UNIVERSITY OF MINNESOTA TWIN CITIES

A brief journey into the past of iterative methods for solving sparse linear systems.

Yousef Saad

Department of Computer Science and Engineering

University of Minnesota

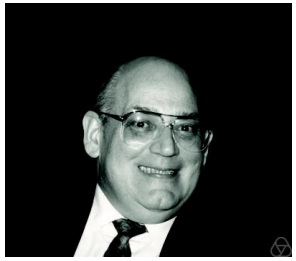
ICERM – 75 years of Math Comp

Nov. 2, 2018

Early days..

The idea of solving large systems by linear equations by iterative methods is certainly not new, dating back to at least Gauss (1823). Later Southwell (1946) and his school gave real impetus to the use of iterative methods (...)

... The basis for much of the present activity in this area of numerical analysis concerned with cyclic iterative methods is a series of papers by Frankel (1950), Geiringer (1949), Reich (1949), Stein and Rosenberg (1948), and Young (1950), all of which appeared when digital computers were emerging with revolutionary force.



Richard Varga, 1962

What is Relaxation?

- Term likely introduced by Southwell (1940s)
- To solve: $Ax = b$ View as a collection of equations:

  $a_i x = \beta_i$ ($a_i = i$-th row)
- Residual: $r = b - Ax.$

Relaxation: Modify i-th component of x into $x_i^{(new)} := x_i + \delta_i$ so that: $r_i^{(new)} = 0.$

$$a_i(x + \delta e_i) = \beta_i \quad \longrightarrow \quad \delta = \frac{r_i}{a_{ii}}$$

- Do this in a certain order ...

➤ Southwell's method (1945) : select for i the one corresponding to the largest entry in r each time

➤ 'cyclic' iterations – term 'stationary' used by Young (1950s): cycle through the components.. Gauss-Seidel, Jacobi, ..

➤ 'chaotic' or 'asynchronous' iterations

Gauss letter to Gerling (1823)

In order to eliminate indirectly,[6] I note that, if 3 of the quantities a, b, c, d are set equal to 0, the fourth gets the largest value when d is chosen as the fourth.[8] Naturally every quantity must be determined from its own equation, and hence d from the fourth. I therefore set $d = -201$ and substitute this value. The absolute terms then become: $+5232, -6352, +1074, +46$; the other terms remain the same.

Now I let b take its turn, find $b = +92$, substitute, and find the absolute terms: $+4036, -4, -3526, -506$. And thus I continue until there is nothing more to correct. Of this whole calculation I actually write only the following table:[9]

	$d=-201$	$b=+92$	$a=-60$	$c=+12$	$a=+5$	$b=-2$	$a=-1$
+ 6	+5232	+4036	+ 16	−320	+ 15	+41	−26
− 7558	−6352	− 4	+ 776	+176	+111	−27	−14
−14604	+1074	−3526	−1846	+ 26	−114	−14	+14
+22156	+ 46	− 506	+1054	+118	− 12	0	+26.

Insofar as I carry the calculation only to the nearest 2000-th of a second, I see that now there is nothing more to correct. I therefore collect

[source: Math. Tables and other aids to computation, vol. 5, # 36 (Oct. 1951). Translation due to G. Forsythe]

Ends letter with:

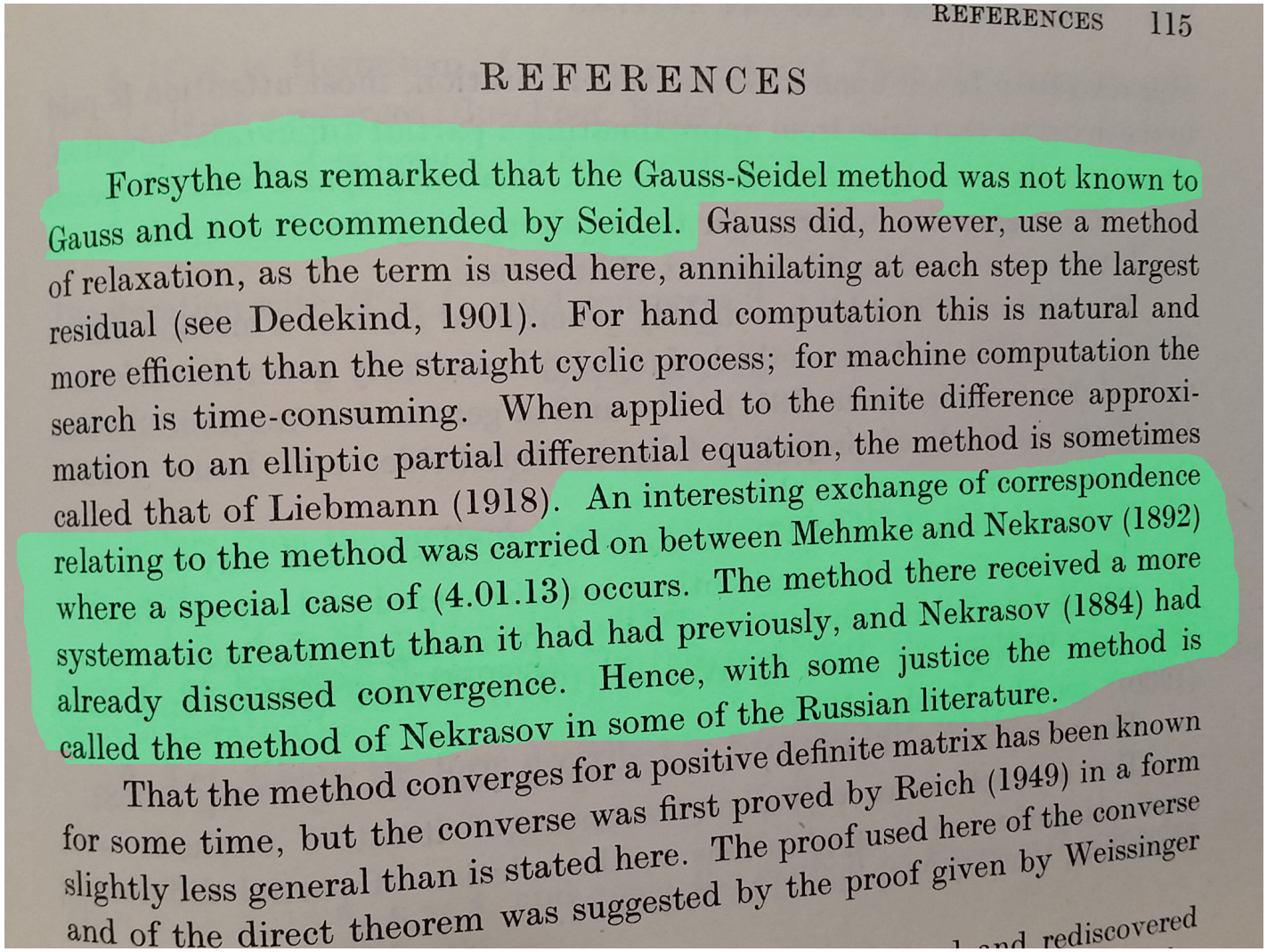
... Almost every evening I make a new edition of the tableau, wherever there is easy improvement. Against the monotony of the surveying business, this is always a pleasant entertainment; one can also see immediately whether anything doubtful has crept in, what still remains to be desired, etc. I recommend this method to you for imitation. You will hardly ever again eliminate directly, at least not when you have more than 2 unknowns.

Ends letter with:

... Almost every evening I make a new edition of the tableau, wherever there is easy improvement. Against the monotony of the surveying business, this is always a pleasant entertainment; one can also see immediately whether anything doubtful has crept in, what still remains to be desired, etc. I recommend this method to you for imitation. You will hardly ever again eliminate directly, at least not when you have more than 2 unknowns. ***The indirect procedure can be done while half asleep, or while thinking about other things.***

➤ Recommends this iterative scheme *(indirect elimination)* over Gaussian elimination for systems of order >2 (!!)

➤ We will contrast this with other recommendations later!



REFERENCES

Forsythe has remarked that the Gauss-Seidel method was not known to Gauss and not recommended by Seidel. Gauss did, however, use a method of relaxation, as the term is used here, annihilating at each step the largest residual (see Dedekind, 1901). For hand computation this is natural and more efficient than the straight cyclic process; for machine computation the search is time-consuming. When applied to the finite difference approximation to an elliptic partial differential equation, the method is sometimes called that of Liebmann (1918). An interesting exchange of correspondence relating to the method was carried on between Mehmke and Nekrasov (1892) where a special case of (4.01.13) occurs. The method there received a more systematic treatment than it had had previously, and Nekrasov (1884) had already discussed convergence. Hence, with some justice the method is called the method of Nekrasov in some of the Russian literature.

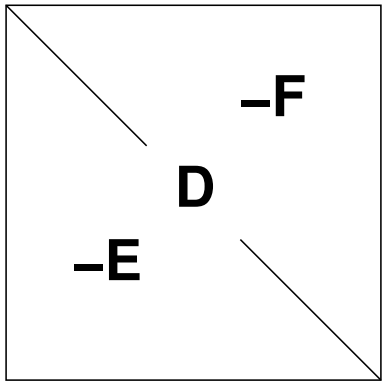
That the method converges for a positive definite matrix has been known for some time, but the converse was first proved by Reich (1949) in a form slightly less general than is stated here. The proof used here of the converse and of the direct theorem was suggested by the proof given by Weissinger

and rediscovered

- Alston Householder [1964]

Matrix form of relaxation methods

- Consider the decomposition $A = D - E - F$



D = diag(A), $-E$ = strict lower part of A and $-F$ its strict upper part.

