (6) but the scalars h_{ij} are now replaced with scalars that satisfy the conditions:

$$v_{j+1} \perp v_1, v_2, \cdots, v_j, \quad \text{and} \quad v_{j+1}^T v_{j+1} = 1$$

instead of (7). In its simplest form, the Arnoldi procedure can be written exactly as the *Hessenberg procedure* shown earlier. However, instead of being calculated as in lines 2 and 4 of the Hessenberg procedure, the scalars h_{ij} and $h_{j+1,j}$ are now calculated as follows:

$$h_{ij} = v_i^T A v_j, \ i = 1, \cdots, j; \quad h_{j+1,j} = \|v\|_2.$$

Actually, in Arnoldi's original article, the vectors v_{j+1} were not normalized, so the scalars $h_{j+1,j}$ were set equal to one and the resulting vectors are orthogonal but not orthonormal.

Another important point to make is that Arnoldi considered mostly the case when $m < n$, in fact $m \ll n$, so he did not consider his algorithm to be a means to transform a matrix to Hessenberg form via a similarity transformation. This constitutes a significant departure from the spirit of the related publications by Krylov in 1931 and Hessenberg in 1942.

The contributions of this particular paper are not well appreciated in the literature as they are often cited from sources other than the original work. It is therefore worthwhile to discuss its content in some detail. The main focus of the paper was not to introduce the orthogonal version of the Hessenberg process but rather to provide an interpretation of the method introduced in the now well-known 1950 article by Lanczos, as a *Galerkin method*. A Galerkin method, or process, is a well-known technique to extract an approximation to a given problem from a subspace. A summary of the approach is provided in the next section. The procedure that is introduced by Arnoldi is just an extension or a by-product of this interepretation. Here is part of what the introduction of the paper says *"An interpretation of Dr. Cornelius Lanczos' iteration method, which he has named "minimized iterations", is discussed in this article, expounding the method as applied to the solution of the characteristic matrix equations both in homogeneous and non-homogeneous form. This interpretation leads to a variation of the Lanczos procedure which may frequently be advantageous by virtue of reducing the volume of numerical work in practical applications.* As can be confirmed by reading the paper, Arnoldi exploits what is referred to as "a classical method for reducing the matrix order", which is nothing but a Galerkin projection method. There are no references to either Krylov or Hessenberg. In fact there are only 3 references: one to the Lanczos paper [25], one to a 1937 article by Aitken on a form of LU factorization known as "pivotal condensation" and finally a reference to a book by Duncan on Galerkin methods [6]. The 'Galerkin' approach is key to Arnoldi's exposition. Most of the paper, sections 1 to 5, is about interpreting Lanczos' method for solving both linear systems ('inhomogeneous case') and eigenvalue problems ('homogeneous case') as Galerkin procedures. Only in section 6 does the author finally introduce his own new addition which he simply calls "The Galerkin treatment ...". A major contribution here is the introduction of a new viewpoint - namely to regard the methods proposed by Lanczos, both for linear systems and eigenvalue problems, as projection type methods. At the time this was rather innovative.

About Arnoldi

Walter Edwin Arnoldi was born on December 14, 1917 in New York (NY), and died on October 5, 1995 in Hartford (CT). He received a degree in Mechanical Engineering from Stevens Institute of Technology, then a Masters from Harvard (in around 1939). Then he worked for the United Technologies Corporation from 1939 until retirement in 1977. His 1951 article appears to be the only scientific publication authored by him but it clearly has had an enormous impact in numerical linear algebra.

The Galerkin approach

As a follow-up for the previous section, we provide a brief summary of the Galerkin projection approach for linear systems and for eigenvalue problems. Suppose we need to solve a linear system of the form $Ax = b$ where $A \in \mathbb{R}^{n \times n}$, $b \in \mathbb{R}^n$ and the unknown x is in $\mathbb{R}^n$. We are given a subspace K of dimension $m \ll n$, with an orthonormal system $V = [v_1, v_2, \cdots, v_m]$ and we wish to find an approximate solution $\tilde{x}$ to the original system that belongs to K. The (orthogonal) projection procedure finds this approximation by expressing

$\tilde{x}$ as $\tilde{x} = Vy$ with $y \in \mathbb{R}^m$ and then it imposes the Galerkin condition

$$b - A\tilde{x} \perp K \rightarrow V^T AVy = V^T b. \qquad (8)$$

The result is a small $m \times m$ system which be easily be solved, assuming $V^T AV$ is nonsingular, to obtain $\tilde{x} = Vy$. A slighty more general approach would entail using a different subspace L of dimension m for the Galerkin condition, i.e., we would write $b - A\tilde{x} \perp L$, leading to what is known as a non-orthogonal (i.e., 'oblique') projection method [37].

The Galerkin procedure works similarly for eigenvalue problems. If the problem to solve is $Ax = \lambda x$, we would find an approximate eigenvalue $\tilde{\lambda} \in \mathbb{C}$ and associated approximate eigenvector $\tilde{u} = Vy$ by imposing the condition [5] $V^H(A - \tilde{\lambda}I)\tilde{u} = 0$ which leads to the projected m-dimensional eigenvalue problem $(V^H AV - \tilde{\lambda}I)y = 0$. Krylov subspace methods are Galerkin-type methods in which K is a Krylov subspace K_m of the form (1) where v is some initial vector, see, e.g., [37], [29] among others for details.

