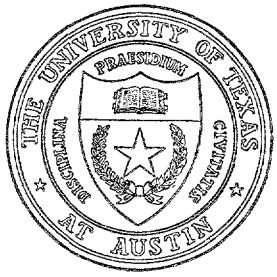
Commentaries of Three Papers of Cornelius Lanczos

by

Kang C. Jea and David M. Young

October 1991 CNA-252



CENTER FOR NUMERICAL ANALYSIS

THE UNIVERSITY OF TEXAS AT AUSTIN

DISCLAIMER

DISCLAIMER

Portions of this document may be illegible in electronic image products. Images are produced from the best available original document.

Commentaries on Three Papers by Cornelius Lanczos

by

Kang C. Jea* and David M. Young

This report contains commentaries on three papers of Cornelius Lanczos listed below. These commentaries will be included in a volume of the collected published Lanczos papers which will be published as part of the Cornelius Lanczos Centenary Celebration at North Carolina State University. The volume is scheduled for publication in December 1993.

A. Cornelius Lanczos [1952]. "Iterative Solution of Systems of Linear Equations by Minimized Iterations", *Journal of Research of the National Bureau of Standards*, **49**, 33-53.

B. Cornelius Lanczos [1953]. "Chebyshev Polynomials in the Solution of Large-Scale Linear Systems", Proceedings of the ACM Conference held in Toronto, California in 1952, Sauls L. Lithograph Co., Washington, DC.

C. Cornelius Lanczos [1958]. "Iterative Solutions of Large-Scale Linear Systems", *J. Soc. Indust. Appl. Math.*, **6**, 91-109.

* Now at the Department of Mathematics, Fu Jen University, Taipei, Taiwan, R.O.C.

Commentary on Lanczos [1952]

by

Kang C. Jea and David M. Young

Commentary on Lanczos [1952]
"Solution of Systems of Linear Equations by Minimized Iterations"

Lanczos [1952] considered the problem of solving the linear system+

$$Au = b \tag{1}$$

where A is a given nonsingular NxN complex matrix and b is a given complex Nx1 column matrix. He considered a method which involves choosing an arbitrary vector $\tilde{b}$ and, for k=1,2,..., the generation of the coefficients $a_0, a_1, \ldots, a_{k-1}$ and $\hat{a}_1, \hat{a}_2, \ldots \hat{a}_k$ to minimize the Euclidean lengths of $p^{(k)}$ and $q^{(k)*}$ where

$$\begin{aligned} p^{(k)} &= [A^k - (a_0 I + a_1 A + \ldots + a_{k-1} A^{k-1})]b \\ \tilde{q}^{(k)*} &= [I - (\hat{a}_1 A + \hat{a}_2 A^2 + \ldots + \hat{a}_k A^k)]b \end{aligned} \tag{2}$$

Lanczos also defined two other vectors $\tilde{p}^{(k)}$ and $\tilde{q}^{(k)*}$ are given by[2]

$$\begin{aligned} \tilde{p}^{(k)} &= [(A^*)^k - (a_0 I + a_1 A^* + \ldots + a_{k-1}(A^*)^{k-1})]\tilde{b} \\ \tilde{q}^{(k)*} &= [I - (\hat{a}_1 A^* + \hat{a}_2 (A^*)^2 + \ldots + \hat{a}_k (A^*)^k)]\tilde{b} \end{aligned} \tag{2a}$$

Evidently we have

$$\begin{aligned} p^{(k)} &= p_k(A)b \\ \tilde{p}^{(k)} &= p_k(A^*)\tilde{b} \end{aligned} \tag{3}$$

where the polynomials $p_k(x)$ are given by

$$p_k(x) = x^k - (a_0 + a_1 x + \ldots + a_{k-1} x^{k-1}) \tag{4}$$

+ Here and elsewhere we have changed the notation slightly.

[2] Here A* denotes the conjugate transpose of A. However, p* denotes a column vector which is, in general, different from p.

Similarly, the vectors $q^{(k)*}$ and $\tilde{q}^{(k)*}$ are given by

$$q^{(k)*} = q_k^*(A)b$$
$$\tilde{q}^{(k)*} = q_k^*(A^*)\tilde{b} \tag{5}$$

where the polynomials $q_k^*(x)$ are given by

$$q_k^*(x) = 1-(\hat{a}_1 x + \hat{a}_2 x^2 + \ldots + \hat{a}_k x^k) \tag{6}$$

Finally, we define the vectors $q^{(k)}$ and $\tilde{q}^{(k)}$ by

$$q^{(k)} = -\frac{1}{\hat{a}_k} q^{(k)*} = q_k(A)b$$
$$\tilde{q}^{(k)} = -\frac{1}{\hat{a}_k} \tilde{q}^{(k)*} = q_k(A^*)\tilde{b} \tag{7}$$

where the polynomials $q_k(x)$ are given by

$$q_k(x) = -\frac{1}{\hat{a}_k} q_k^*(x) = -\frac{1}{\hat{a}_k}[1-(\hat{a}_1 x + \ldots + \hat{a}_k x^k)] \tag{8}$$