- Gauss-Seidel amounts to: $(D - E)x^{(k+1)} = Fx^{(k)} + b$

[Decomposition used: $A = (D - E) - F$]

- Jacobi iteration: $Dx^{(k+1)} = (E + F)x^{(k)} + b$

[Decomposition used: $A = D - (E + F)$]

Over-relaxation (Young and others) is based on the decomposition:

$$\omega A = (D - \omega E) - (\omega F + (1 - \omega) D)$$

$\rightarrow$ successive overrelaxation, (SOR):

$$(D - \omega E) x^{(k+1)} = [\omega F + (1 - \omega) D] x^{(k)} + \omega b$$

- Up to the early 1980s, this was the state of the art
- Young-Frankel theory: Over-relaxation - Best ω (turned out to be > 1)...
- Two major books: R. Varga (1962) and D. Young (1971)

Young man: "Go Iterative"

SOLVING LINEAR ALGEBRAIC EQUATIONS CAN BE INTERESTING

GEORGE E. FORSYTHE[1]

1. **Introduction.** The subject of this talk is mathematically a lowly one. Consider a system of n linear algebraic equations in n unknowns, written

$$Ax = b. \tag{1}$$

Here A is a square matrix of order n, whose elements are given real numbers a_{ij} with a determinant $d(A) \neq 0$; x and b denote column vectors, and the components of b are given real numbers. (Complex numbers would offer no essential difficulty.) It is desired to calculate the components of the unique solution $x = A^{-1}b$; here A^{-1} is the inverse of A.

Such problems arise in the most diverse branches of science and technology, either directly (e.g., the normal equations of the least-squares adjustment of observations) or in an approximation to another problem (e.g., the difference-equation approximation to a self-adjoint boundary-value problem for a partial differential equation)

Bull. Amer. Math. Society, 59 (1953)

Great survey.. Original title "Solving Linear Equations is not trivial"

➤ Something new on the Horizon: Conjugate gradient.

It is my hope, on the other hand, to arouse the mathematician's interest by showing (sec. 2) the diversity of approaches to the solution of (1), and by mentioning (secs. 3 to 6) some problems associated with selected iterative methods. The newest process on the roster, the method of conjugate gradients, is outlined in sec. 7. Sec. 8 touches on the difficult general subject of errors and "condition," while a few words are hazarded in sec. 9 about the effect of machines on methods.

George Forsythe, 1953

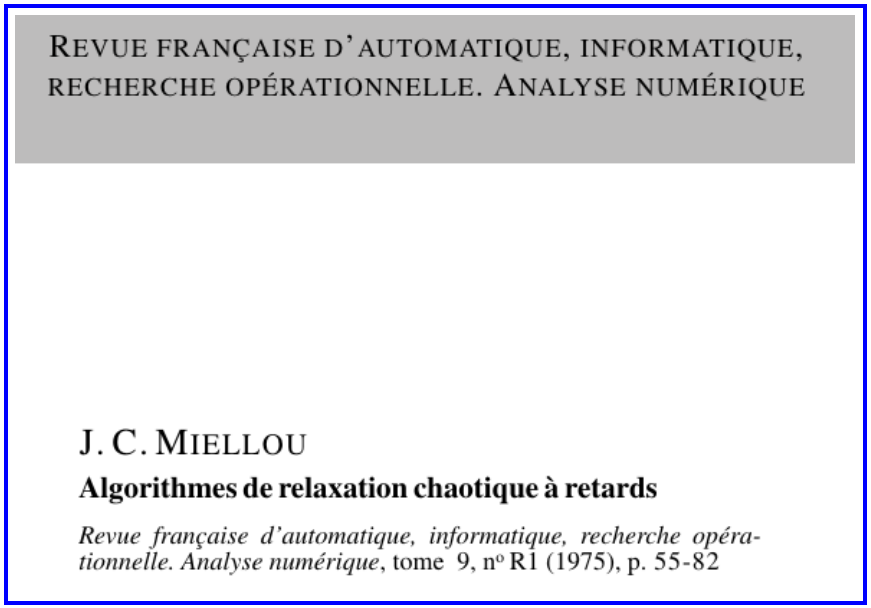
➤ George Forsthye (1917-1972) - Joined Stanford in 1957 (Math) - founded the CS dept. there (1965)

A word on chaotic iterations...