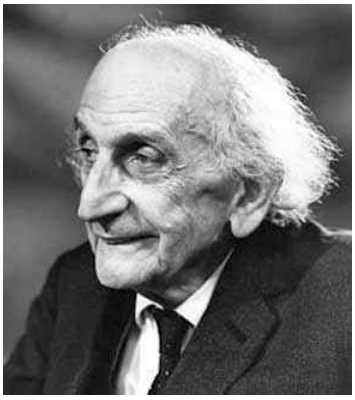
Breakthrough: the 1950 article by Lanczos

The article by Lanczos [25] referred to by Arnoldi is a truly impressive piece of work by its originality and vision. The subspace K in the Galerkin approach utilized by Arnoldi method and the symmetric Lanczos procedures are Krylov subspaces of form K_m in (1). Both methods obtain an orthonormal basis of the subspace K_m by a form of Gram-Schmidt procedure. An interesting observation here is that when A is Hermitian then $V^H AV$ is also Hermitian so the Hessenberg matrix H_m obtained from the Arnoldi procedure must be tridiagonal. This is an important result leading to a major simplification of the algorithm, namely the symmetric Lanczos algorithm, and it is obtained from the Arnoldi process by using a straightforward argument. Lanczos did not make this observation as he was focusing on symmetric systems, but he arrived at his algorithm from a different path. The paper addresses several problems, emphasizing solutions based on approximating functions by polynomials, pade-type approximations, generating functions, etc. It is rather dense in interesting ideas leading up to the algorithm which was presented as a realistic alternative to the techniques shown in the first part of the paper, for the situation when the matrix is not small.

Here is what the Acknowledgment section of the article states: *The present investigation contains the results of years of research in the fields of network analysis, flutter problems, vibration of antennas, solution of systems of linear equations, encountered by the author in his consulting and research work for the Boeing Airplane Co., Seattle, Wash. The final conclusions were reached since the author's stay with the Institute for Numerical Analysis, of the National Bureau of Standards.* The note ends by thanks expressed at his supervisers at Boeing as well at NBS.

About Lanczos

Cornelius Lanczos was Born under the name Kornél Löwy on February 2, 1893 in Székesfehérvár (Hungary) and died on June 25, 1974 in Budapest, Hungary. Lanczos had a rather turbulent life to say the least, as historical events unfolding around him were often precarious and forced him to move a few times. Yet from a scientific viewpoint what a productive and rich life it was.



Early in his career, Lanczos held a few positions in Germany: Freiburg '21, Frankfurt '24, Berlin '28, Frankfurt '29. His position in Berlin in 1928-29 was an assistantship to Albert Einstein. The two maintained a correspondance for many years thereafter. In 1931, Lanczos was confronted with the rise of violent anti-semitic sentiment in Germany and decided to join Purdue University,

[5] The matrix V^H is the transpose conjugate of V, i.e, $V^H = \overline{V}^T$.

at first with a visiting position in the physics department and then as a full professor in 1932. He left Purdue in 1944 for Boeing aircraft company and then joined the Institute for Numerical Analysis at the National Bureau of Standards in 1949.

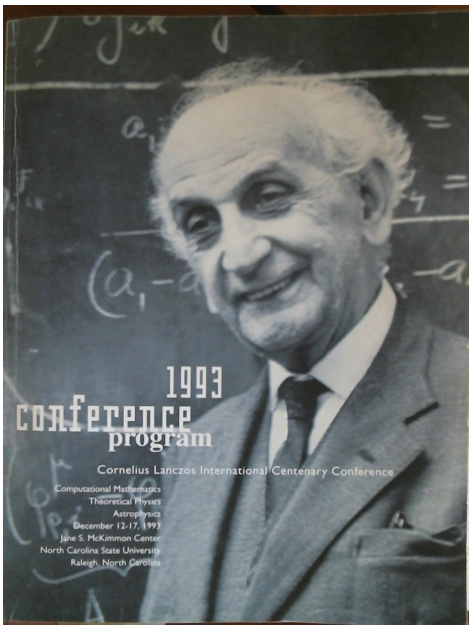
There he worked until 1952 when, during the McCarthy era, the atmosphere of suspicion became unbearable and he had to leave again, accepting an invitation for a visiting position at the Theoretical Physics Department of the Dublin Institute for Advance Study in Ireland. This invitation was extended to him by Erwin Schrödinger, the quantum physicist, who himself fled Austria in 1933 for political reasons. A year later he received a permanent position at the same institute and remained in Dublin until 1974 when he passed away during a visit to the Eötvös Lóránd University in Budapest.

The years 1949 and 1952, which he spent at the Institute for Numerical Analysis (INA) at the National Bureau of Standards (NBS), were particularly productive for Lanczos. Among his colleaues at the INA were a number of well-known figures in Numerical Analysis, including Olga Taussky-Todd, John Todd, and George Forsythe. Also, his well-known work on what we call the Lanczos algorithm was from this period.