Using the polynomial operators $q_k(A)$ we can generate a sequence of approximate solutions to (1). Thus, we let

$$u^{(k-1)} = -\frac{1}{q_k(0)} A^{-1}(q_k(A) - q_k(E))b \qquad ,k = 1,2,\ldots \tag{9}$$

where E is the null matrix (We note that $q_k(A) - q_k(E)$ contains a factor A). For each k the residual $r^{(k-1)}$ corresponding to $u^{(k-1)}$ is given by

$$r^{(k-1)} = b - Au^{(k-1)} = q^{(k)*} \tag{10}$$

It can be shown that the length of $r^{(k-1)}$ is minimum. An alternative sequence of approximate solutions to (1) can be generated by

$$\tilde{u}^{(k-1)} = -\frac{1}{p_k(0)} A^{-1}(p_k(A) - p_k(E))b \qquad ,k = 1,2,\ldots \tag{11}$$

The corresponding residuals $\tilde{r}^{(k-1)}$ are given by

$$\tilde{r}^{(k-1)} = b - A\tilde{u}^{(k-1)} = -\frac{1}{a_0} p^{(k)} \tag{12}$$

Thus $\tilde{r}^{(k-1)}$ has in general a larger length than $r^{(k-1)}$ except at the end of the process when $\tilde{r}^{(k-1)} = r^{(k-1)} = 0$, and $\tilde{u}^{(k-1)} = u^{(k-1)} = A^{-1}b$, the exact solution.

The generation of the two sets of coefficients for the polynomials $p_k(x)$ and $q_k(x)$ can be shown to be closely related and to lead to two-term recurrence relations for the corresponding polynomials $p_k(x)$ and $q_k(x)$. Thus we have

$$\begin{aligned} p_{k+1}(x) &= \rho_k p_k(x) + x q_k(x) \\ q_{k+1}(x) &= \sigma_k q_k(x) + p_{k+1}(x) \end{aligned} \tag{13}$$

where

$$\rho_k = -\frac{(q_k(A)b, (A^*)^{k+1}\tilde{b})}{(p_k(A)b, (A^*)^{k+1}\tilde{b})} = -\frac{(q_k(A^*)\tilde{b}, A^{k+1}b)}{(p_k(A^*)\tilde{b}, A^k b)}$$

and

$$\sigma_k = -\frac{(p_{k+1}(A)b, (A^*)^{k+1}\tilde{b})}{(q_k(A)b, (A^*)^{k+1}\tilde{b})} = -\frac{(p_{k+1}(A^*)\tilde{b}, A^{k+1}b)}{(q_k(A^*)\tilde{b}, A^{k+1}b)}$$

Evidently, the vectors $p^{(k)}, \tilde{p}^{(k)}, q^{(k)}$ and $\tilde{q}^{(k)}$ can be generated by using (13). That is

$$\begin{aligned} p^{(k+1)} &= \rho_k p^{(k)} + A q^{(k)}; & \tilde{p}^{(k+1)} &= \rho_k \tilde{p}^{(k)} + A^* \tilde{q}^{(k)} \\ q^{(k+1)} &= \sigma_k q^{(k)} + p^{(k+1)}; & \tilde{q}^{(k+1)} &= \sigma_k \tilde{q}^{(k)} + \tilde{p}^{(k+1)} \end{aligned} \tag{14}$$

where

$$\rho_k = -\frac{(\tilde{p}^{(k)}, Aq^{(k)})}{(p^{(k)}, \tilde{p}^{(k)})} = -\frac{(p^{(k)}, A^*\tilde{q}^{(k)})}{(p^{(k)}, \tilde{p}^{(k)})}$$

$$\sigma_k = -\frac{(Ap^{(k+1)}, \tilde{p}^{(k)})}{(Aq^{(k)}, \tilde{p}^{(k)})} = -\frac{(p^{(k+1)}, \tilde{p}^{(k+1)})}{(Aq^{(k)}, \tilde{p}^{(k)})} \qquad (14')$$

Moreover, it can be shown that the approximate solution $u^{(k+1)}$, given by (9) can be written in the form

$$u^{(k+1)} = \sum_{i=0}^{k} \eta_i q^{(i)}$$

$$= u^{(k)} + \eta_k q^{(k)}, k = 0,1,2,.... \qquad (15)$$

where

$$\begin{cases} \eta_0 = -1/\rho_0 \\ \eta_k = \eta_{k-1}/\rho_k \quad , \quad k = 1,2,.... \end{cases}$$