- Very early impact of the vision of parallel processing –
- Ostrowski [1955] “free steering” Schechter [1959] “block chaotic”
- Chazan-Miranker’s paper [1969] :

“The problem of chaotic relaxation was suggested to the authors by J. Rosenfeld who has conducted extensive numerical experiments with chaotic relaxation [J. Rosenfeld (1967)]. *Rosenfeld found that the use of more processors decreased the computation time for solving the linear system. (...) The chaotic form of the relaxation eliminated considerable programming and computer time in coordinating the processors and the algorithm. His experiments also exhibited a diminution in the amount of overrelaxation allowable for convergence in a chaotic mode."*

- Grenoble/Besançon school:
- J-C. Miellou [1975]
- Francois Robert [discrete iterations]
- P. Spiteri, P. Bahi, Charnay, Musy, ...,


REVUE FRANÇAISE D'AUTOMATIQUE, INFORMATIQUE, RECHERCHE OPÉRATIONNELLE. ANALYSE NUMÉRIQUE

J. C. MIELLOU

Algorithmes de relaxation chaotique à retards

Revue française d'automatique, informatique, recherche opérationnelle. Analyse numérique, tome 9, n° R1 (1975), p. 55-82

- Ideas were ahead of their time by a generation.
- Chaos is back: Dongarra et al [2012, 2015], D. Szyld & Frommer (2000), Garay, Magoules, Szyld, (2017), Elhadded, Garay, Magoules, Szyld, (2017), ..
- 'Chaotic' $\rightarrow$ 'Asynchronous' [sounds better]

Meanwhile on the opposite camp....

- Major advances on direct methods for sparse matrices.
- 1960s – Introduction of link between sparse matrices and graphs [Seymour Parter, 1961]
- Graphs played a major role thereafter
- Fill-ins and paths [Rose-Tarjan Theorem, 1975]
- Elimination Trees [YSMP group at Yale, late 1970s]
- Packages: YSMP, SPARSPAK.
- *Major* contribution: book by George and Liu (1981)

Huge progress of direct solvers

SIAM Review
Vol. 31, No. 1, pp. 1–19, March 1989

001

THE EVOLUTION OF THE MINIMUM DEGREE ORDERING ALGORITHM*

ALAN GEORGE† AND JOSEPH W.H. LIU‡

Abstract. Over the past fifteen years, the implementation of the minimum degree algorithm has received much study, and many important enhancements have been made to it. This paper describes these various enhancements, their historical development, and some experiments showing how very effective they are in improving the execution time of the algorithm. A shortcoming is also presented that exists in all of the widely used implementations of the algorithm, namely, that the quality of the ordering provided by the implementations is surprisingly sensitive to the initial ordering. For example, changing the input ordering can lead to an increase (or decrease) of as much as a factor of three in the cost of the subsequent numerical factorization. This sensitivity is caused by the lack of an effective tie-breaking strategy, and the authors' experiments illustrate the importance of developing such a strategy.

Key words. sparse matrices, minimum degree algorithm, computational complexity, ordering algorithms

- Traced progress from 1970s to late 1980s.

TABLE 1
Statistics on various versions of minimum degree algorithm.

Version	Minimum Degree Algorithm	Off-diagonal Factor Nonz	Ordering Time
Md#1	Final minimum degree	1,180,771	43.90
Md#2	Md#1 without multiple elimination	1,374,837	57.38
Md#3	Md#2 without element absorption	1,374,837	56.00
Md#4	Md#3 without incomplete deg update	1,374,837	83.26
Md#5	Md#4 without indistinguishable nodes	1,307,969	183.26
Md#6	Md#5 without mass elimination	1,307,969	2289.44

- To this day the two sides are in competition
- Some applications dominated by direct solvers...

... and others by iterative techniques.

Not always clear that iterative methods were worth a try:

- Bodewig [1956] says that they were:

  "... nearly always too slow (except when the matrix approaches a diagonal matrix), for most problems they do not converge at all, they cannot easily be mechanised and so they are more appropriate for computing by hand than for computing by machines, and they do not take advantage of the situation when the equations are symmetric".

- Only potential advantage he saw over direct methods:

  "Rounding errors do not accumulate, they are restricted to the last operation".

➤ David M. Young says (1st chapter of his book, 1971):

The use of direct methods even for solving very large problems has received increased attention recently (see for example Angel, 1970). In some cases their use is quite appropriate. However, there is the danger that if one does not properly apply iterative methods in some cases one will incorrectly conclude that they are not effective and that direct methods must be used. It is hoped that this book will provide some guidance (...)

● David M. Young Jr. [1923–2008]

A comparison from the other side (George & Liu's book)

1.3 Iterative versus direct methods *(...) The above remarks should make it clear that unless the question of which class of methods should be used is posed in a quite narrow and well defined context, it is either very complicated or impossible to answer.* [...gives reference to Varga and Young and says no books on direct solvers]

In addition, there are situations where it can be shown quite convincingly that direct methods are far more desirable than any conceivable iterative scheme

➤ Do not mention the relative ineffectiveness of direct solvers for large 3D problems [though it is clear it was known]

Nested dissection: cost for a regular mesh

- In 2-D consider an $n \times n$ problem, $N = n^2$
- In 3-D consider an $n \times n \times n$ problem, $N = n^3$

	2-D	3-D
space (fill)	$O(N \log N)$	$O(N^{4/3})$
time (flops)	$O(N^{3/2})$	$O(N^2)$

- Significant difference in complexity between 2-D and 3-D
- See George & Liu's book [1981]

- *Very* common misconception: *3-D problems are harder just because they are bigger.* In fact they are intrinsically harder.