Lanczos' work showed deep insight rooted in approximation theory as well as physics. Often his work in numerical linear algebra exploited the viewpoint of polynomial approximation to find approximate solutions to various probllems.

The contributions of Lanczos to the field of numerical analysis were monumental. Remarkably, Lanczos was initially a theoretical physicist and he kept working in both fields of mathematics and physics to the end of his life. His dissertation in 1920 was on the use of quarternions for the treatment of special relativity and electrodynamics. His first major contribution was in general relativity when he published an exact solution to Einstein's field equations for gravity. In 1942 he independently developed (along with G. C. Danielson) what is now known as the Fast Fourrier Transform (FFT). The authors discuss the cost-effectiveness of the method but did not realize that the operation count was of orfer $N \log N$. This versatility is reminescent of a characteristic of the big mathematicians of earlier times. Not surprisingly, his research and teaching were often characterized by the inclusion of physics insight to mathematical arguments [13].

Perhaps the best testimony to the breath and impact of his work can be gauged from the talks given at the 1993 Cornelius Lanczos International Centenary conference, where well-known mathematical physicists (e.g., Roger Penrose) gave presentations alongside numerical analysts (e.g., Gene Golub).



This dual view of applied mathematics, which Lanczos liked to call 'workable mathematics' [13] is undoubtedly what gave Lanczos' contribution to physics, applied mathematics, and numerical linear algebra, their unique character and their depth. In numerical linear algebra, his algorithm on 'minimized iteration' has had a major impact across disciplines. My own research in the 1970s included a study of the Lanczos algorithm. At that time I came across a breakthrough article that showed how his method was successfully put to work in the study of normal modes of oceans [4], and then nearly 50 years later a modification of the same method was instrumental in the study normal modes of planets [39], [40].

About the INA

The Institute for Numerical Analysis was an institute located on the campus of the University of California at Los Angeles (UCLA) and created by NBS in 1947 with a goal of taking advantage

of the then nascent computer technology. At the origin of the INA was the desire by the Census Bureau to transfer funds to NBS to facilitate the purchase of a suitable computer to help with the 1950 census. Note that the National Bureau of Standards, which changed its name to the "National Institute of Standards and Technology (NIST)" in 1988, was created in 1901 by the US Departnent of Commerce and that the Census bureau was also part of the Department of Commerce. With the concurrent interest of the Office of Naval Research (ONR), it was agreed that NBS and ONR should jointly establish a center, to develop, as well as use, computers.

The INA was the brain child of John Curtis (NBS) with the support of Mina Rees (ONR). In 1946 John Curtis drafted a plan (called a 'prospectus') to re-organize the Applied Mathematics Division (AMD) at NBS in 1947 and the INA was one of the components of this plan. Funds for this purpose were transferred in September 1946 and the institute opened shortly thereafter.

The first INA director was Douglas Hartree a well-known British physicist (of Hartree-Fock fame). Hartree perceived a serious problem with the negative attitude of mathematicians toward numerical analysis [19] and saw the INA as an opportunity to "overcome this attitude". The INA attracted the best worldwide scientists in areas related to applied mathematics and it played a major role in promoting numerical analysis as a field. Thus, the list of distinguished scientists who visited the INA or had been partly affiliated with it includes [19]: J. von Neumann, R. Courant, E. Teller, S. Lefschetz, N. Wiener, D. Hartree, and others. Well-known numerical analysts affiliated with the INA include J. Todd, O. Taussky-Todd, G. Forsythe, M Hestness, E. Stiefel, and C. Lanczos. In 1954, the INA ceased to operate as an NBS institute and became part of UCLA.

The impact that the INA had in computer science, numerical analysis and numerical linear algebra cannot be overstated. For example, George Forsythe joined Stanford University in 1957 and in 1959 started one of the first Computer Science departments in the US, a task that was rather challenging at the time [23]. Donald Knuth [23] stated that *"It is generally agreed that he, more than any other man, is responsible for the rapid development of computer science in the world's colleges and universities. His foresight, combined with his untiring efforts to spread the gospel of computing, have had a significant and lasting impact..."* The establishment of a Computer Science department at Stanford had a powerful trend setting effect for other universities. In 1970, another former member of the INA named Marvin Stein [6], served as founding department head for my own (computer science) department at the University of Minnesota. Marvin joined the INA in 1948 and contributed to the development of the Conjugate Gradient method while working under the supervision of Magnus Hestenes.