Futhermore, it can be shown that the $p^{(k)}$ and $\tilde{p}^{(k)}$ are biorthogonal sets of vectors and that the $q^{(k)}$ and $\tilde{q}^{(k)}$ are bi-conjugate sets of vectors in the sense that

$$(p^{(i)}, \tilde{p}^{(j)}) = 0 \quad \text{for } i \neq j, \qquad (16)$$

$$(q^{(i)}, A\tilde{q}^{(j)}) = 0 \quad \text{for } i \neq j. \qquad (17)$$

The above process is called the "p-q algorithm" by Lanczos [1952]; it is commonly known as the "Lanczos method". The procedure can be summarized as follows:

Algorithm 1. Lanczos method (A non-Hermitian)

Do for $k = 0,1,2...$ until $p^{(k)} = 0$

1. $p^{(0)} = q^{(0)} = b$
2. Choose $\tilde{b}$ arbitrarily and set $\tilde{p}^{(0)} = \tilde{q}^{(0)} = \tilde{b}$
3. Let $u^{(0)} = 0, \eta_{-1} = -1,$ then $r^{(k)} = b - Au^{(0)} = b$

compute $p^{(k+1)}, \tilde{p}^{(k+1)}, q^{(k+1)},$ and $\tilde{q}^{(k+1)}$ by (14)

$\eta_k = \eta_{k-1} / \rho_k$

$u^{(k+1)} = u^{(k)} + \eta_k q^{(k)}$

$r^{(k+1)} = r^{(k)} - \eta_k A q^{(k)}$

end do

We remark that elimination of the $q_k(x)$ in (11) leads to the following three-term recurrence relation which involves $p_k(x)$ alone:

$$p_{k+1}(x) = (x - \alpha_k)p_k(x) - \beta_{k-1}p_{k-1}(x) \tag{18}$$

where

$$\alpha_k = -(\rho_k + \sigma_{k-1})$$
$$\beta_{k+1} = \rho_k \sigma_k$$

This in turn leads to a method which involves the generation of two sets of vectors $p^{(k)}$ and $\tilde{p}^{(k)}$ satisfying (14).

Similarly, one can eliminate the $p_k(x)$ in (13) and get three-term recurrence relations involving only the $q_k(x)$. Thus we have

$$q_{k+1}(x) = (x - \hat{\alpha}_k)q_k(x) - \hat{\beta}_{k-1}q_{k-1}(x) \tag{19}$$

where

$$\hat{\alpha}_k = -(\rho_k + \sigma_k)$$
$$\hat{\beta}_{k-1} = \rho_k \sigma_{k-1}$$

The equation (19) leads to a method which involves the generation of two sets of bi-conjugate vectors $q^{(k)}$ and $\tilde{q}^{(k)}$ which satisfy the relation (17).

If A is symmetric positive definite (SPD), and if we choose $\tilde{b} = b$, then $p^{(k+1)} = \tilde{p}^{(k+1)}$ and $q^{(k+1)} = \tilde{q}^{(k+1)}$ in (14). Therefore, the work required with Algorithm 1 is reduced in half. Thus we have

Algorithm 2. Lanczos method (A SPD)

1. $p^{(0)} = q^{(0)} = b$
2. Let $u^{(0)} = 0, \eta_{-1} = -1,$ then $r^{(0)} = b - Au^{(0)} = b$
3. Do for $k = 0,1,2...$ until $p^{(k)} = 0$

$$p^{(k+1)} = \rho_k p^{(k)} + Aq^{(k)}$$

$$q^{(k+1)} = \sigma_k q^{(k)} + p^{(k+1)}$$

where

$$\rho_k = -\frac{(p^{(k)}, Aq^{(k)})}{(p^{(k)}, p^{(k)})}$$

$$\sigma_k = -\frac{(Ap^{(k+1)}, p^{(k)})}{(Aq^{(k)}, p^{(k)})} = -\frac{(p^{(k+1)}, p^{(k+1)})}{(Aq^{(k)}, p^{(k)})}$$

$$\eta_k = \eta_{k-1} / \rho_k$$

$$u^{(k+1)} = u^{(k)} + \eta_k q^{(k)}$$

$$r^{(k+1)} = r^{(k)} - \eta_k Aq^{(k)}$$

end do

Thus (16) and (17) become the following

$$(p^{(i)}, p^{(j)}) = 0 \qquad \text{for } i \neq j, \tag{20}$$

$$(q^{(i)}, Aq^{(j)}) = 0 \qquad \text{for } i \neq j. \tag{21}$$

It is easy to show that Algorithm 2 is closely related to the conjugate gradient (CG) method developed by Hestenes and Stiefel [1952] for solving a linear system (1). If one starts with $u^{(0)} = 0$ then one can show inductively that the residual vectors $r_{CG}^{(k)}$ generated by the CG method and the $p^{(k)}$ lie in the same Krylov space. Moreover, the residual vectors are also pairwise orthogonal and thus the $r_{CG}^{(k)}$ and the $p^{(k)}$ have the same direction. Similarly, the direction vectors $p_{CG}^{(k)}$ for the CG method are pairwise conjugate and have the same directions as the $q^{(k)}$ in Algorithm 2.

