ON THE INVERSION OF MATRICES AND LINEAR OPERATORS

W. V. PETRYSHYN

1. **Introduction.** The purpose of this short paper is to discuss an iterative method of order $p \geqq 2$ for inversion of nonsingular matrices and bounded linear operators in finite- and infinite-dimensional Banach and Hilbert spaces. Here we extend as well as unify the results of a number of authors [1], [2], [3], [4], [5], [6], [11]. The method is essentially the hyper-power method of order $p \geqq 2$ considered by John [5] for matrices and by Altman [2] for operators. When $p=2$ the method reduces to the procedure first suggested for matrices by Schulz [11] and later discussed by Hotelling [6] and Ansorge [3] and recently by Dück [4] and Albrecht [1]. The convergence is proved here under a condition which is weaker than the one assumed by Altman [2] and the error estimates derived here are better than the estimates derived in [2]. Furthermore, we show the relationship that exists among these various estimates. At the end, using the results concerning the K-p.d. matrices and operators derived by the author in [8], [10] we show how to choose the initial approximation to the inverse of a given matrix or operator so that the convergence condition is satisfied. This gives the answer to the practically difficult problem pointed out by Newman [7] for the class of matrices and operators considered in the last section of this paper.

2. **The method and the error estimates.** Let B denote a Banach space with the norm $|\ |$, I an identity mapping of B, and $R(A)$ the range space of a linear bounded operator A defined on B which, in what follows, we shall assume to be *continuously invertible*, i.e., A has a bounded inverse A^{-1} defined on $R(A)=B$. We say that a given linear bounded operator X_0 satisfies a $\delta(T_0)$-*condition* if the spectrum $\delta(T_0)$ of the operator $T_0 \equiv I - X_0 A$ (i.e., the set of all complex numbers λ for which the operator $(\lambda I - T_0)$ is not continuously invertible) lies in the interior of the unit circle centered at zero. If p is a positive integer with $p \geqq 2$, then starting with X_0 we construct a sequence X_n of approximations to the inverse A^{-1} of a given continuously invertible operator A by the following procedure: If X_n is the iterant constructed at the nth step of the process, then the succeeding iterant

Received by the editors July 24, 1964.

X_{n+1} is determined by

$$\tilde{X}_n = (I + T_n + T_n^2 + \cdots + T_n^{p-2})X_n, \tag{1}$$

$$X_{n+1} = X_n + T_n\tilde{X}_n, \tag{2}$$

where T_n is defined by

$$T_n \equiv I - X_nA, \qquad n = 0, 1, 2, \cdots. \tag{3}$$

It follows from (2) that

$$T_{n+1} = I - X_{n+1}A = I - X_nA - T_n\tilde{X}_nA = T_n(I - \tilde{X}_nA). \tag{4}$$

On the other hand, in view of the identity,

$$(I + C + \cdots + C^r)(I - C) = I - C^{r+1} \tag{5}$$

valid for any linear bounded operator C in B, we obtain from (1)

$$I - \tilde{X}_nA = T_n^{p-1}. \tag{6}$$

This and (4) imply that $\{X_{n+1}\}$ determined by (1)–(2) is such that

$$T_{n+1} = T_n^p, \qquad n = 0, 1, 2, \cdots, n, \cdots. \tag{7}$$

Consequently, the process (1)–(2) is of order p and is essentially the hyperpower method studied in [2].

THEOREM 1. *Let the initial approximation X_0 satisfy the $\delta(T_0)$-condition. Then the sequence of successive approximations $\{X_{n+1}\}$ determined by* (1)–(2) *converges in the operator norm to A^{-1} and there exists an integer $r>0$ such that $|T_n|<1$ for all $n\geqq r$ and the error estimate $|E_{n+1}| \equiv |A^{-1}-X_{n+1}|$ for $n\geqq r$ is given by*

$$|A^{-1} - X_{n+1}| \leqq \frac{|T_{n+1}X_{n+1}|}{1 - |T_{n+1}|}. \tag{8}$$

If instead of $\delta(T_0)$-condition we assume the stronger condition

$$|T_0| = |I - X_0A| < 1, \tag{9}$$

then in addition to (8) *the following, though less precise but more practical, three error estimates are valid.*

