On the Origins of Linear and Non-Linear Preconditioning

Martin J.Gander[1]

1 Linear Preconditioning

On December 26, 1823, Gauss sent a letter to his friend Gerling [10] to explain how he computed an approximate least squares solution based on angle measurements between the locations Berger Warte, Johannisberg, Taufstein and Milseburg. The system is symmetric, see Figure 1; it comes from the normal equations, and Gauss explains (translation by Forsythe [6]):

> "In order to eliminate indirectly, I note that, if 3 of the quantities a, b, c, d are set to 0, the fourth gets the largest value when d is chosen as the fourth. 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."

Die Bedingungsgleichungen sind also:

$$
\begin{aligned}
0 &= +\quad 6 + 67a - 13b - 28c - 26d \\
0 &= -\quad 7558 - 13a + 69b - 50c - 6d \\
0 &= -14604 - 28a - 50b + 156c - 78d \\
0 &= +22156 - 26a - 6b - 78c + 110d; \\
&\text{Summe} = 0.
\end{aligned}
$$

Um nun indirect zu eliminiren, bemerke ich, dass, wenn 3 der Grössen a, b, c, d gleich 0 gesetzt werden, die vierte den grössten Werth bekommt, wenn d dafür gewählt wird. Natürlich muss jede Grösse aus ihrer eigenen Gleichung, also d aus der vierten, bestimmt werden. Ich setze also $d = -201$ und substituire diesen Werth. Die absoluten Theile werden dann: $+5232$, -6352, $+1074$, $+46$; das Übrige bleibt dasselbe.

Fig. 1 Letter of Gauss from 1823 explaining what is now known as the Gauss-Seidel method.

[1] Université de Genève, Section de mathématiques, e-mail: martin.gander@unige.ch

With the new right hand side, Gauss then chooses again the variable to update which gives the largest value, and we recognize the well known Gauss-Seidel method, with the extra feature that at each step a particular variable is chosen to be updated, instead of just cycling through all the variables. Note also that the matrix is singular, but consistent (summing all equations gives zero, as indicated by Gauss' comment 'Summe=0' in Figure 1), and the method gives one particular solution. Gauss concludes his letter with the statement in Figure 2 (translation by Forsythe [6]):

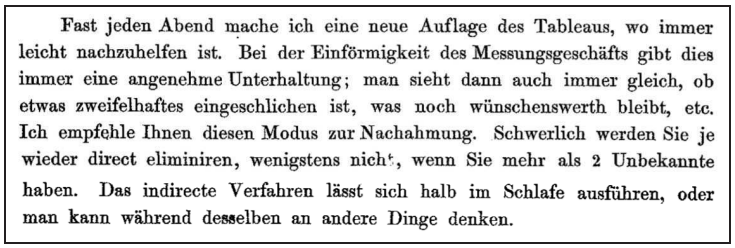

Fast jeden Abend mache ich eine neue Auflage des Tableaus, wo immer leicht nachzuhelfen ist. Bei der Einförmigkeit des Messungsgeschäfts gibt dies immer eine angenehme Unterhaltung; man sieht dann auch immer gleich, ob etwas zweifelhaftes eingeschlichen ist, was noch wünschenswerth bleibt, etc. Ich empfehle Ihnen diesen Modus zur Nachahmung. Schwerlich werden Sie je wieder direct eliminiren, wenigstens nicht, wenn Sie mehr als 2 Unbekannte haben. Das indirecte Verfahren lässt sich halb im Schlafe ausführen, oder man kann während desselben an andere Dinge denken.

Fig. 2 Gauss explains how relaxing these relaxations are.

> "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."

A general description of the method was then given by Seidel in [17], who also proved convergence of the method for the case of the normal equations, proposed to do the relaxations cyclically, and also to distribute them to two computers (humans) to do parallel computing[1].