It should be noted that the Lanczos method minimizes the Euclidean length $\|r^{(k)}\| = (r^{(k)}, r^{(k)})^{1/2}$ of the residual vector which is A-norm of the error vector. Moreover, $\|r^{(k)}\|$ is the same as the A-norm $\|\varepsilon^{(k)}\|_A = \|A\varepsilon^{(k)}\| = (A\varepsilon^{(k)}, A\varepsilon^{(k)})^{1/2}$ of the error vector $\varepsilon^{(k)} = u^{(k)} - \bar{u}$, where $\bar{u} = A^{-1}b$. We remark that the usual form of the CG method minimizes the $A^{1/2}$ – norm, $\|\varepsilon^{(k)}\|_{A^{1/2}} = (\varepsilon^{(k)}, A\varepsilon^{(k)})^{1/2}$, of $\varepsilon^{(k)}$.

A real linear system (1) is said to be symmetrizable if there exists an SPD matrix Z such that Z and ZA are SPD. Otherwise, the system is nonsymmetrizable. Young and Jea [1980] considered a method called the "idealized generalized conjugate gradient method", (IGCG method), designed to handle the nonsymmetrizable case. The method involves the choice of an auxiliary matrix Z, and the determination of $u^{(n)}$ by the conditions

$$u^{(n)} - u^{(0)} \varepsilon K_n(r^{(0)}) \tag{22}$$

$$(Zr^{(n)}, v) = 0, \quad \text{for all } v \varepsilon K_n(r^{(0)}, A) \tag{23}$$

Here, $K_n(r^{(0)}, A)$ is the Krylov space spanned by the vectors $r^{(0)}, Ar^{(0)} \ldots A^{n-1}r^{(0)}$. Young and Jea [1981] showed that if Z and ZA are positive real in the sense that $Z+Z^T$ and $ZA+(ZA)^T$ are SPD then $u^{(n)}$ is uniquely determined by (22) and (23). Three equivalent forms of the IGCG method were given, namely, ORTHODIR, ORTHOMIN, and ORTHORES. These procedures reduce to simplified forms called ORTHODIR*, ORTHOMIN*, and ORTHORES*, respectively, if A is symmetrizable. We remark that ORTHOMIN* is the usual two-term form of the CG method defined by Hestenes and Stiefel [1952], while ORTHORES* is a three-term form of the CG method given by Engeli, et.al. [1959] and by Concus, Golub, and O'Leary [1976]. The method ORTHODIR* is defined in Young and Jea [1980].

As one method for handling the nonsymmetrizable case involving real systems, Jea and Young [1983] considered an expanded system of the form

$$\hat{A}\hat{u} = \hat{b} \tag{24}$$

where

$$\hat{A} = \begin{bmatrix} A & 0 \\ 0 & A^T \end{bmatrix}$$

$$\hat{u} = \begin{bmatrix} u \\ \tilde{u} \end{bmatrix}$$

and

$$\hat{b} = \begin{bmatrix} b \\ \tilde{b} \end{bmatrix}$$

The expanded system includes the original system (1) and a fictitious system $A^T\tilde{u} = \tilde{b}$, where $\tilde{b}$ is arbitrarily chosen.

Evidently, if

$$\hat{Z} = \begin{bmatrix} 0 & I \\ I & 0 \end{bmatrix} \tag{25}$$

then $\hat{Z}\hat{A}$ is symmetric and $\hat{Z}\hat{A} = \hat{A}^T\hat{Z}$. Because of this, as shown by Jea and Young [1983] the IGCG method for solving (26) with $\hat{Z}$ as given in (27) is greatly simplified. The three simplified versions converted back to N-vectors are called Lanczos/ORTHODIR, Lanczos/ORTHOMIN, and Lanczos/ORTHORES. The name "Lanczos" is added to each of these three procedures because it can be shown that Lanczos/ORTHODIR is equivalent to the three-term form of the Lanczos method where $q^{(k)}$ and $\tilde{q}^{(k)}$ correspond (19). We remark that Lanczos/ORTHOMIN is essentially the biconjugate gradient (BCG) method considered by Fletcher [1976].