$$|A^{-1} - X_{n+1}| \leqq \frac{|T_n|}{1 - |T_n|}\,|X_{n+1} - \tilde{X}_n|, \tag{10}$$

$$|A^{-1} - X_{n+1}| \leqq |T_n|^{p-1}\frac{|T_nX_n|}{1 - |T_n|}, \tag{11}$$

$$|A^{-1} - X_{n+1}| \leqq |T_0|^{p^{n+1}} \frac{|X_0|}{1 - |T_0|}, \tag{12}$$

whose degree of precision decreases in the given order.

PROOF. To prove the convergence of X_{n+1} to A^{-1} note that since X_0 satisfies the $\delta(T_0)$-condition the spectral radius $r(T_0) \equiv \sup_{\gamma \in \delta(T_0)} |\lambda|$ of $T_0 \equiv I - X_0 A$ is less than 1. On the other hand it is known [12] that $r(T_0) = \lim_m (|T_0^m|)^{1/m}$ and that for every positive integer m

$$r(T_0) \leqq (|T_0^m|)^{1/m}. \tag{13}$$

Since $|T_0^m|$ are positive numbers and $r(T_0) < 1$, the radical test for convergence of an infinite series of numbers implies that the series $\sum_{m=0}^{\infty} |T_0^m|$ converges. This shows that $|T_0^m| \to 0$, as $m \to \infty$, and hence, in view of (7) and the fact that

$$A^{-1} - X_{n+1} = T_0^{p^{n+1}} A^{-1},$$

proves the convergence of $\{X_{n+1}\}$ to A^{-1} in the operator norm.

To obtain the estimate (8) we first note that since $|T_0^m| \to 0$, as $m \to \infty$, there must exist an integer $r > 0$ such that

$$|T_r| = |T^{p^r+1}| < 1 \tag{14}$$

whence, since $T_{r+i} = T_r^{p^i}$ for any integer $i \geqq 0$, we see that

$$|T_{r+i}| \leqq |T_r|^{p^i} < 1, \tag{15}$$

i.e., $|T_n| < 1$ for every $n \geqq r$. Put $E_{n+1} \equiv A^{-1} - X_{n+1}$ and consider the equality

$$\begin{aligned} \{E_{n+1} - T_{n+1}E_{n+1}\} &= X_{n+1}A(A^{-1} - X_{n+1}) \\ &= (I - X_{n+1}A)X_{n+1} \\ &= T_{n+1}X_{n+1}. \end{aligned} \tag{16}$$

Since $|T_{n+1}| < 1$, (16) and the properties of the norm yield the inequality

$$\{1 - |T_{n+1}|\} |E_{n+1}| \leqq |E_{n+1} - T_{n+1}E_{n+1}| = |T_{n+1}X_{n+1}|$$

from which we derive the estimate (8).

To prove the other assertions of Theorem 1 let us first observe that, in view of (13) with $m = 1$, the condition (9) implies the $\delta(T_0)$-condition of X_0 so that in this case the first part of Theorem 1 remains valid for every n. To obtain (9) from (8) note that, by virtue of (7), (1), and (2), a simple manipulation shows that

$$\begin{aligned} T_{n+1}X_{n+1} &= T_n(T_n^{p-1}X_{n+1}) \\ &= T_n\{(I + T_n + \cdots + T_n^{p-2} + T_n^{p-1})X_{n+1} \\ &\qquad - (I + T_n + \cdots + T_n^{p-2})X_{n+1}\} \\ &= T_n\{(I + T_n + \cdots + T_n^{p-1})X_{n+1} \\ &\qquad - (I + T_n + \cdots + T_n^{p-2})(X_n + T_n\tilde{X}_n)\} \\ &= T_n\{I + T_n + \cdots + T_n^{p-1}\}(X_{n+1} - \tilde{X}_n). \end{aligned}$$