Iterative methods for linear systems

Early methods for solving linear systems of the form

$$b - Ax = 0 \tag{9}$$

were dominated by 'relaxation' techniques. Starting with some inital guess $x^{(0)}$ to the solution vector, iterative methods generate a sequence of approximations $x^{(k)}$, for $k = 1, 2, \cdots$, to the solution, which we will denote by x_*. Relaxation methods work by modifying one component of the current iterate, say $x^{(k)}$, at a time. Thus, we will modify the i-th component of $x^{(k)}$, for some i, in such a way that the i-th component of the new residual vector $r = b - Ax$ for the newly obtained x, i.e., $x^{(k+1)}$ is equal to zero, i.e., we need to enforce the condition $e_i^T(b - Ax^{(k+1)}) = 0$. This is repeated on another component i until convergence is reached, e.g., until the norm of the residual vector $b - Ax^{(k)}$ is small enough. This basic idea seems to have been first developed by Gauss ca 1817, followed by Jacobi (1850), and Seidel (1874). These can be viewed as basic Galerkin-type methods whereby the subspace K is taken to be the one-dimensional subspace $K = \mathbf{span}\{e_i\}$, where e_i denotes the i-th column of the identity matrix. These methods were the main iterative methods for solving linear systems until up to the early 1970s. Here is what Richard Varga writes in his seminal book [47]

As an example of the magnitude of problems that have been successfully solved on digital

[6] Marvin Stein was a member of the university of Minnesota from 1955 until he retired in 1997. He passed away on Feb 27, 2015 at the age of 90.

computers by cyclic iterative methods, the Bettis Atomic Power laboratory of the Westinghouse Electric Corporation had in daily use in 1960 a two-dimensional program which would treat as a special case, Laplacean-type matrix equations of order 20,000. He then adds as a footnote: ... *Even more staggering is Bettis' use of a 3-Dimensional program called "TNT-1", which treats coupled matrix equations of order 108,000.* So in 1960, one could solve a discretized elliptic partial differential equation of size $\approx 100,000$ equations - and this was the state-of-the-art.

Another track of iterative methods, not as visible, was one based on descent-type methods, which amounted to *One-dimensional projection processes*. The best known method among these is the steepest descent algorithm introduced by Cauchy in [1847] for solving nonlinear equations. It was Kantorovitch who introduced it in the form we know today for linear systems for symmetric positive definite matrices in 1945 by applying Cauchy's approch to the minimization of the objective function

$$J(x) = \frac{1}{2}x^T Ax - b^T x. \qquad (10)$$

Cimmino's method [1938] and Kaczmarz's method [1937] also developed independently what we can term 'Line-search' methods in the direction of a row or column of A.

All these techniques can be viewed in a unified way as one-dimensional projection (Galerkin) processes. Given an initial guess x with residual vector $r = b - Ax$, and two nonzero vectors d (search direcrtion) and e (constraint direction), the idea is to apply a projection method on the one-dimensional subspace $K = \mathbf{span}\{d\}$ and orthogonally to the one-dimensional space $L = \mathbf{span}\{e\}$. This means that the new iterate is defined by the following equations

$$\tilde{x} := x + \alpha d \quad \text{and} \quad b - A\tilde{x} \perp e. \qquad (11)$$

Since we have $b - A\tilde{x} = r - \alpha Ad$, this defines[7] $\alpha \equiv (r, e)/(Ad, e)$ and the new iterate, provided the denominator is not zero. The process is repeated with a new pair of directions each time, until convergence is reached. For example, at each iteration of the steepest descent algorithm we define d and e to be each equal to the current residual $r = b - Ax$. In the minimal residual method for a general nonsingular matrix A, at each iteration d is defined to be equal to r but e is defined to be equal to Ar. In Kaczmarzs method d is set to $A^T e_i$, while e is set to be equal to e_i and this is repeated in a cycle for $i = 1, \cdots, n$.

[7]Here (x, y) denotes the Euclidean inner product in $\mathbb{R}^n$.

Polynomial iteration

In some of the methods just described the iteration takes the form $x_{k+1} = x_k + \beta_k r_k$, where $r_k = b - Ax_k$ is the current residual. It can be seen that the residual vector r_{k+1} satisfies $r_{k+1} = (I - \beta_k A)r_k$ and so by induction $r_{k+1} = p_{k+1}(A)r_0$ where p_{k+1} is the *residual polynomial* $p_{k+1}(t) = (1 - \beta_k t)...(1 - \beta_0 t)$, which is a polynomial of degree $k + 1$ that satisfies $p_{k+1}(0) = 1$. In 1950 Frankel considered a 'second-order' iteration of the form

$$x_{k+1} = x_k + \beta_k d_k, \quad \text{where} \quad d_k = r_k - \alpha_k d_{k-1}, \qquad (12)$$

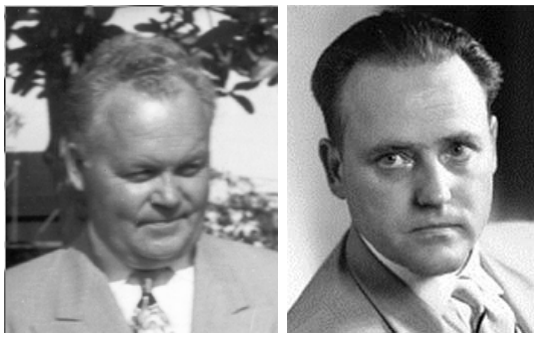
starting with $d_{-1} \equiv 0$. This leads to a more general polynomial iteration than those described above. For example, it is easy to defined scalars that will yield 'optimal' polynomials based on Chebyshev polynomials of the first kind, as was observed in the seminal work of Golub and Varga [15].

This 'approximation theory' viewpoint was adopted by several authors and was part of the Lanczos approach to solving linear systems [27], [26], [42], [5]. In essence, Krylov subspace methods will implicitly exploit residual polynomials that are optimal in different ways.