In 1845, Jacobi presented in [12] the variant of Gauss' method now known as the Jacobi method, where one simultaneously relaxes all the variables. He acknowledges the computations that were performed by his friend Dr. Seidel. Realizing that the method can be slow or even fail if the system is not diagonally dominant enough, Jacobi then presents the groundbreaking idea of preconditioning using Jacobi rotations, see Figure 3:

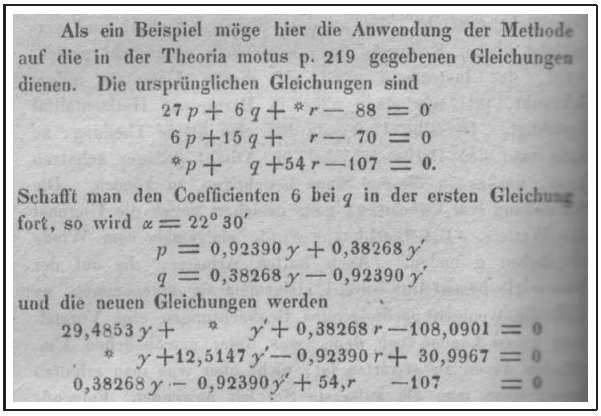
> "As an example we use the method for the equations from Theoria motus p. 219. The original equations are (see Figure 3). If we remove the coefficient 6 in front of q in the first equation, the angle of rotation is $\alpha = 22^0 30'$, and the new equations are..."

After preconditioning, it takes then only three Jacobi iterations to obtain three accurate digits!

In modern notation, a stationary iterative method for the linear system

$$A\mathbf{u} = \mathbf{f} \tag{1}$$

[1] "... sich unter zwei Rechner so vertheilen lässt ..."



Als ein Beispiel möge hier die Anwendung der Methode auf die in der Theoria motus p. 219 gegebenen Gleichungen dienen. Die ursprünglichen Gleichungen sind

$$\begin{aligned} 27p + 6q + *r - 88 &= 0 \\ 6p + 15q + r - 70 &= 0 \\ *p + q + 54r - 107 &= 0. \end{aligned}$$

Schafft man den Coefficienten 6 bei q in der ersten Gleichung fort, so wird $\alpha = 22^\circ 30'$

$$\begin{aligned} p &= 0{,}92390\,y + 0{,}38268\,y' \\ q &= 0{,}38268\,y - 0{,}92390\,y' \end{aligned}$$

und die neuen Gleichungen werden

$$\begin{aligned} 29{,}4853\,y + * y' + 0{,}38268\,r - 108{,}0901 &= 0 \\ * y + 12{,}5147\,y' - 0{,}92390\,r + 30{,}9967 &= 0 \\ 0{,}38268\,y - 0{,}92390\,y' + 54{,}r - 107 &= 0 \end{aligned}$$

Fig. 3 Jacobi's idea of preconditioning the linear system using Jacobi rotations.

is obtained from a splitting of the matrix $A = M - N$, followed by the iteration

$$M\mathbf{u}^{n+1} = N\mathbf{u}^n + \mathbf{f}. \tag{2}$$

For Jacobi, we would have $M = \text{diag}(A)$, for Gauss-Seidel $M = \text{tril}(A)$, a Schwarz domain decomposition method with minimal overlap would have M block diagonal, and for multigrid, M represents a V-cycle or W-cycle. Rewriting the stationary iterative method (2) as

$$\mathbf{u}^{n+1} = M^{-1}N\mathbf{u}^n + M^{-1}\mathbf{f} = (I - M^{-1}A)\mathbf{u}^n + M^{-1}\mathbf{f},$$

we see that the method converges fast if the spectral radius $\rho(I - M^{-1}A)$ is small, and it is cheap, if systems with M can easily be solved.

In 1951, Stiefel and Rosser[2] gave both a presentation at a symposium on simultaneous linear equations and the determination of eigenvalues at the National Bureau of Standards (UCLA), and realized that they presented the same method. The method of Forsythe, Hestenes and Rosser appeared in a short note in [7], and the method of Stiefel in a comprehensive and elegant exposition on iterative methods in [18]. Hestenes, who was also present at the symposium, and Stiefel then wrote together during Stiefel's stay at the National Bureau of Standards the famous 1952 conjugate gradient paper [11][3]. Independently in 1952, Lanczos had also invented essentially the same method [15], based on his earlier work on eigenvalues problems [14], where he already pointed out that solving linear systems with this method was just a special case.

[2] Rosser was working with Forsythe and Hestenes at that time

[3] "An iterative algorithm is given for solving a system $Ax = k$ of n linear equations in n unknowns. The solution is given in n steps."