We also remark that, as noted by Jea and Young [1983], there is no guarantee that the three methods, Lanczos/ORTHODIR, Lanczos/ORTHOMIN, and Lanczos/ORTHORES, will not break down. However, there exists an integer $t \leq N$ such that if any one of the three methods does not

break down within $t+1$ iterations then $u^{(t+1)} = \bar{u}$. Also, the Lanczos/ORTHOMIN method converges if and only if the Lanczos/ORTHORES method converges, and if both converge, then the Lanczos/ORTHODIR method converges and in that case and all three methods are equivalent. From this it would appear that the Lanczos/ORTHODIR method is the most robust. However, Lanczos/ORTHODIR appears to be subject to roundoff error.

A number of other methods have been proposed which are related to the Lanczos method. These include oblique projection method considered by Saad [1982], the conjugate gradient-squared (CGS) method of Sonneveld [1989] and the CGSTAB method of Van der Vorst and Sonneveld [1990], to mention only a few.

References

Concus, P., Golub, G.H. and O'Leary, D.P. [1976] "A Generalized Conjugate Gradient Method for the Numerical Solution of Elliptic Partial Differential Equations," in *Sparse Matrix Computation* (J.R. Bunch and D.J. Rose, eds.), Academic Press, New York, 309-332.

Engeli, M., Ginsburg, T.H, Rutishauser, H. and Stiefel, E. [1959] "Refined Iterative Methods for Computation of the Solution and the Eigenvalues of Self-adjoint Boundary Value Problems," *Mitt. Inst. Angew. Math.* ETH, Zurich, Nr. 8, Basel-Stuttgart.

Fletcher R., [1976] "Conjugate Gradient Methods for Indefinite Systems," *Numerical Analysis* Dundee 1975, G.A. Watson ed., New York: Springer, Lecture Notes in Mathematics, no. 506, 73-89.

Hestenes, M.R. and Stiefel, E.L. [1952] "Methods of Conjugate Gradients for Solving Linear Systems," *J. Res. Nat. Bur. Standards*, **49**, 409-436.

Jea, K.C. and Young, D.M. [1983] "On the Simplification of Generalized Conjugate-Gradient Methods for Nonsymmetrizable Linear Systems," *Linear Algebra and its Applications*, 52/53, 399-417.

Lanczos, C. [1950] "An Iteration Method for the Solution of the Eigenvalue Problem of Linear Differential and Integral Operators," *J. Res. Nat. Bur. Standards*, **45**, 255-282.

Lanczos, C. [1952] "Solution of Systems of Linear Equations by Minimized Iterations," *J. Res. Nat. Bur. Standards*, **49**, 33-53.

Luenberger, D.G. [1970] "The Conjugate Residual Method for Constrained Minimization Problems." *SIAM J. Numer. Anal.* 7, 1970, 390-398.

Saad, Y. [1982] "The Lanczos Biorthogonalization Algorithm and Other Oblique Projection Methods for Solving Large Unsymmetric Systems," *SIAM J. Numer. Anal.*, **19**, no. 3, 485-506.

Saad, Y. and Schultz, M.H., [1986] "GMRES: A Generalized Minimal Residual Algorithm for Solving Nonsymmetric Linear Systems," *SIAM J.*, **7**, no. 3, 856-869.

Sonneveld, P. [1989] "CGS, a Fast Lanczos-type Solver for Nonsymmetric Linear Systems," *SIAM J. Sci. Stat. Comput.*, **10**, no. 1, 36-52.

van der Vorst, H.A. and Sonneveld, P. [1990] "CGSTAB: A More Smoothly Converging Variant of CG-S," submitted to SIAM J. Sci. Stat. Comp, to appear.

Young, D.M. and Jea, K.C. [1980] "Generalized Conjugate-Gradient Acceleration of Nonsymmetrizable Iterative Methods," *Linear Algebra and its Applications*, **34**, 159-194.

Young, D.M. and Jea, K.C. [1981] "Generalized Conjugate-Gradient Acceleration of Nonsymmetrizable Iterative Methods. Part 2: The Nonsymmetrizable Case," The University of Texas at Austin, Center for Numerical Analysis, Report CNA-163.

Commentary on Lanczos [1953]

by

David M. Young and Kang C. Jea

Commentary Lanczos [1953].
"Chebyshev Polynomials in the Solution of Large-Scale Linear Systems"

In this paper, Lanczos considered iterative methods for solving the linear system

$$Ay = b \tag{1}$$

where A is a given square nonsingular matrix of order N which is symmetric and positive definite (SPD), all of whose eigenvalues lie in the interval (0,1]. For the more general case where A is an arbitrary nonsingular complex matrix, one can consider the normal equations

$$Cy = c \tag{2}$$

where $C = A^H A$, $c = A^H b$, and A^H is the conjugate transpose of A. The system (2) can be scaled by dividing both sides by the factor μ where

$$\mu = \max_i \sum_{j=1}^{N} |c_{ij}| \tag{3}$$

Thus one obtains the scaled system

$$C_0 y = c_0 \tag{4}$$

where $C_0 = \mu^{-1}C$, $c_0 = \mu^{-1}c$ and the eigenvalues of C_0 are positive and do not exceed one.