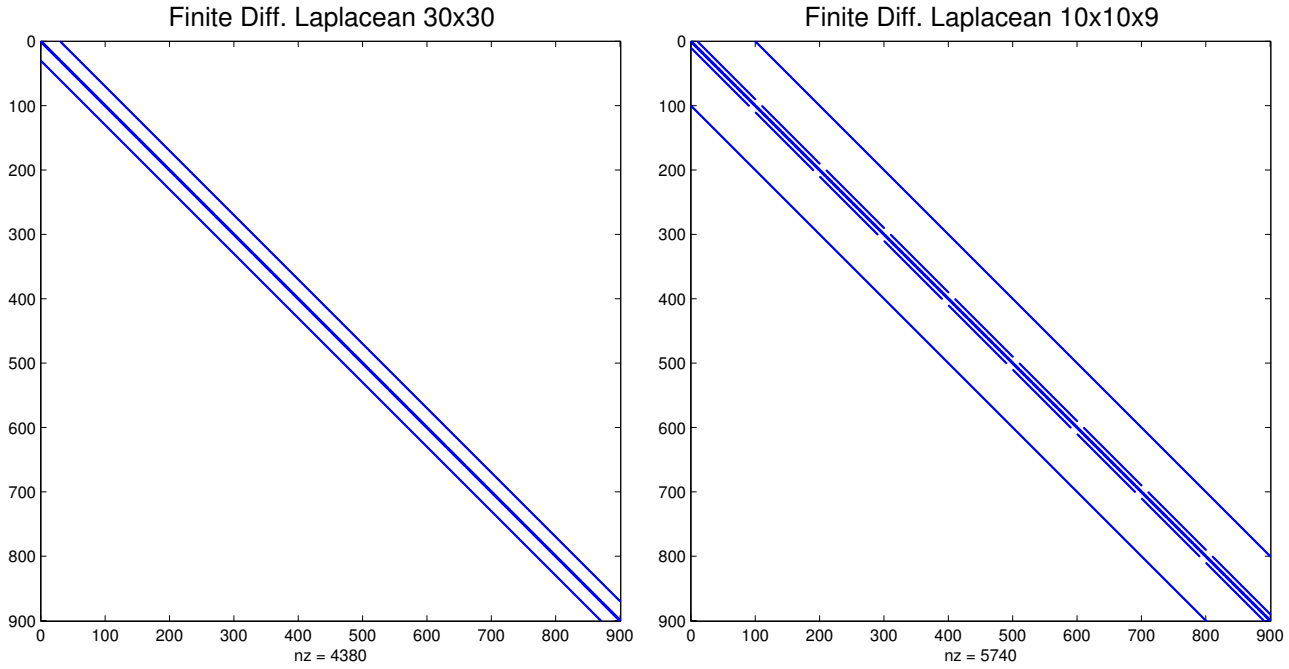
DEMO

Direct solution of two systems of size $N = 122,500$

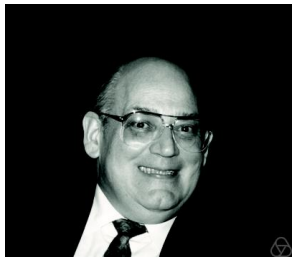
First: Laplacean on a 350×350 grid (2D);

Second: Laplacean on a $50 \times 50 \times 49$ grid (3D)

➤ Pattern of similar [much smaller] coefficient matrices.



As an example of the magnitude of problems that have been successfully solved on digital 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. adds a footnote: *... Even more staggering is Bettis' use of a 3-Dimens. program called "TNT-1", which treats coupled matrix equations of order 108,000.*



- State of the art in 1960 was a 20,000 $\times$ 20,000 Laplace equation.
- Could do this in seconds on a Laptop about 15 years ago.

One-dimensional projection processes

➤ Steepest descent: Cauchy [1847]. But Kantorovitch [1945] introduced it in the form we know today for linear systems for SPD matrices:

$$\boxed{\min_x \ \tfrac{1}{2} x^T A x - b^T x}$$

➤ Cimmino's method [1938] and Kaczmarz's method [1937] were also 'Line-search' type methods in the direction of a row^T or a column of A.

➤ All these techniques can be viewed as one-dimensional projection processes

One-dimensional projection processes

- Write: $\boxed{\tilde{x} := x + \alpha d}$
- Let $r = b - Ax, \tilde{r} = b - A\tilde{x}$.