So what is this famous conjugate gradient (CG) method ? To solve approximately $A\mathbf{u} = \mathbf{f}$, A symmetric and positive definite, CG finds at step n using the Krylov space[4]

$$\mathscr{K}_n(A,\mathbf{r}^0) := \{\mathbf{r}^0, A\mathbf{r}^0, \ldots, A^{n-1}\mathbf{r}^0\}, \quad \mathbf{r}^0 := \mathbf{f} - A\mathbf{u}^0$$

an approximate solution $\mathbf{u}^n \in \mathbf{u}^0 + \mathscr{K}_n(A,\mathbf{r}^0)$ which satisfies

$$||\mathbf{u} - \mathbf{u}^n||_A \longrightarrow \min, \qquad ||\mathbf{u}||_A^2 := \mathbf{u}^T A\mathbf{u}.$$

Using Chebyshev polynomials, one can prove the following convergence estimate for CG:

Theorem 1. *With $\kappa(A) := \frac{\lambda_{\max}(A)}{\lambda_{\min}(A)}$ the condition number of A, the iterate $\mathbf{u}^n$ of CG satisfies the convergence estimate*

$$||\mathbf{u} - \mathbf{u}^n||_A \leq 2\left(\frac{\sqrt{\kappa(A)}-1}{\sqrt{\kappa(A)}+1}\right)^n ||\mathbf{u} - \mathbf{u}^0||_A.$$

We see that the conjugate gradient method converges very fast, if the condition number $\kappa(A)$ is not very large.

The success of CG motivated researchers to design similar methods searching in a Krylov space for solutions when the system matrix is not symmetric and positive definite. There are two classes of such methods: the first class are the Minimum Residual methods (MR) which search for $\mathbf{u}^n \in \mathbf{u}_0 + \mathscr{K}_n(A,\mathbf{r}^0)$ such that

$$||\mathbf{f} - A\mathbf{u}^n||_2 \longrightarrow \min.$$

MINRES (Paige, Saunders 1975) is such an algorithm, designed for symmetric systems which are not positive definite. GMRES (Saad, Schultz 1986) does the same for arbitrary systems, and QMR (Freund, Nachtigal 1991) tries to solve the minimization problem approximately. The second class of methods is based on orthogonalization (OR): they search for $\mathbf{u}^n \in \mathbf{u}_0 + \mathscr{K}_n(A,\mathbf{r}^0)$ such that

$$\mathbf{f} - A\mathbf{u}^n \perp \mathscr{K}_n(A,\mathbf{r}^0).$$

SymmLQ (Paige, Saunders 1975) does this for symmetric indefinite systems, FOM (Saad 1981) for general systems, and BiCGstab (Van Der Vorst 1992) does it approximately. All these methods converge well, if the spectrum of the matrix A is clustered around 1 provided the matrices are normal ($AA^T = A^TA$).

If the spectrum of A is not clustered around 1, the old idea of Jacobi can be used: find a preconditioner, a matrix M, such that the preconditioned system

$$M^{-1}A\mathbf{u} = M^{-1}\mathbf{f}$$

[4] The name is going back to Krylov [13] studying the solution of systems of second order ordinary differential equations, and the now called Krylov space only appears implicitly there

has a spectrum which clusters much better around 1 than the spectrum of the matrix A itself. For CG, using Theorem 1 one can even say more specifically that M should make the condition number $\kappa(M^{-1}A)$ much smaller than $\kappa(A)$. In all cases however it should be inexpensive to apply M^{-1}.

It is sometimes possible to directly design preconditioners with good properties: excellent examples in domain decomposition are the additive Schwarz method (Dryja and Widlund 1987), FETI (Farhat and Roux 1991) and Balancing Domain Decomposition (Mandel and Brezina 1993), but it takes a lot of experience and intuition to do so.

A systematic approach for constructing preconditioners is to recall what we have seen for stationary iterative methods: we needed M such that the spectral radius $\rho(I-M^{-1}A)$ is small, and it is inexpensive to apply M^{-1}. The last point is identical with preconditioning, and note that

$$\rho(I-M^{-1}A) \text{ small} \quad \Longleftrightarrow \quad \text{the spectrum of } M^{-1}A \text{ is close to one!}$$

It is therefore natural to first design a good M for a stationary iterative method, and then use it as a preconditioner for a Krylov method.