This and the properties of the operator norm imply that

$$(17) \quad \begin{aligned} &|T_{n+1}X_{n+1}| \\ &\qquad \leqq |T_n|\{1 + |T_n| + \cdots + |T_n|^{p-1}\}|X_{n+1} - \tilde{X}_n|. \end{aligned}$$

On the other hand, since $|T_{n+1}| = |T_n^p| \leqq |T_n|^p$, we have

$$(18) \quad \begin{aligned} 1 - |T_{n+1}| &\geqq 1 - |T_n|^p \\ &= (1 - |T_n|)(1 + |T_n| + \cdots + |T_n|^{p-1}). \end{aligned}$$

Furthermore, in view of (7), the stronger condition (9) implies that $|T_n| < 1$ for every positive integer n. Thus, using this and (17) and (18), we derive from (8) the estimate (10):

$$(10_0) \quad |A^{-1} - X_{n-1}| \leqq \frac{|T_{n-1}X_{n-1}|}{1 - |T_{n+1}|} \leqq \frac{|T_n|}{1 - |T_n|}|X_{n-1} - \tilde{X}_n|.$$

The estimate (11) follows immediately from (10). In fact, by (1) and (2), $|X_{n+1} - \tilde{X}_n| = |T_n^{p-1}X_n| \leqq |T_n|^{p-2}|T_nX_n|$ so that, by virtue of (10_0), we have for each n the inequality

$$(19) \quad \frac{|T_{n+1}X_{n+1}|}{1 - |T_{n+1}|} \leqq \frac{|T_n||X_{n+1} - \tilde{X}_n|}{1 - |T_n|} \leqq |T_n|^{p-1}\frac{|T_nX_n|}{1 - |T_n|}$$

from which (11) follows.

The estimate (12) is then derived from (11) or (19) since by mathematical induction we obtain from (19) the inequality

$$(20) \quad \begin{aligned} \frac{|T_{n+1}X_{n+1}|}{1 - |T_{n+1}|} &\leqq \frac{|T_n||X_{n+1} - \tilde{X}_n|}{1 - |T_n|} \leqq |T_n|^{p-1}\frac{|T_nX_n|}{1 - |T_n|} \\ &\leqq |T_n|^{p-1}|T_{n-1}|^{p-1}\frac{|T_{n-1}X_{n-1}|}{1 - |T_{n-1}|} \leqq \cdots \\ &\leqq \{|T_n||T_{n-1}| \cdots |T_0|\}^{p-1}\frac{|T_0X_0|}{1 - |T_0|}. \end{aligned}$$

Since $T_{n-j}=T_0^{p^{n-j}}$ for $j=0, 1, \cdots, n$; $p\geqq 2$, and

$$p^n+p^{n-1}+\cdots+p+1=p^n\left(1+\frac{1}{p}+\cdots+\frac{1}{p^n}\right)=\frac{p^{n+1}-1}{p-1}$$

we have the equality

$$\{|T_n||T_{n-1}|\cdots|T_0|\}^{p-1}=\{|T_0|^{(p^n+p^{n-1}+\cdots+p+1)}\}^{p-1}$$
$$=|T_0|^{(p^{n+1}-1)}.$$

Hence, using the last equality, we derive from (20) the estimate (12) and the fact that the degree of precision of the estimates for the hyperpower method (1)–(2) decreases as we go from (8) to (12) in the given order. Thus, the proof of Theorem 1 is complete.

Let us observe that when B is a finite-dimensional normed linear vector space, then the continuously invertible operator A is in this case simply a nonsingular matrix $A=(a_{ij})$ and, since in this case the spectrum $\delta(T_0)$ of the matrix $T_0\equiv I-X_0A$ consists only of eigenvalues of T_0, the $\delta(T_0)$-condition simply means that the eigenvalues of T_0 are of modulus less than 1. Furthermore, as is well known, this condition is not only sufficient but also necessary for the matrix T_0 to be convergent. Thus for matrices of finite order we have the following result

COROLLARY 1. *If $B=B_N$ is a normed linear vector space of dimension N with norm $|\ |$ and $A=(a_{ij})$, $1\leqq i,j\leqq N$, is a nonsingular matrix in B_N, then the sequence of iterants X_{n+1} determined by*

$$(1_0)\qquad \tilde{X}_n=(I+T_n+\cdots+T_n^{p-2})X_n,$$

$$(2_0)\qquad X_{n+1}=X_n+T_n\tilde{X}_n$$

converges in the matrix norm[1] *to A^{-1} if and only if X_0 is so chosen that the eigenvalues of the matrix $T_0\equiv I-X_0A$ are of modulus less than* 1.