Then write 'Petrov-Galerkin' condition: $\boxed{\tilde{r} \perp e \rightarrow r - Ad \perp e}$

$\rightarrow$ $\boxed{\alpha = \frac{(r,e)}{(Ad,e)}}$

- Steepest descent: $d = e = r$
- Minimal residual $d = r$, $e = Ar$
- Residual norm steepest descent $d = r$, $e = A^T r$ [SD for normal equations
- Kaczmarz method: $d = A^T e_i$, $e = e_i$ for $i = 1, \cdots, n$.
- Kaczmarz method equivalent to Gauss-Seidel for

$$\boxed{AA^T u = b \text{ with } (x = A^T u)}$$

- Very popular in the 70s for Computer Tomography [ART method]
- Cimmino's method == Jacobi method for $A^T Ax = A^T b$

Polynomial iteration

➤ One-dim projection methods and Richardson iterations are of the form

$$x_{k+1} = x_k + \beta_k d_k$$

➤ Frankel [1950] considers a 'second-order' iteration:

$$x_{k+1} = x_k + \beta_k d_k \text{ where } d_k = r_k - \alpha_k d_{k-1}$$

➤ With constant coefficients $\longrightarrow$ Chebyshev iteration.

➤ General form is similar to that of Conjugate Gradients

➤ Many papers adopted an 'approximation theory' viewpoint [Lanczos, '52]

Krylov methods take off: The CG algorithm

➤ Magnus Hestenes [UCLA] and Eduard Stiefel [ETH, Zürich] developed the method of Conjugate Gradient independently

M. Hestenes

Article:

Methods of conjugate gradients for solving linear systems, Nat. Bur. Standards, 1952.

E. Stiefel

➤ Lanczos developed a similar method [different notation and viewpoint:]
Solution of systems of linear equations by minimized iterations, Nat. Bur. Standards (1952)
C. Lanczos

➤ In effect a minimal residual (MR) method, implemented with the Lanczos procedure.

➤ Note: Same journal - Lanczos came out in July '52, Hestenes and Stiefel in Dec. '52.

➤ CG: Single most important advance in the 1950s - not too well received initially

- Viewed as an unstable, direct method.
- Engeli [1959]: viewed it as iterative it takes $2n$ or $3n$ to 'converge'
- ... until the early 1970s : paper by John Reid + analysis by Kaniel
- With preconditionining [Meijerink and van der Vorst, '77] the combination IC (Incomplete Cholesky) and CG became a de facto iterative solver for SPD case.

A mystery: why not other acceleration?

- Acceleration techniques: Richardson, Aitken, then
- Shanks, Wynn's ϵ- algorithm (Brezinski et al, Sidi, ...)

...were available but not used much by researchers in iterative methods

- Related methods often provided examples to developers of extrapolation algorithms [e.g., Wynn [1956] test with Gauss-Seidel]
- Forsythe mentions them briefly in his 1953 survey
- Recent work: C. Brezinski, M. Redivo-Zaglia tested several extrapolation algorithms applied to Kaczmarz iteration (To appear).

Krylov methods: the 'nonsymmetric' period

➤ Lanczos [MR paper 1952] : shows a method that is essentially the BiCG algorithm - then says : *let us restrict our attention to symmetric case ... (Normal eqns.)* A pity!

➤ Forward to 1976: Fletcher introduces BiCG – From here: CGS [Sonneveld, 1989] Bi-CGSTAB [van der Vorst, 1992], QMR, TFQMR [Freund, 1991], ...

Krylov methods: the 'nonsymmetric' period

➤ Lanczos [MR paper 1952] : shows a method that is essentially the BiCG algorithm - then says : *let us restrict our attention to symmetric case ... (Normal eqns.)* A pity!

➤ Forward to 1976: Fletcher introduces BiCG – From here: CGS [Sonneveld, 1989] Bi-CGSTAB [van der Vorst, 1992], QMR, TFQMR [Freund, 1991], ...

➤ "Orthogonal projection" track: ORTHOMIN [Vinsome, 1976], GCR [Axelsson, Vinsome, Eisenstat Elman & Schultz,] [ORTHODIR, ORTHORES, Young et al.], GMRES [YS-Schultz 1986], ...

➤ + Theory [Faber Manteuffel theorem, 1984], Convergence [Eisenstat-Elman-Schultz, 1983], Greenbaum-Strakos, ...

What is a projection method?

- Initial Problem: $b - Ax = 0$
- Given two subspaces K and L of $\mathbb{R}^N$ define the *approximate problem:*

Find $\tilde{x} \in K$ such that $b - A\tilde{x} \perp L$

- Leads to a small linear system ('projected problem')
- Typically: sequence of projection steps are applied
- With a nonzero initial guess x_0, the approximate problem is

Find $\tilde{x} \in x_0 + K$ such that $b - A\tilde{x} \perp L$

Two important particular cases.