Theorem 2. *Using an MR Krylov method with preconditioner M never gives worse (and usually much better) residual reduction than just using the stationary iteration.*

Proof. The stationary iterative method computes

$$\mathbf{u}^n = (I-M^{-1}A)\mathbf{u}^{n-1} + M^{-1}\mathbf{f} = \mathbf{u}^{n-1} + \mathbf{r}_{stat}^{n-1},$$

where we introduced $\mathbf{r}_{stat}^n := M^{-1}\mathbf{f} - M^{-1}A\mathbf{u}^n$. Multiplying this equation by $-M^{-1}A$ and adding $M^{-1}\mathbf{f}$ on both sides then gives

$$\mathbf{r}_{stat}^n = (I-M^{-1}A)\mathbf{r}_{stat}^{n-1} = (I-M^{-1}A)^n\mathbf{r}^0. \tag{3}$$

The preconditioned Krylov method will use the Krylov space

$$\mathscr{K}_n(M^{-1}A,\mathbf{r}^0) := \{\mathbf{r}^0, M^{-1}A\mathbf{r}^0, \ldots, (M^{-1}A)^{n-1}\mathbf{r}^0\}$$

to search for $\mathbf{u}^n \in \mathbf{u}^0 + \mathscr{K}_n(M^{-1}A,\mathbf{r}^0)$, i.e. it will determine coefficients α_i s.t.

$$\mathbf{u}^n = \mathbf{u}^0 + \sum_{i=1}^{n} \alpha_i (M^{-1}A)^{i-1}\mathbf{r}^0.$$

Multiplying this equation by $-M^{-1}A$ and adding $M^{-1}\mathbf{f}$ on both sides then gives

$$\mathbf{r}_{kry}^n = p^n(M^{-1}A)\mathbf{r}^0, \tag{4}$$

p^n a polynomial of degree n with $p^n(0)=1$. Since the MR Krylov method finds the polynomial which minimizes the residual in norm, it is at least as good as the specific polynomial $(I-M^{-1}A)^n$ chosen by the stationary iterative method in (3).

The classical alternating and parallel Schwarz methods are such stationary iterative methods, and also RAS [3] and optimized Schwarz methods [8], and the Dirichlet-Neumann and Neumann-Neumann methods [16]. They all are convergent as stationary iterative methods, while for example additive Schwarz is not [5, 9].

2 Non-Linear Preconditioning

In contrast to linear preconditioning, non-linear preconditioning is a much less explored area of research. In the context of domain decomposition, a seminal contribution for non-linear preconditioning was made by Cai, Keyes and Young at DD13 [2], namely the Additive Schwarz Preconditioned Inexact Newton method (ASPIN), see also Cai and Keyes [1]. The idea is:

> "The nonlinear system is transformed into a new nonlinear system, which has the same solution as the original system. For certain applications the nonlinearities of the new function are more balanced and, as a result, the inexact Newton method converges more rapidly."

Instead of solving $F(\mathbf{u}) = \mathbf{0}$, one solves instead $G(F(\mathbf{u})) = \mathbf{0}$ where according to the authors the function G should have the properties: 1) if $G(\mathbf{v}) = \mathbf{0}$ then $\mathbf{v} = \mathbf{0}$, 2) $G \approx F^{-1}$ in some sense, 3) $G(F(\mathbf{v}))$ is easy to compute, and 4) applying Newton, $(G(F(\mathbf{v})))'\mathbf{w}$ should also be easy to compute. The authors then define the ASPIN preconditioner as follows: for $F : \mathbb{R}^m \to \mathbb{R}^m$, define J (overlapping) subsets Ω_j for the indices $\{1,2,\ldots,m\}$, such that $\bigcup_j \Omega_j = \{1,2,\ldots,m\}$, and corresponding restriction matrices R_j, e.g. $\Omega_1 = \{1,2,3\} \implies R_1 = [I\ 0]_{3\times m}$, I the 3×3 identity matrix. Define the solution operator $T_j : \mathbb{R}^m \to \mathbb{R}^{|\Omega_j|}$ such that

