APPLICATIONES MATHEMATICAE
22,4 (1995), pp. 499–513

K. MOSZYŃSKI (Warszawa)

ON SOLVING LINEAR ALGEBRAIC EQUATIONS WITH AN ILL-CONDITIONED MATRIX

1. Introduction. Consider a system of N linear algebraic equations

$$Ax = b \tag{1.1}$$

with an invertible $N \times N$ real matrix A. The matrix A has spectral decomposition

$$A = \sum_{j=1}^{p} (\lambda_j P_j + N_j) \tag{1.2}$$

where:

- P_j, $j = 1, \ldots, p$, are the spectral projectors,
- N_j, $j = 1, \ldots, p$, are the spectral nilpotents,
- λ_j, $j = 1, \ldots, p$, are the eigenvalues of A.

The following conditions hold:

- $P_k P_l = P_l P_k = \delta_{kl} P_k$,
- $P_k N_l = N_l P_k = \delta_{kl} N_k$ for $k, l = 1, \ldots, p$;
- if $s_k = \dim(P_k \mathbb{R}^N)$, then $s_1 + \ldots + s_p = N$ and $N_j^{s_j} = 0$;
- $\sum_{j=1}^{p} P_j = I_N$, where I_N is the $N \times N$ identity matrix.

It is easy to see that

$$A^{-1} = \sum_{j=1}^{p} \frac{1}{\lambda_j} \left[P_j + \sum_{s=1}^{s_j - 1} \frac{(-1)^s}{\lambda_j^s} N_j^s \right]; \tag{1.3}$$

1991 *Mathematics Subject Classification*: Primary 65F30.
Key words and phrases: linear algebraic systems, matrix decomposition.
Supported by Research Grant No 211689101 of the Polish Scientific Research Council (KBN).

hence, for the solution x of (1.1) we get

$$x = A^{-1}b = \sum_{j=1}^{p} \frac{1}{\lambda_j}\left[P_j b + \sum_{s=1}^{s_j-1} \frac{(-1)^s}{\lambda_j^s} N_j^s b\right]. \tag{1.4}$$

We can read the formula (1.4) as follows: if we consider some class of matrices A, close to symmetric matrices, for example a class for which $\|N_j^s\|/|\lambda_j^s| \le C\|P_j\|/s_j$, $j = 1,\dots,p$; $s = 1,\dots,s_j$, with some not very large constant C independent of s and j, then the part of the matrix A corresponding to the eigenvalues of the smallest moduli has the strongest influence on the solution x. If A is *ill-conditioned*, the influence of large eigenvalues may be negligible, while they will disturb any process of numerical solution of (1.1). In such a case, (approximate) decomposition of A into parts corresponding to eigenvalues of *small* and *large* moduli seems to be useful. We can look at such an operation as a kind of *preconditioning*. But, in general, preconditioning is not the only purpose of decomposing A. Another purpose is to enable *parallel computing*. Clearly, such a decomposition is closely related to (approximate) invariant subspaces of A, or equivalently, to some matrices (approximately) commuting with A.

2. Generalities. Put $X = \mathbb{R}^N$ and $Y = \operatorname{span}\{q_1,\dots,q_r\} \subset X$, where $q_1,\dots,q_r$ are linearly independent elements of X such that $r \le N$. Denote by $Q = [q_1 \ \dots \ q_r]$ the matrix with columns $q_1,\dots,q_r$; it is an $N \times r$ matrix of rank r.

(2.1) Proposition. *For any $N \times N$ matrix U, $Y = UX$ iff there exist linearly independent vectors $f_1,\dots,f_r$ in $X = \mathbb{R}^N$ such that $U = QF$, where*

$$F = \begin{bmatrix} f_1^T \\ \vdots \\ f_r^T \end{bmatrix}.$$

Moreover, $U^2 = U$ (U is a projector) iff $FQ = I_r$. ∎

Let U be a matrix satisfying the conditions of (2.1). Then $U = QF$, and the $r \times r$ matrices Q^TQ and FF^T are both invertible.

If U is *nearly a projector*, then FQ should be at least invertible.

We are interested in matrices (approximately) commuting with A. Assume first that U *exactly* commutes with A:

$$UA - AU = 0,$$

i.e.

$$QFA - AQF = 0 \tag{2.2}$$

and

$$AQ = QC_1, \quad Q = [q_1 \ \dots \ q_r], \tag{2.3}$$

with the $r \times r$ matrix $C_1 = FAF^T(FF^T)^{-1}$; in other words, $Y = \operatorname{span}\{q_1, \dots, q_r\}$ is an invariant subspace of A. From (2.2) we immediately deduce

$$FA = C_2F \tag{2.4}$$

with the $r \times r$ matrix $C_2 = (Q^TQ)^{-1}Q^TAQ$. From (2.3) we get $C_1 = C_2$. Condition (2.4) means that $Z = \operatorname{span}\{f_1, \dots, f_r\}$ is an invariant subspace of A^T.