Krylov methods take off: The Conjugate Gradient (CG) algorithm, 1952

The conjugate gradient method was developed independently Magnus Hestenes [UCLA] and Eduard Stiefel [ETH, Zürich], [18]. In preparation for the semi-centennial anniversary celebration of the National Bureau of Standards, several researchers were invited in the summer of 1951 and the list of participants at INA was unusually long [19]. Among these visitors was Eduard Stiefel a mathematician from ETH Zürich. As talks were being prepared for the upcoming symposium organized to celebrate the anniversary of NBS, soon after Stiefel's arrival it was discovered that the two groups, Stiefel on the one hand, and Hestenes on the other had developed the same

method independently [19] – using very different arguments. Stiefel extended his stay at the INA through June 1952 to work on the publication of a joint paper.



Magnus Hestenes and Eduard Stiefel

At almost the same time Lanczos developed a similar method using different notation and a very different viewpoint [27] based on a biorthogonalization process described in his 1950 article on eigenvalue problems [25]. Here is what Hestenes and Todd say about this [19, p. 61]: *"It occurred to none of us at that time that these relations could be used effectively in an algorithm for solving linear equations in n steps. We were not aware of this connection until the conjugate gradient routine had been devised by geometrical considerations. It is clear therefore that the conjugate gradient algorithm is an easy consequence of results given by Lanczos. This led Lanczos to devise an alternative version of the conjugate gradient algorithm, which he called a Method of Minimized Iterations"*. Lanczos' method dealt with the nonsymmetric case, and in exact arithmetic it is a form of the bi-CG algorithm [8] in which the approximate solution is extracted from what Lanczos called a "q expansion", which is nothing but an expansion of the solution that exploits the bi-conjugate basis[8]. His paper appeared in the same journal, and the institution was also the same, i.e., the INA. Lanczos' paper appeared in July 1952, and the one by Hestenes and Stiefel in December 1952. It is a mystery as to why Lanczos did not collaborate with the Hestenes group at the INA. He was clearly a little ahead.

[8]The nonsymmetric Lanczos algorithm - from which the bi-conjugate gradient (bi-CG) method can be derived - generates two sets of directions that are orthogonal to each other (bi-orthogonal) instead one set of orthogonal vectors. The approximate solution to a linear system can be expanded in one these sets of vectors.

The CG algorithm and the Lanczos tridiagonalization algorithms were discoveries of the utmost importance in numerical linear algebra - even though this may not have been understood at the time. The class of Krylov Subspace Methods has been cited in various sources as one of the top 10 algorithms of the 20th century, see for example [45].

About Hestenes

Magnus Rudolph Hestenes was born on February 13, 1906, in Bricelyn, Minnesota and he died on May 31, 1991, in Los Angeles, CA. He received a Ph.D. from the University of Chicago in 1932. Then he joined the faculty of UCLA in 1947 and kept his position there until his retirement in 1973. He was associated with the INA and listed as 'UCLA liaison' member. His work dealt with calculus of variations, optimal control, gradient-type methods for linear systems and eigenvalue problems

About Stiefel

Eduard L. Stiefel was born on April, 21 1909 in Zürich and died on November 25, 1978 in Zürich. He received a Ph.D. in 1935 from the Swiss Federal Institute of Technology (ETH Zurich). He became full professor at ETH Zurich in 1943, and founded the Institute for Applied Mathematics in 1948. Here is what Hestenes and Todd say about him [19, p. 29]: *"Eduard L. Stiefel was a very versatile mathematician. He began his career as a topologist and made notable contributions in this field. (...) He was well versed both in pure and in applied mathematics. He anticipated the coming of high-speed digital computers and was instrumental in the development of such a computer at ETH. This led him to an intensive study of numerical methods. (...) Stiefel received many honors and participated in the government of Zurich."*

Aftermath of CG article:

The CG Method did not initially receive the attention it deserved. The main reason for this is that it was regarded as a *direct (n-step) method* for solving linear systems and as such it did not compare well with Gaussian elimination as it was costly and more importantly, unstable. In his book, Householder [20, Ch.5.7] presents the method in the section of direct methods, stating

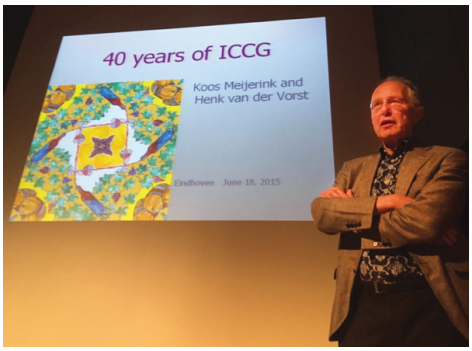
in effect that the method was truly novel - but he did not mention stability issues. In a 1959 article, Engeli, Ginsburg, Rutishauser and Stiefel [8] discussed a new viewpoint: They regarded the CG algorithm as an iterative procedure. However, one had to wait until the early 1970s before this idea started being adopted. Specifically, Reid [35] promoted the use the Conjugate Gradient method as an iterative technique but now with an emphasis put on solving large sparse linear systems. At the same time Chris Paige in his PhD thesis [31], see also [32], [33], proposed the first comprehensive study of the Lanczos process in the presence of inexact arithmetic while Kaniel [22] analyzed the theoretical convergence of the CG and Lanczos algorithms based on spectral distributions. One may say that these seminal publications really marked the beginning of modern Krylov methods. A little later, the book by Parlett [34] played a significant role in reviving the Lanczos algorithm for eigenvalue problems.