1. $L = AK$. Then $\|b - A\tilde{x}\|_2 = \min_{z \in K} \|b - Az\|_2$
   $\rightarrow$ class of minimal residual methods: CR, GCR, ORTHOMIN, GMRES, CGNR, ...
2. $L = K$ $\rightarrow$ class of Galerkin or orthogonal projection methods. When A is SPD then $\|x^* - \tilde{x}\|_A = \min_{z \in K} \|x^* - z\|_A$

➤ CG and the Minimized Iteration method of Lanczos are projection methods on the Krylov subspace:

$$K_m = \mathbf{span}\{r_0, Ar_0, \cdots, A^{m-1}r_0\}$$

➤ Hence the link with polynomials (Lanczos)

Preconditioning

➤ Idea of preconditioning quite old – Golub & O'Leary (1989) trace term 'preconditioning' back to Turing [1948] – and Forsythe [1953] uses the term explicitly: *With the concept of "Ill conditioned" systems* $\boldsymbol{Ax} = \boldsymbol{b}$ *goes the idea of "preconditioning" them. Gauss [1823] and Jacobi [1845] made early contributions to this subject."*

➤ Wachpress seems first to use preconditioned CG - using ADI preconditioner [1963]

➤ Cesari [1937] also used polynomial preconditioning to speed-up Richardson iteration.

➤ Polynomial preconditioning suggested by Lanczos [1952] and Stiefel [1959] for CG/CG-like schemes.

The rise of ILU and ICCG

➤ Early work on incomplete factorizations:

* Buleev [1960], Oliphant [1962]

* Stone (SIP method) [1968], Dupont Kendall Rachford [1968]

* Axelsson [generalized SSOR method] [1972]

$$A = (\tilde{D} - L)\tilde{D}^{-1}(\tilde{D} - U)$$

* Similar idea by Dupont Kendall Rachford [1968]

➤ All these methods discussed splittings and worked on stencils for Finite Difference matrices...

➤ Then came IC / ILU [general sparse matrices]

Preconditioning: the impact of ICCG

MATHEMATICS OF COMPUTATION, VOLUME 31, NUMBER 137
JANUARY 1977, PAGES 148–162

An Iterative Solution Method for Linear Systems of Which the Coefficient Matrix is a Symmetric *M*-Matrix

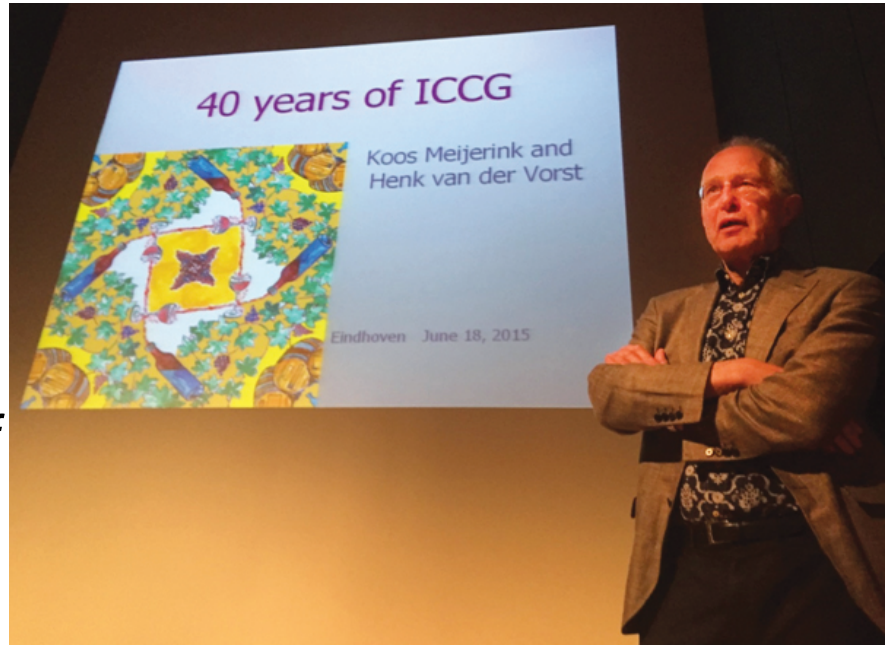
By J. A. Meijerink and H. A. van der Vorst

Abstract. A particular class of regular splittings of not necessarily symmetric *M*-matrices is proposed. If the matrix is symmetric, this splitting is combined with the conjugate-gradient method to provide a fast iterative solution algorithm. Comparisons have been made with other well-known methods. In all test problems the new combination was faster than the other methods.

- Incomplete Cholesky Conjugate Gradient (ICCG) paper appeared in 1977 (Math. Comp). Work done much earlier
- Paper had a tremendous impact in following years.

40 years of ICCG

The 'Preconditioning 2015' conference banquet featured a dinner speech by Koos Meijerink and Henk van der Vorst titled *"40 years of preconditioning"*.



* Time for a commercial: Preconditioning 2019 will be in Minneapolis July 1-3, 2019.

[see www.cs.umn.edu/~saad/Precon19]

Preconditioning

➤ Level-of-fill ILU developed (independently) by people from applications in reservoir simulations [Watts,..]

➤ Most successful preconditioners: those developed specifically for the application [from the PDE or physics – e.g., MG, AMG, DD methods, ...]