$$R_j F(\mathbf{v} - R_j^T T_j(\mathbf{v})) = 0. \tag{5}$$

Then ASPIN solves using inexact Newton

$$\sum_{j=1}^{J} R_j^T T_j(\mathbf{u}) = 0. \tag{6}$$

It is not easy to understand where this transformation comes from[5]. Let us first look at a fixed point iteration like Gauss-Seidel or Jacobi for this nonlinear problem. If we denote the unknowns corresponding to the subsets Ω_j by $\mathbf{u}_j$, the corresponding block Jacobi fixed point iteration would be to solve for $n = 0,1,2,\ldots$

$$\begin{array}{l} F_1(\mathbf{u}_1^{n+1},\mathbf{u}_2^n,\ldots,\mathbf{u}_J^n) = 0 \\ F_2(\mathbf{u}_1^n,\mathbf{u}_2^{n+1},\ldots,\mathbf{u}_J^n) = 0 \\ \qquad\qquad \vdots \\ F_J(\mathbf{u}_1^n,\mathbf{u}_2^n,\ldots,\mathbf{u}_J^{n+1}) = 0 \end{array} \quad \Longrightarrow \quad \begin{array}{l} \mathbf{u}_1^{n+1} = G_1(\mathbf{u}_2^n,\ldots,\mathbf{u}_J^n) \\ \mathbf{u}_2^{n+1} = G_2(\mathbf{u}_1^n,\mathbf{u}_3^n,\ldots,\mathbf{u}_J^n) \\ \qquad\qquad \vdots \\ \mathbf{u}_J^{n+1} = G_J(\mathbf{u}_1^n,\mathbf{u}_2^n,\ldots,\mathbf{u}_{J-1}^n) \end{array} \tag{7}$$

[5] "ASPIN may look a bit complicated ..." (Cai, Keyes 2002).

where we denoted the solutions of the non-linear equation F_j by G_j. At the fixed point, which solves $F(\mathbf{u}) = 0$, we must have $\mathbf{u} = G(\mathbf{u})$, and thus instead of solving $F(\mathbf{u}) = 0$ using Newton's method, one can instead solve $\mathbf{u} - G(\mathbf{u}) = 0$ using Newton's method. This gives us a very general idea of non-linear preconditioning: one first designs a fixed point iteration (like the stationary iterative method in the linear case); but then one does not use this method directly, one applies Newton's method to the equation at the fixed point (like one applies a Krylov method to the fixed point of the stationary iterative method).

Theorem 3. *ASPIN in the case of no algebraic overlap (which means minimal geometric overlap of one mesh size) is identical to solving with an inexact Newton method the non-linear block Jacobi iteration equations at the fixed point.*

Proof. The definition of the solution operator in (5) shows that we can use it to replace G_j in (7), namely

$$\mathbf{u}_j^{n+1} = R_j \mathbf{u}^n - T_j(\mathbf{u}^n).$$

Now in the case of no algebraic overlap (minimal geometric overlap), the sum in (6) just composes the operators T_j in a large vector, there is never actually a sum computed, and thus (6) represents precisely (7) at the fixed point, i.e.

$$0 = \mathbf{u} - G(\mathbf{u}) = \mathbf{u} - \sum_{j=1}^{J} R_j^T (R_j \mathbf{u} - T_j(\mathbf{u})) = \sum_{j=1}^{J} R_j^T T_j(\mathbf{u}),$$

where we used that $\mathbf{u} - \sum_{j=1}^{J} R_j^T R_j \mathbf{u} = 0$ in the case of zero algebraic overlap.

Remark 1. In the case of more overlap, ASPIN has the same problem as the additive Schwarz method in the overlap, it is inconsistent and can only be used as a preconditioner [5, 9], where a Krylov method must correct this inconsistency. In the case of ASPIN, Newton must to the same; ASPIN then does not correspond to a consistent fixed point iteration in the case of more than minimal overlap.