It took about 25 years for the conjugate gradient method to be accepted as a viable procedure for solving linear systems iteratively. A history of these first 25 years of the CG method can be found in the detailed survey article by Golub and O'Leary [14]. Another major development, namely the idea of preconditioning, will come in the late 1970 and will give rise to an approach that became one the preferred iterative approaches for solving large sparse symmetric positive definite linear systems.

Preconditioning

The idea of preconditioning is rather old. Golub and O'Leary [14] trace the term 'preconditioning' back to Turing [1948]. Polynomial preconditioners were invoked in many paper including an early 1937 article by Cesari [3]. Lanczos mentioned it in his 1952 paper [27] as did Stiefel in a review paper that appeared in 1959 [42]. Forsythe [10] uses the term explicitly in 1953: *With the concept of "Ill conditioned" systems $Ax = b$ goes the idea of "preconditioning" them. Gauss [1823] and Jacobi [1845] made early contributions to this subject."*

Later came the idea of incomplete factorizations by Buleev [2], Varga [46], Oliphant [30], Stone [43], and others.



Henk van der Vorst, at the dinner speech (given with Koos Meijerink) at 'Preconditioning 2015'

However, preconditioners based on Incomplete LU (ILU) factorizations became popular with a paper by Meijerink and Van der Vorst that appeared in 1977 [28]. Thus, the Incomplete Cholesky Conjugate Gradient (ICCG) became a very popular approach for solving sparse symmetric positive definite matrices. Interestingly, the article itself, which made a huge impact, took over 7 years after its first submission before appearing as the story was told by the authors at the banquet of the *2015 Preconditioning* conference.

Krylov methods: the 'nonsymmetric' period

As mentioned earlier the method developed by Lanczos in his 1952 paper [27] worked for nonsymmetric linear systems. Fast forward to 1976 and one will find essentially the same method under the name Bi-Conjugate Gradient (BiCG) proposed by Fletcher [9]. From these basic Krylov subspace methods that exploit bi-orthogonality, a number of other algorithms were later invented: Conjugate Gradient Squared (CGS) [41], Bi-Conjugate Gradient Stabilized (Bi-CGSTAB), [44], Transpose-Free Quasi Minimal Residual method (TFQMR) [11], and a few others.

At the same time as these methods were being developed another, independent, track of nonsymmetric Krylov Subspace Methods was also emerging. These were "Orthogonal projection" techniques which aimed to minimize the residual norm of the iterates. A method dubbed 'ORTHOMIN' [48] seems to have been the first in this category followed by a number of others: Axelson's method [1], 'ORTHODIR' [21], the Generalized Conjugate Residual (GCR) method [7], and the Generalized Minimal Residual method

(GMRES) [38]. This period was characterized by a significant flury of activity whereby reseachers developed elaborate variants of algorithms from both tracks, i.e., orthogonal and bi-orthogonal. An outstanding coverage of this class of methods can be found in the rather exhaustive volume by Meurant and Tebbens [29], which has over 1000 references, and provides an in-depth look at both theoretical and practical aspects. Activity in Krylov subspace accelerators started to subside toward the mid to late 1990s as it was understood that preconditioners played a bigger role in the success of this class of methods than the accelerators.

Conclusion

This quick look at Krylov methods may have led the reader to one or more of the following striking observations. The first observation concerns the magnificent progress of ideas from Krylov's original article whose significance today may not be well appreciated but which was a decisive breathrough for its time. The second observation, is the fascinating mix of characters behind the methods developed. Among them are people like Hessenberg and Arnoldi who made very pointed and narrow contributions, yet impactful ones. Others were true giants in our field, whose discoveries triggered shockwaves across several scientific disciplines. In this regard, the stature of Lanczos as a scientist is difficult to match. As alluded to earlier, he was a scientist from a gone-by era, of the kind who worked by themselves (all of his articles are single authored except three) and who had an amazing ability to shape knowledge. Lanczos was remarkable not only for his truly exceptional contributions, and his broad grasp of science, but also by his character: A humble and kind individual who did not seek recognition – someone who endured difficult times in his life and understood struggles of others. On the day of his burial, Rabbi Sandor Scheiber of Budapest said: [13, p. 123] *"... He never spoke of himself, but stood up for the recognition of others... He was troubled by the misery of people and that he could not do enough to relieve it."*

The third observation one can make by taking a look at current research in science and engineering is that today the field of Krylov subspace methods is as vibrant as ever. Despite the relatively long history of KSMs, their users keep finding new ways of exploiting them to solve new problems in various applications. In fact, with the emergence of data related methodologies, it is inevitable that we will see an increased interest in KSMs. After all the core idea behind Krylov Subspace Methods is that they are in essence nothing but a form of dimension reduction techniques similar to those invoked to deal with large datasets.