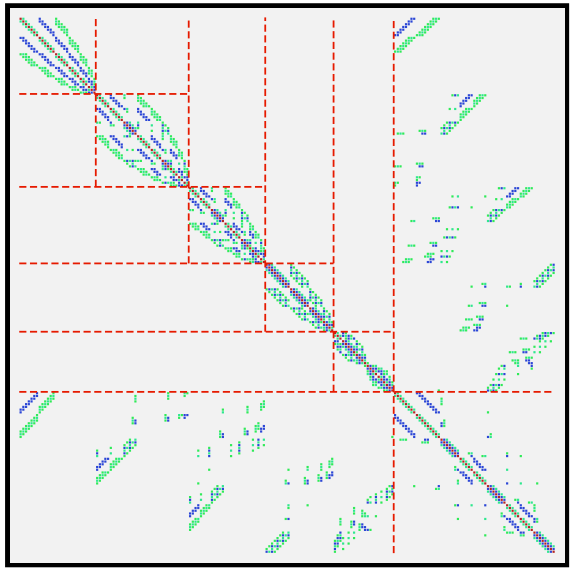
➤ For general sparse matrices: many ideas/methods [Approximate inverses, Benzi-Meyer, Grote-Huckle-..,] Multilevel ILU [Bank,YS-Zhang, ...]

- ‘(One sided) Row reordering: put largest entries on diagional [idea from sparse direct solution methods] Duff+ Koster [2001].
- Complex shifts: Erlangga Osterlee Vuik [2006], Beawens et al [2000?]

How does an industrial preconditioner look like today?

A: A mixture of

- AMG for solving Poisson-like components
- ILU for some other parts / fall-back option
- Domain Decompositon ideas,...
- GPU-specific parts



In a nutshell ...

Relaxation	Gauss (1820s) Jacobi, Seidel (1870s) Nekrasov (1890s), Dedekind (1900) Southwell (1940s), Ostrowski, Young (1950s), Varga (1950s), ...
Multigrid	Federenko (1964), Brandt (1970s)
CG/ CG-like	Hesteness-Stiefel (1952s), Lanczos (1952), Fletcher (1976), ..
Nonsymm. Krylov	Vinsome (orthomin, gcr, 1976), Young (1980s), Elman & al. (1980s), YS (1986, GMRES) Sonneveld (CGS, 1980), van der Vorst (Bicgstab, 1990s), QMR, TFQMR (Freund, 1990s)
Precond- itioning	Oliphant Buleev (1960s), DKR (1968), SIP (1967), SSOR-like (Evans, Axelsson) IC/ILU (van der Vorst 1977), ...

What has changed in past 4 decades?

- Research still active: Preconditioners for some types of problems [Helmholtz, Maxwell, Structures,...], parallel implementations,
- Less active: accelerators, ..
- Numerical analysis and Numerical Linear Algebra has is gradually disapperaring from computer science departments
- Topics no longer present in CS &Math. departments: sparse matrix techniques, sparse direct solvers, preconditioners, multigrid, ...
- *but* these may reappear in other areas.... e.g. computational statistics, ...,

What next?

➤ By and large, past solution techniques (iterative and direct) were aimed at solving PDEs

➤ Driven in part by demand: Aerospace & Automobile ind., ...

Q: What new demands show up at the horizon?

A: See titles at any meeting on 'computational X' and you will observe that in quite a few talks 'computational' is replaced by 'machine learning'

➤ Linear systems $\rightarrow$ matrix functions, approximating inverses, random sampling, online SVD, SVD updates, fast QR, ...

➤ Other ideas can be adapted .. For example Peaceman-Rachford's ADI [1955] gave rise to ADMM – a great success.

➤ Similarly, Steepest Descent gave rise to 'Stochatic Gradient Descent' (SGD) = a basic algorithm in Deep-learning.

➤ ... and stochastic variants of the Kaczmarz method are showing up [e.g. D Needell, R Ward, N Srebro - 2014, ...]

➤ Note: asynchronous iterations can be viewed as a form of *stochastic relaxaxion* – see analysis by J. C. Strikwerda [2002]

➤ Important new consideration: randomness

➤ In this context: CG - and related methods no longer work; Standard optimality (e.g. CG, GMRES) does not help.

Change should be welcome

➤ In the words of “Who moved my cheese?” [Spencer Johnson, 2002]:

> *If you do not change, you can become extinct !*
>
> *The quicker you let go of old cheese, the sooner you find new cheese*

Further reading: (historical surveys)

1 O. Axelsson. A survey of preconditionied iterative methods for linear systems of algebraic equations. *BIT*, 25:166–187, 1985.

2 G. E. Forsythe. Solving linear algebraic equations can be interesting. *Bull. Amer. Math. Soc.*, 59:299–329, 1953.

3 G. H. Golub and D. P. O'Leary. Some history of the conjugate gradient and Lanczos algorithms: 1948-1976. *SIAM review*, 31:50–102, 1989.

4 Y. Saad and H. A. van der Vorst. Iterative solution of linear systems in the 20th century. *Journal of Computational and Applied Mathematics*, 123:1–33, 2000.