In case of convergence we have the error estimate analogous to (8). *If in addition we assume that X_0 satisfies the condition* (9), *then the error estimates* (10)–(12) *are also valid.*

Special cases. (i) If $p=2$, then $\tilde{X}_n=X_n$ for each n and (1)–(2) reduces in this case to the quadratically convergent method suggested for finite matrices by Schulz [11]

[1] If $X^*=(X_1, \cdots, X_N)$ is the transpose of any column vector X in B, then for the norm $|X|$ we may take, for example, $|X|\equiv|X|_\infty=\mathrm{Max}_{1\leq i\leq N}|X_i|$ or $|X|\equiv|X|_1=\sum_{i=1}^N|X_i|$ or some other norm $|X|$. The corresponding matrix norms would be $|A|\equiv|A|_\infty=\max_i\sum_{j=1}^N|a_{ij}|$ or $|A|\equiv|A|_1=\max_j\sum_{i=1}^N|A_{ij}|$ or some other norm $|A|$ corresponding to $|X|$.

(i) $$X_{n+1} = X_n + (I - X_nA)X_n = X_n(2 - AX_n).$$

Theorem 1 establishes the convergence of (i) and determines for it the four error estimates both for matrices and operators. Let us note that when A is a matrix and X_0 is such that condition (9) is satisfied the estimates (10) and (11) for the method (i) were derived by Dück [4] under the assumption that X_n is nonsingular for each n and later by Albrecht [1] without this assumption. The estimate (12) for (i) was derived by Hotelling [6] (see also Ansorge [3], John [5], and Newman [7]).

(ii) If $p=3$, then $\bar{X}_n=(2-X_nA)X_n$ and $X_{n+1}=X_n+T_n\bar{X}_n$ or equivalently the sequence $\{X_{n+1}\}$ is determined by the cubically convergent method

(ii) $$X_{n+1} = \{3 - 3X_nA + (X_nA)^2\}X_n.$$

For matrices satisfying the condition (9) the error estimate (11) was derived in this case by Albrecht [1] who also showed its connection with the *improved Newton's method.* In case of operators the method (ii) was investigated by Altman [2] who, imposing the condition (9), proved its convergence with the error estimate (12) and showed that among all the hyperpower methods the method (ii) is best in the sense that the same number of multiplications gives a better accuracy in terms of the error estimate (12) for $p=3$ than for any other $p\geqq 2$.

3. **Inversion of K-p.d. matrices and operators.** We shall now apply Theorem 1 to the problem of inverting a nonsingular matrix or a continuously invertible operator of a K-p.d. type in a finite- or infinite-dimensional Hilbert space H with the inner product $(\ ,\ )$ and norm $\|\ \|$.

Let A be a linear bounded operator in H of the form

(21) $$A \equiv D + Q,$$

where D is K-symmetric and K-positive definite (K-p.d.) and Q is K-symmetric,[2] i.e., there is a positive definite Hermitian operator K and some positive constant $\eta_1>0$ such that for all u and v in H

(22) $$(Du, Kv) = (Ku, Dv), \qquad (Qu, Kv) = (Ku, Qv),$$

(23) $$(Du, Ku) \geqq \eta_1^2(u, u).$$

Note that since D and K are bounded operators in H there is a constant $\eta_2>0$ such that

[2] For various properties and general theory of K-p.d. matrices of finite order and bounded and unbounded operators see the author's papers [8], [9], [10].