Acknowledgments

A historical paper like this one cannot be written without available resources provided by others. The authors of these articles and testimonals should be commended and thanked for graciously taking the time and effort to contribute to our knowledge of the science and the lives of the people behind it. In this regard, the most fascinating and informative read for me in preparing this article has been the book on Lanczos's life by Barbara Gellai [13]. I also read with considerable interest the article by Hestenes and Todd [19], an impressive document that discussed in great detail the history and impact of the Institute of Numerical Analysis at NBS.

I would like to thank Ron Boisvert for inviting me to take part in the symposium titled *75th Anniversary of Mathematics and Statistics at NIST*, and Barry Schneider for the idea that I contribute this article to CiSE. The manuscript benefitted from numerous helpful suggestions made by Lorena Barba. I wish to also thank Ahmed Sameh for his comments and encouragements on an earlier draft. This work was supported by NSF Grant DMS-1912048

REFERENCES

1. O. Axelsson, *Conjugate gradient type-methods for unsymmetric and inconsistent systems of linear equations*, Linear Algeba Appl., 29 (1980), pp. 1–16.
2. N. I. Buleev, *A numerical method for the solution of two-dimensional and three-dimensional equations of diffusion*, Math. Sb, 51 (1960), pp. 227–238. (in Russian).
3. L. Cesari, *Sulla risoluzione dei sistemi di equazioni lineari per approssimazioni successive*, Atti Accad. Naz. Lincei Rend. Cl. Sci. Fis. Mat. Nat., Ser. 6a, 25 (1937), pp. 422–428.

4. A. K. Cline, G. H. Golub, and G. W. Platzman, *Calculation of normal modes of oceans using a Lanczos method*, in Sparse Matrix Computations, J. R. Bunch and D. C. Rose, eds., Academic Press, 1976, pp. 409–426.
5. C. de Boor and J. R. Rice, *Extremal polynomials with applications to Richardson iteration for indefinite systems*, SIAM J. Sci. Stat. Comput., 3 (1982), pp. 47–57.
6. W. J. Duncan, *Galerkin's method in mechanics and differential equations*, R. & M 1798, 1938.
7. S. C. Eisenstat, H. C. Elman, and M. H. Schultz, *Variational iterative methods for nonsymmetric systems of linear equations*, SIAM J. Numer. Anal., 20 (1983), pp. 345–357.
8. M. Engeli, T. Ginsburg, H. Rutishauser, and E. Stiefel, *Refined Iterative Methods for Computation of the Solution and the Eigenvalues of Self-Adjoint Boundary Value Problems*, Birkhäuser, Basel/Stuttgart, 1959.
9. R. Fletcher, *Conjugate gradient methods for indefinite systems*, in Proceedings of the Dundee Biennal Conference on Numerical Analysis 1974, G. A. Watson, ed., New York, 1975, Springer Verlag, pp. 73–89.
10. G. E. Forsythe, *Solving linear algebraic equations can be interesting*, Bulletin of the American Mathemathical Society, 59 (1953), pp. 299–329.
11. R. W. Freund, *A Transpose-Free Quasi-Minimal Residual algorithm for non-Hermitian linear systems*, SIAM J. Sci. Comput., 14 (1993), pp. 470–482.
12. S. Fujino and E. Heil, *Who was k. hessenberg?*, Kyoto University Research Information Repository - RIMS Kokyuroku, (1996), pp. 208–217. (In Japanese).
13. B. Gellai, *The Intrinsic Nature of Things. The Life and Science of Cornelius Lanczos*, American Mathematical Society, Providence, Rhode Island, 2010.
14. G. H. Golub and D. P. O'Leary, *Some history of the conjugate gradient and Lanczos algorithms: 1948-1976*, SIAM Review, 31 (1989), pp. 50–102.
15. G. H. Golub and R. S. Varga, *Chebyshev semi-iterative methods, successive overrelaxation iterative methods, and second order Richardson iterative methods*, Numerische Mathematik, 3 (1961), p. 157168.
16. K. Hessenberg, *Auflsung linearer eigenwertaufgaben mit hilfe der hamilton-cayleyschen gleichung*, 1941. Doc. Ing. Thesis, T. H. Darmstadt - Note: this citation inferred from other works has an incorrect title and year. See the 1942 reference for corrections.
17. ———, *Die berechnung der eigenwerte und elgenlösungen linearer gleichungsysteme*, 1942. Doc. Ing. Thesis, T. H. Darmstadt.
18. M. R. Hestenes and E. L. Stiefel, *Methods of conjugate gradients for solving linear systems*, J. of Res. of the Nat. Bur. of Stand., Section B, 49 (1952), pp. 409–436.
19. M. R. Hestenes and J. Todd, *Nbs-ina the institute for numerical analysis - ucla 1947-1954*, Tech. Rep. NIST Special Publication 730, National Institute of Standards and Tectinology, Washington, DC, 1991.
20. A. S. Householder, *The Theory of Matrices in Numerical Analysis*, Blaisdell Pub. Co., 1964. Reprinted by Dover publishing Inc., NY, 1975.
21. K. C. Jea and D. M. Young, *Generalized conjugate-gradient acceleration of nonsymmetrizable iterative methods*, Linear Algeba Appl., 34 (1980), pp. 159–194.
22. S. Kaniel, *Estimates for some computational techniques in linear algebra*, Math. of Comput., 20 (1966), pp. 369–378.
23. D. E. Knuth, *George forsythe and the development of computer science*, Commun. ACM, 15 (1972), p. 721726.
24. A. N. Krylov, *On the numerical solution of equations whose solution determine the frequency of small vibrations of material systems (in russian)*, Izv. Akad. Nauk. SSSR Otd Mat. Estest., 1 (1931), pp. 491–539.
25. C. Lanczos, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. of Res. of the Nat. Bur. of Stand., 45 (1950), pp. 255–282.
26. ———, *Chebyshev polynomials in the solution of large-scale linear systems*, in Proceedings of the ACM, 1952, pp. 124–133.
27. ———, *Solution of systems of linear equations by minimized iterations*, J. of Res. of the Nat. Bur. of Stand., 49 (1952), pp. 33–53.
28. J. A. Meijerink and H. A. van der Vorst, *An iterative solution method for linear systems of which the coefficient matrix is a symmetric M-matrix*, Math. of Comput., 31 (1977), pp. 148–162.
29. G. Meurant and J. D. Tebbens, *Krylov Methods for Nonsymmetric Linear Systems - From theory to computations*, Springer Series in Computational Mathematics, vol. 57, Springer, 2020.
30. T. A. Oliphant, *An extrapolation process for solving linear systems*, Quart. Appl. Math., 20 (1962), pp. 257–267.
31. C. C. Paige, *The computation of eigenvalues and eigenvectors of very large sparse matrices*, PhD thesis, Lon-