3 Conclusion

We have explained how first stationary iterative methods were invented for linear systems of equations by Gauss and Jacobi, and how Jacobi had already the idea of preconditioning in 1845. With the invention of Krylov methods, stationary iterations have lost their importance as solvers, but good splittings from stationary iterative methods found great use as preconditioners for Krylov methods. In the case of non-linear problems, one can follow the same principle: one first conceives a fixed point iteration for the non-linear problem, like a non-linear iterative domain decomposition method, or the full approximation scheme from multigrid. One then however does not use this fixed point iteration as a solver, one solves instead the

equations at the fixed point: *this is the meaning of non-linear preconditioning*. This observation allowed the authors in [4] to devise a new non-linear preconditioner called RASPEN, which avoids the problem ASPIN has in the overlap, and also introduces the coarse grid correction in a consistent way by using the full approximation scheme from multigrid. It is also shown in [4] that one can actually use the exact Jacobian, since the non-linear subdomain solvers provide this information already, and extensive numerical experiments in [4] show that RASPEN performs significantly better as non-linear preconditioner than ASPIN.

References

1. X.-C. Cai and D. E. Keyes. Nonlinearly preconditioned inexact Newton algorithms. *SIAM Journal on Scientific Computing*, 24(1):183–200, 2002.
2. X.-C. Cai, D. E. Keyes, and D. P. Young. A nonlinear additive Schwarz preconditioned inexact Newton method for shocked duct flow. In *Proceedings of the 13th International Conference on Domain Decomposition Methods*, pages 343–350. DDM.org, 2001.
3. X.-C. Cai and M. Sarkis. A restricted additive Schwarz preconditioner for general sparse linear systems. *SIAM journal on scientific computing*, 21(2):792–797, 1999.
4. V. Dolean, M. J. Gander, W. Kheriji, F. Kwok, and R. Masson. Nonlinear preconditioning: How to use a nonlinear Schwarz method to precondition Newton's method. *submitted*, 2016.
5. E. Efstathiou and M. J. Gander. Why Restricted Additive Schwarz converges faster than Additive Schwarz. *BIT Numerical Mathematics*, 43(5):945–959, 2003.
6. G. E. Forsythe. Notes. *Mathematical Tables and Other Aids to Computation*, 5(36):255–258, 1951.
7. G. E. Forsythe, M. R. Hestenes, and J. B. Rosser. Iterative methods for solving linear equations. In *Bulletin of the American Mathematical Society*, volume 57(6), pages 480–480, 1951.
8. M. J. Gander. Optimized Schwarz methods. *SIAM Journal on Numerical Analysis*, 44(2):699–731, 2006.
9. M. J. Gander. Schwarz methods over the course of time. *Electron. Trans. Numer. Anal*, 31(5):228–255, 2008.
10. C. F. Gauss. Letter to Gerling, December 26, 1823. In *Werke*, volume 9, pages 278–281. Göttingen, 1903.
11. M. R. Hestenes and E. Stiefel. *Methods of conjugate gradients for solving linear systems*, volume 49. NBS, 1952.
12. C. G. J. Jacobi. Ueber eine neue Auflösungsart der bei der Methode der kleinsten Quadrate vorkommenden lineären Gleichungen. *Astronomische Nachrichten*, 22(20):297–306, 1845.
13. A. N. Krylov. On the numerical solution of the equation by which in technical questions frequencies of small oscillations of material systems are determined. *Izvestija AN SSSR (News of Academy of Sciences of the USSR), Otdel. mat. i estest. nauk*, 7(4):491–539, 1931.
14. C. Lanczos. *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*. United States Governm. Press Office Los Angeles, CA, 1950.
15. C. Lanczos. Solution of systems of linear equations by minimized iterations. *J. Res. Nat. Bur. Standards*, 49(1):33–53, 1952.
16. A. Quarteroni and A. Valli. *Domain Decomposition Methods for Partial Differential Equations*. Oxford Science Publications, 1999.
17. L. Seidel. Über ein Verfahren, die Gleichungen, auf welche die Methode der kleinsten Quadrate führt, sowie lineäre Gleichungen überhaupt, durch successive Annäherung aufzulösen. In *Abhandlungen der Mathematisch-Physikalischen Klasse der Königlich Bayerischen Akademie der Wissenschaften, Band 11, III. Abtheilung*, pages 81–108. 1874.
18. E. Stiefel. Über einige Methoden der Relaxationsrechnung. *Zeitschrift für angewandte Mathematik und Physik ZAMP*, 3(1):1–33, 1952.