$$(Du, Ku) \leqq \eta_2^2(u, u). \tag{24}$$

It was already pointed out in the introduction that the main disadvantage of the method (1)–(2) is that it is a very difficult practical problem to find the initial approximation X_0 to A^{-1} which satisfies the $\delta(T_0)$-condition or the stronger condition (9). Recently, Altman [2] showed that if A is a self-adjoint and positive definite operator in H such that

$$m(u, u) \leqq (Au, u) \leqq M(u, u), \qquad u \in H, \tag{25}$$

where $0 < m < M$ and m, M are the smallest and largest eigenvalues of A, respectively, then the initial approximation X_0 defined by $X_0 = \alpha I$ satisfies the condition (9) if α is any constant satisfying the condition $0 < \alpha < 2/M$.

Using our results in [8], [10] we shall show here how to choose X_0 so that the $\delta(T_0)$-condition is satisfied.

Indeed, let X_0 be a given initial approximation to A^{-1} and compute the successive approximations $\{X_{n+1}\}$ by the scheme (1)–(2). Then the following theorem is valid.

THEOREM 2. (a) *Let D be K-symmetric and K-p.d., Q be K-symmetric, and $G \equiv D - Q$ be K-p.d. If the operator $A \equiv D + Q$ is K-p.d. and the initial approximation X_0 to A^{-1} is chosen to be $X_0 = D^{-1}$, then the sequence X_{n+1} determined by* (1)–(2) *converges in the operator norm to the inverse A^{-1}. Furthermore, there is an integer $r > 0$ such that for $n \geqq r$ the error estimate is given by the formula*

$$\|A^{-1} - X_{n+1}\| \leqq \frac{\|T_{n+1}X_{n+1}\|}{1 - \|T_{n+1}\|}. \tag{8_1}$$

(b) *If in addition to the above condition on G and A we assume that*

$$r(D^{-1}Q) < \frac{\eta_1}{\eta_2}, \tag{9_1}$$

then the estimates (10)–(12) *remain also valid.*

PROOF. The proof of Theorem 2 follows from Theorem 1 and the Main Theorem in [8]. Let us first observe that the conditions satisfied by D imply that it is continuously invertible so that the choice $X_0 = D^{-1}$ is always possible. Furthermore, A is K-symmetric and since, by hypothesis, A is also K-p.d. it follows that A is continuously invertible. If we now take $X_0 = D^{-1}$, then $T_0 = I - D^{-1}A = -D^{-1}Q$ and hence by Theorem 1, it is sufficient to show that $r(T_0) = r(D^{-1}Q)$

<1. This, however, follows as a special case of the Main Theorem [8] according to which (under present conditions on D, Q, and G) $r(D^{-1}Q)<1$ if and only if A is K-p.d.

To prove the second part of Theorem 2 let us introduce a new inner product and norm in H by

$$[u, v] = (Du, Kv), \qquad |u|^2 = [u, u], \qquad u, v \in H, \tag{26}$$

and denote the resulting space by H_0. Since D is K-symmetric and K-p.d. the metric (26) is well defined and, in view of (23) and (24), the norms $\|\ \|$ and $|\ |$ are equivalent. It is easy to see that $T_0 = -D^{-1}Q$, considered as an operator in H_0, is Hermitian for, indeed, we have

$$\begin{aligned}[T_0u, v] &= (DT_0u, K_v) = -(Qu, Kv) = -(Ku, Qv) \\ &= (Du, K(D^{-1}Q)v) = [u, T_0v]\end{aligned}$$

for all u and v in H_0. Hence, as is known [12], $r(T_0) = r(D^{-1}Q)$ is given by the H_0-norm of T_0, i.e.,

$$r(T_0) = |T_0| = |D^{-1}Q|. \tag{27}$$

Consequently, the definition of a norm of a bounded operator, the relation (27), and the inequalities (23) and (24) imply that

$$\left(\frac{\eta_1}{\eta_2}\right)\|T_0\| \leqq r(T_0) \leqq \left(\frac{\eta_2}{\eta_1}\right)\|T_0\| \tag{28}$$

whence, in view of (9_1), we derive the condition (9). Thus, the rest of Theorem 2 follows from Theorem 1.