don University, Institute of Computer Science, London, England, 1971.

32. ———, *Error analysis of the Lanczos algorithm for tridiagonalizing a symmetric matrix*, J. Inst. Math. Appl., 18 (1976), pp. 341–349.
33. ———, *Accuracy and effectiveness of the Lanczos algorithm for the symmetric eigenproblem*, Linear Algebra and its Applications, 34 (1980), pp. 235–258.
34. B. N. PARLETT, *The Symmetric Eigenvalue Problem*, no. 20 in Classics in Applied Mathematics, SIAM, Philadelphia, 1998.
35. J. K. REID, *On the method of conjugate gradients for the solution of large sparse systems of linear equations*, in Large Sparse Sets of Linear Equations, J. K. Reid, ed., Academic Press, 1971, pp. 231–254.
36. Y. SAAD, *On the rates of convergence of the Lanczos and the block Lanczos methods*, SIAM J. Numer. Anal., 17 (1980), pp. 687–706.
37. Y. SAAD, *Numerical Methods for Large Eigenvalue Problems-classics edition*, SIAM, Philadelphia, 2011.
38. Y. SAAD AND M. H. SCHULTZ, *GMRES: a generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.
39. J. SHI, R. LI, Y. XI, Y. SAAD, AND M. V. DE HOOP, *Computing planetary interior normal modes with a highly parallel polynomial filtering eigensolver*, in SC16: Proceedints of the International Conference for High Performance Computing, Networking, Storage and Analysis, IEEE, 2018, pp. 894–906. SC18, Dallas, TX, Nov. 11-16, 2018.
40. J. SHI, R. LI, Y. XI, Y. SAAD, AND M. V. DE HOOP, *Planetary normal mode computation: Parallel algorithms, performance, and reproducibility*, IEEE Transactions on Parallel and Distributed Systems, 32 (2021), pp. 2609–2622.
41. P. SONNEVELD, *CGS: a fast Lanczos-type solver for nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 10 (1989), pp. 36–52.
42. E. L. STIEFEL, *Kernel polynomials in linear algebra and their applications*, U. S. National Bureau of Standards, Applied Mathematics Series, 49 (1958), pp. 1–24.
43. H. S. STONE, *Iterative solution of implicit approximations of multidimensional partial differential equations*, SIAM J. Numer. Anal., 5 (1968), pp. 530–558.
44. H. A. VAN DER VORST, *Bi-CGSTAB: A fast and smoothly converging variant of Bi-CG for the solution of nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 12 (1992), pp. 631–644.
45. H. A. VAN DER VORST, *Krylov subspace iteration*, Computing in Science and Engineering, 2 (2000), pp. 32–37. Special issue on the 'Top 10 Algorithms of the Century'.
46. R. VARGA, *Factorization and normalized iterative methods*, in Boundary problems in differential equations, R. Langer, ed., Univ. of Wisconsin Press, Madison, 1960, pp. 121–142.
47. R. S. VARGA, *Matrix Iterative Analysis*, Prentice Hall, Englewood Cliffs, NJ, 1962.
48. P. K. W. VINSOME, *ORTHOMIN: an iterative method for solving sparse sets of simultaneous linear equations*, in Proceedings of the Fourth Symposium on Resevoir Simulation, Society of Petroleum Engineers of AIME, 1976, pp. 149–159.
49. J. H. WILKINSON, *The Algebraic Eigenvalue Problem*, Clarendon Press, Oxford, 1965.