The true solution of the linear system (4) is given by $\bar{y} = C_0^{-1}c_0$. The main object of the paper is to seek polynomials $G_0(x), G_1(x), \ldots$, such that for each m a good approximation to $\bar{y}$ is given by

$$y_m = G_m(C_0)c_0 \tag{5}$$

The residual vector, r_m, corresponding to y_m is given by

$$
\begin{aligned}
r_m &= c_0 - C_0 y_m \\
&= (I - C_0 G_m(C_0)) c_0 \\
&= F_{m+1}(C_0) c_0
\end{aligned} \tag{6}
$$

where

$$F_{m+1}(x) = 1 - xG_m(x) \tag{7}$$

is the "residual" polynomial of degree $m+1$. The polynomials $F_{m+1}(x)$ are chosen so that $F_{m+1}(0)=1$ and so that $F_{m+1}(x)$ is small everywhere else in the interval [0,1]. Lanczos used two approaches to construct these polynomials:

(1) Fejer kernel approach

Corresponding to the Fejer kernel for Fourier series, Lanczos considered the polynomials

$$F_{m+1}(x) = \frac{1 - T_{m+2}(1-2x)}{(m+2)^2 2x} \tag{8}$$

where $T_k(x)=\cos(k \cos^{-1} x)$ is the kth Chebyshev polynomial of the first kind. Then by (7)

$$G_m(x) = \frac{T_{m+2}(1-2x) + 2(m+2)^2 x - 1}{2(m+2)^2 x^2} \tag{9}$$

Moreover, y_m and r_m given by (5) and (6), respectively, can be evaluated effectively by using the new vectors g_m where

$$g_m = (\frac{m+2}{2})^2 G_m(C_0) c_0 \tag{10}$$

and by the use of a three-term recursion relation satisfied by Chebyshev polynomials. Moreover, by analyzing the residual polynomials $F_{m+1}(x)$, Lanczos showed that if all eigenvalues λ of C_0 statisfy

$$\lambda \geq \lambda_0 = (\frac{2.55}{m+2})^2 \tag{11}$$

then the component of the residual vector corresponding to the eigenvector associated with λ cannot be more than σ_0,(σ_0 ~.05), times the corresponding component of the right hand side b. For example, if m=6 and if the matrix C_0, regardless of its size, contains no eigenvalue smaller than λ_0, then the approximate solution y_6 will be accurate to within 5%. Since λ_0=0.1016, the permissible spread of the eigenvalues of C_0 is about 1:10. (Also one can say that the "condition" of C_0 and the "skewness" of C_0 is about 10.)

(2) Dirichlet kernel approach

Corresponding to the Dirichlet kernel for Fourier series, Lanczos considered the polynomials

$$\overline{F}_{m+1}(x)=\frac{1}{m+1}(1-x)S_m(1-2x) \tag{12}$$

where $S_m(x)$ denotes the Chebyshev polynomial of the second kind defined by

$$S_m(x)=\frac{1}{m+1}T'_{m+1}(x)=\frac{\sin(m+1)\theta}{\sin\theta} \tag{13}$$

where x=cosθ. Then the associated polynomials $\overline{G}_m(x)$ are given by

$$\overline{G}_m(x)=\frac{1-\overline{F}_{m+1}(x)}{x}=\frac{1}{m+1}\overline{g}_m(x) \tag{14}$$

and where the $\overline{g}_m(x)$ satisfy the recurrence relation

$$\overline{g}_{m+1}(x)=2(1-2x)\overline{g}_{m-1}(x)+(m+1) \tag{15}$$

It can be shown from (12) that the maximum relative error is now $\sqrt{\sigma_0}$ instead of σ_0. However, if we repeat the entire cycle with the residual obtained after one cycle as the new right hand side, then the previously obtained bound σ_0 is reached. Moreover, if the eigenvalues λ of C_0 satisfy

$$\lambda \geq \tilde{\lambda}_0=\left(\frac{1.28}{m+1}\right)^2 \tag{16}$$

then the component of the residual vector after 2 cycles corresponding to the eigenvector associated with λ cannot be more than 5% of the corresponding component of the right hand side b. Hence, if we take m=7, then $\tilde{\lambda}_0$ can be as small as 0.025. Thus, with a spread of 1:40 for the eigenvalues of C_0, an accuracy of 5% can be obtained by use of two cycles. Greater allowable spreads and greater accuracy can be obtained using larger values of m and/or more cycles. However, the allowable eigenvalue spread increases relatively slowly with m and with the number of cycles.