REMARK 1. Suppose that A is a self-adjoint operator in H which satisfies the condition (25) and $\alpha>0$. If we take $D \equiv \alpha^{-1}$, $Q \equiv A - \alpha^{-1}$, and $K \equiv I$, then (23) and (24) are satisfied with $\eta_1^2 = \eta_2^2 = \alpha^{-1}$ and $A = D+Q$ and $G \equiv D-Q = 2\alpha^{-1} - A$. This shows that, by virtue of (25), G is positive definite if and only if α satisfies the condition $0<\alpha<2/M$. Furthermore, since $\eta_1 = \eta_2$, (28) reduces to the equality

$$r(T_0) = |T_0| = \|T_0\|. \tag{28_1}$$

Hence, under the hypotheses of Theorem 2(a), the Main Theorem in [8] applied to this case implies the validity of (9_1) and thus the corresponding result of Altman [2] follows as a very special case of Theorem 2.

If H is a finite-dimensional Hilbert space, i.e., a unitary finite-dimensional vector space, then the following corollary to Theorem 2 is valid.

COROLLARY 2. *If H is a finite-dimensional unitary vector space, D and Q are K-symmetric matrices and D is also K-p.d., and $X_0=D^{-1}$, then $\{X_{n+1}\}$ determined by the hyperpower method* (1)–(2) *converges to the inverse of $A\equiv D+Q$ if and only if A and $G\equiv D-Q$ are K-p.d. Furthermore, the error estimate* (8_1) *is also valid here for $n\geqq r>0$. If additionally the eigenvalues of the matrix $D^{-1}Q$ satisfies the condition analogous to* (9_1), *then the error estimates* (10)–(12) *remain valid.*

At the end let us remark that, as was shown in [10], the class of K-symmetric and K-p.d. matrices is equivalent to the class of matrices having positive eigenvalues and a complete set of corresponding eigenvalues or a class of weakly positive matrices or a class of matrices of the form H_1H_2, where H_1 and H_2 are two Hermitian and positive definite matrices.

REFERENCES

1. J. Albrecht, *Bemerkungen zum Iterationsverfahren von Schulz zur Matrixinversion*, Z. Angew. Math. Mech. **41** (1961), 262–263.

2. M. Altman, *An optimum cubically convergent iterative method of inverting a linear bounded operator in Hilbert space*, Pacific J. Math. **10** (1960), 1107–1113.

3. R. Ansorge, *Über ein Iterationsverfahren von G. Schulz zur Ermittlung der Reziproken einer Matrix*, Z. Angew. Math. Mech. **39** (1959), 164–165.

4. W. Dück, *Fehlerabschätzungen für das Iterationsverfahren von Schulz zur Bestimmung der Inversen einer Matrix*, Z. Angew. Math. Mech. **40** (1960), 192–194.

5. F. John, *Advanced numerical analysis*, Lecture Notes, Institute of Math. Sci., New York University, 1956.

6. H. Hotelling, *Some new methods in matrix calculation*, Amer. Math. Statist. **14** (1943), 1–34.

7. M. Newman, *Matrix computations*, Survey of numerical analysis, McGraw-Hill, New York, 1962, pp. 222–254.

8. W. V. Petryshyn, *On the generalized overrelaxation method for operator equations*, Proc. Amer. Math. Soc. **14** (1963), 917–924.

9. ———, *On a class of K-p.d. and non-K-p.d. operators and operator equations*, J. Math. Anal. Appl. **10** (1965), 1–24.

10. ———, *On extrapolated Jacobi or simultaneous displacements method in the solution of matrix and operator equations*, Math. Comp. **19** (1965), 37–55

11. G. Schulz, *Iterative Berechnung der reziproken Matrix*, Z. Angew. Math. Mech. **13** (1933), 57–59.

12. A. E. Taylor, *Introduction to functional analysis*, Wiley, New York, 1957.

UNIVERSITY OF CHICAGO