Lanczos considered the use of polynomial preconditioning to handle cases where the eigenvalue spread is very large. Given m and an SPD matrix C_0 whose eigenvalues lie in the interval $[\beta,1]$ where $\beta>0$ and $\beta<<1$, he constructed a polynomial $R_m(x)$ so that the eigenvalues of $Q_m(C_0)=R_m(C_0)C_0$ lie in the interval $[\delta,1]$ where δ is considerably larger than β; δ might be as large as 0.1. Thus the eigenvalue spread of the preconditioned system is much smaller. This preconditioning procedure can be combined with either of the two iteration procedures described above to get a high degree of error reduction even when C_0 has a very large eigenvalue spread.

The polynomial $Q_m(x)$ is chosen to have the form

$$Q_m(x)=\frac{T_m(1+\varepsilon)-T_m(1+\varepsilon-2x)}{T_m(1+\varepsilon)+1} \tag{17}$$

The maximum absolute value of this polynomial is 1 while the minimum absolute value is

$$\frac{T_m(1+\varepsilon)-1}{T_m(1+\varepsilon)+1} \tag{18}$$

By proper choice of ε one can prevent $|Q_m(x)|$ from dropping below δ (provided m is large enough). The author suggests that for many problems involving larger eigenvalue spreads it may not be possible to attain good accuracy, especially in cases where the input data is obtained from physical measurements. However, he indicates that even in such cases useful information can often be obtained from applying the methods.

The paper concludes with some examples of problems where the availability of a program based on the given methods would be useful. One class of problems involves solving systems of "moderate skewness"; another class of problems involves the determination of a few eigenvalues of a matrix.

The paper contains a number of interesting ideas which are particularly innovative, considering that they were developed in the early 1950's. It does seem however, that, especially for some problems involving partial differential equations where accurate or exact data is assumed, one can indeed obtain high accuracy even when the matrix is very ill conditioned. This can be done for many problems using Chebyshev acceleration (a.k.a. "semi-iterative methods;" see e.g. Varga [1957,1962] and Golub and Varga [1961].) Here instead of the Fejer or the Dirichlet polynomials one can use Chebyshev polynomials for the $F_{m+1}(x)$. The $F_{m+1}(x)$ are one at x=1 and are chosen to minimize their maximum absolute values in $[\theta,1]$ where θ is the smallest eigenvalue of C_0. The iterants y_m satisfy a three-term recurrence relation and are easily computed. Moreover, the smallest eigenvalue θ can be determined adaptively as described by Hageman and Young [1981].

References

Golub, G.H., and Varga, R.S. [1961], "Chebyshev Semi-iterative Methods, Successive Overrelaxation Iterative Methods, and Second-order Richardson Iterative Methods, Parts I and II." *Numer. Math.* **3**, 147-168.

Hageman, L.A. and Young, D.M. [1981], *Applied Iterative Methods*, Academic Press, New York.

Lanczos, C. [1953], "Chebyshev Polynomials in the Solution of Large-Scale Linear Systems", *Proceedings of the Association for Computing Machinery*, Toronto, 1952, 124-133, Sauls Lithograph Co., Washington, DC.

Varga, R.S. [1957], "A Comparison of the Successive Overrelaxation Method and Semi-iterative Methods Using Chebyshev Polynomials," *J. Soc. Indus. Appl. Math.* **5**, 39-46.

Varga, R.S. [1962], *Matrix Iterative Analysis*, Prentice Hall, Englewood Cliffs, N.J.

Commentary on Lanczos [1958]

by

Kang C. Jea and David M. Young

Commentary on Lanczos [1958]
"Iterative Solution of Large-Scale Linear Systems"

As the title implies, this paper is concerned with the iterative solution of large-scale linear systems. The first part of the paper is primarily devoted to a discussion of the limitations in the attainable accuracy which may result when some of the input data, such as the elements of the matrix or the elements of the right hand side, are not exact but are instead determined from physical measurements and thus are of limited accuracy. Another source of error often arises when the matrix is "ill-conditioned", i.e. if the condition number of the matrix is large. (The condition number is also referred to as "skewness" or "eigenvalue spread").

The primary emphasis of the paper is on the case where the coefficient matrix A of the system is symmetric and positive definite (SPD). The first step is to determine an upper bound for the largest eigenvalue of A. This is done by a Bernoulli type method involving several steps of a modified power method. The linear system

$$Ay=b \tag{1}$$

is then divided by the estimated largest eigenvalue to give the scaled system

$$A_0 y = b_0 \tag{2}$$

where A_0 is SPD and where the largest eigenvalue of A_0 is not greater than one.

The method for solving (2) is similar to that described in Lanczos [1953], and is based on the use of the polynomials $Q_0(x)$, $Q_1(x)$,..., where

$$Q_{m-1}(x) = \frac{\sin^2(m\theta/2)}{\sin^2(\theta/2)} \tag{3}$$

and where $\sin^2(\theta/2) = x$. As in Lanczos [1953] it can be shown that if all eigenvalues λ of A_0 lie in the range

$$\lambda_0 = \left(\frac{2.56}{m+2}\right)^2 \leq \lambda \leq 1 \tag{4}$$

then a solution is obtained with a relative error of about 5%. Thus for example, if the condition number of A is 100 one can obtain a relative error of 5% with 24 iterations. If the condition number of A is 1000 such accuracy could be obtained in 80 iterations. The author claims that for physical systems, larger condition numbers would "hardly be permissible".

As an improvement on the method just described for solving SPD systems, the author developed an "additional algorithm". We illustrate by an example. Suppose one desires to obtain a relative error of 5%. If one estimates the condition number to be 100, one can carry out 24 iterations and see if the length of the residual r_1 does not exceed 5% of the length of b_0. If the residual is as small as expected and if one is satisfied with 5% accuracy the solution thus obtained is considered to be satisfactory. If one desires higher accuracy or if the length of r_1 indicates that the condition number is greater than 100 the cycle is repeated. Actually Lanczos repeats the cycle twice obtaining the approximations y_1,y_2, and y_3. He then chooses α_1, α_2, and α_3 so that

$$\alpha_1 + \alpha_2 + \alpha_3 = 1, \tag{5}$$

and such that (r,r) is minimized where r=b-Ay is the residual corresponding to

$$y = \alpha_1 y_1 + \alpha_2 y_2 + \alpha_3 y_3. \tag{6}$$

An algorithm is given for finding α_1, α_2, and α_3.

The above procedure can be used to obtain an improved estimate of the condition number. This information can in turn be used to modify the iteration process to obtain improved convergence.

The last part of the paper is concerened with the case where A is real but nonsymmetric. In this case the system Ay=b is replaced by the double system

$$\begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \begin{bmatrix} 0 \\ y \end{bmatrix} = \begin{bmatrix} b \\ 0 \end{bmatrix} \tag{7}$$

or

$$\hat{A}\hat{y} = \hat{b} \tag{8}$$

where

$$\hat{A} = \begin{bmatrix} 0 & A \\ A^T & 0 \end{bmatrix} \tag{9}$$

and

$$\hat{y} = \begin{bmatrix} 0 \\ y \end{bmatrix}, \quad \hat{b} = \begin{bmatrix} b \\ 0 \end{bmatrix} \tag{10}$$

It should be noted that if A is SPD then the eigenvalues μ of the double matrix $\hat{A}$ lie on two disjoint intervals $-\beta \le \mu \le -\alpha < 0$ and $0 < \alpha \le \mu \le \beta$ where α and β are the smallest and largest eigenvalues of A, respectively. This suggests that the convergence depends on the ratio $(\beta / \alpha)^2$ instead of β / α, the condition number of A, and hence that the convergence properties are more like those obtained for the normal equations rather than for the original system (1).

In any case the polynomials considered here are defined by

$$Q_{2m-2}(x) = \frac{1-(-1)^m T_{2m}(x)}{2x^2} = \frac{\sin^2 m\phi}{\sin^2 \phi} \tag{11}$$

Here, $x=\cos\theta=\sin\varphi$ where $\phi = (\pi / 2) - \theta$. Although the polynomials of odd order $Q_{2m-1}(x)$ are not be used for the solution, they are of interest since they are involved in the three-term recurrence relation, by which the polynomials of even order are generated. They are defined by

$$Q_{2m-1}(x) = \frac{(-1)^m T_{2m+1}(x) - (2m+1)x}{2x^2} \tag{12}$$

The subsequent analysis is similar to that used for the SPD case. However, it is found that the number of iterations required is approximately linear in the condition number rather than approximately proportional to the square root of the condition number as in the SPD case. The author indicates that the method will remain economical only if the machine used is of high speed, if the system has many zero elements, or if the condition number of the matrix is "very moderate".

The paper contains a number of interesting and innovative ideas. One important idea is that of estimating the condition number of a matrix and applying an iterative procedure involving the matrix which is based on the estimate. If the convergence is slower than expected, an improved estimate of the condition number is obtained and the iterative procedure is continued based on the new estimate. The process is continued until convergence is obtained. This idea was used by Hageman and Young [1981] for the development of adaptive procedures for the acceleration of iterative methods based on Chebyshev polynomials.

References

Hageman, L.A. and Young, D.M. [1981], *Applied Iterative Methods*, Academic Press, New York.

Lanczos, C. [1953], "Chebyshev Polynomials in the Solution of Large-Scale Linear Systems", *Proceedings of the Association for Computing Machinery*, Toronto, 1952, 124-133, Sauls Lithograph Co., Washington, DC.

Lanczos, C. [1958], "Iterative Solution of Large-Scale Linear Systems", *J. Soc. Indust. Appl. Math.*, **6**, No. 1, 91-109.