If FQ is *invertible*, then (2.2) implies

$$AQ = QC_3 \tag{2.5}$$

with $C_3 = FAQ(FQ)^{-1}$, and now from (2.4) it follows that $C_3 = C_2 = C_1$. Finally, if FQ is invertible, then by (2.2) we obtain

$$FA = C_4F \tag{2.6}$$

with $C_4 = (FQ)^{-1}FAQ$. Formula (2.5) together with the previous conditions implies

$$C_1 = C_2 = C_3 = C_4 = C. \tag{2.7}$$

Denote by $\sigma(B)$ the spectrum of a matrix B. It is easy to see that

$$\sigma(C_i) \subset \sigma(A), \quad i = 1, 2, 3, 4. \tag{2.8}$$

In fact, if $(\lambda - C_1)x = 0$ for $x \neq 0$, then $Q(\lambda - C_1)x = 0$ and hence, applying (2.2), we get $(Q\lambda - AQ)x = 0$ or $(\lambda - A)Qx = 0$ with $Qx \neq 0$, because $Q = [q_1 \ \dots \ q_r]$ and $q_1, \dots, q_r$ are linearly independent. In view of (2.8) if A is invertible, then so is $C = C_1 = C_2 = C_3 = C_4$. If U commutes with A, then C contains entire information concerning A relating to the invariant subspace $Y = UX$. This fact is expressed more precisely by the following:

(2.9) Proposition. *If U commutes with A, then*

$$C = \sum_{j=1,\, UP_j \neq 0}^{p} (\lambda_j P_j^C + N_j^C),$$

where

$$P_j^C = (Q^TQ)^{-1}Q^TP_jQ = FP_jF^T(FF^T)^{-1}, \tag{2.10}$$

$$N_j^C = (Q^TQ)^{-1}Q^TN_jQ = FN_jF^T(FF^T)^{-1}, \quad j = 1, \dots, p, \tag{2.10$'$}$$

are the spectral projectors and nilpotents of C.

Proof. Observe first that formulae (2.3)–(2.8) follow from commutativity of U and A. Since U commutes with $(\lambda - A)^{-1}$ (if $(\lambda - A)^{-1}$ exists),

the condition similar to (2.7) holds for $P_j = \frac{1}{2\pi i}\int_\Gamma (\lambda - A)^{-1}\,d\lambda$. Here Γ is a Jordan curve containing only one eigenvalue λ_j in its interior domain, and such that $\sigma(A) \cap \Gamma$ is empty. We then have (2.10). By an easy transformation we prove that U commutes with N_j as well, and (2.10′) follows. Now

$$\begin{aligned} P_k^C P_l^C &= (Q^TQ)^{-1}Q^TP_kQFP_lF^T(FF^T)^{-1} \\ &= (Q^TQ)^{-1}Q^TUP_kP_lF^T(FF^T)^{-1} = (Q^TQ)^{-1}Q^TU\delta_{kl}P_kF^T(FF^T)^{-1} \\ &= \delta_{kl}(Q^TQ)^{-1}Q^TQFP_kF^T(FF^T)^{-1} = \delta_{kl}P_k^C, \qquad k,l = 1,\dots,p, \end{aligned}$$

i.e. P_k^C are spectral projectors. Similarly we can prove that N_j^C are the spectral nilpotents of C. This completes the proof. ∎

Now assume $UA - AU = R$ and $R \neq 0$. Then neither (2.7) nor (2.9) is true. The only thing we may expect is that (2.7) and (2.9) hold approximately if R is small enough (and $(FQ)^{-1}$ exists). More precisely, if U_0, Q_0, F_0 commute with A, $(F_0Q_0)^{-1}$ exists and $U = (Q_0 + \Delta Q)(F_0 + \Delta F)$, then (2.7) and (2.9) hold asymptotically for U as $\|\Delta Q\| \to 0$ and $\|\Delta F\| \to 0$.

Moreover, if $(F_0Q_0)^{-1}$ exists and $\|\Delta Q\|$ and $\|\Delta F\|$ are small enough, then C_j, $j = 1,2,3,4$, are invertible. Since FF^T, Q^TQ, and FQ are invertible, it follows that FAF^T, Q^TAQ, and FAQ are invertible as well.

It may be of some interest to know the inverses of Q^TAQ, FAF^T, and FAQ. It is easy to verify that, under the above assumptions,

$$\begin{aligned} (Q^TAQ)^{-1} &= FA^{-1}F^T(Q^TQFF^T)^{-1}(I - \Delta_1)^{-1} \\ &= FA^{-1}F^T(Q^TQFF^T)^{-1} + O(R) \end{aligned} \tag{2.11}$$

with $\Delta_1 = Q^TRA^{-1}F^T(Q^TQFF^T)^{-1} = O(R)$, and

$$\begin{aligned} (FAF^T)^{-1} &= (I + \Delta_2)^{-1}(Q^TQFF^T)^{-1}Q^TA^{-1}Q \\ &= (Q^TQFF^T)^{-1}Q^TA^{-1}Q + O(R) \end{aligned} \tag{2.12}$$

with $\Delta_2 = (Q^TQFF^T)^{-1}Q^TA^{-1}RF^T = O(R)$; finally, if $(FQ)^{-1}$ exists, then

$$(FAQ)^{-1} = (FA^{-1}Q)(FQ)^{-2}(I - \Delta_3)^{-1} = FA^{-1}Q(FQ)^{-1} + O(R) \tag{2.13}$$

with $\Delta_3 = FRA^{-1}Q(FQ)^{-2} = O(R)$.

Let now σ_0 be *a spectral set*, i.e. $\sigma_0 \subset \sigma(A)$. Consider U of the following form:

$$U = \sum_{\lambda_j \in \sigma_0} P_j + \varepsilon, \tag{2.14}$$

where P_j are spectral projectors of A and ε is a small matrix. We have $R = UA - AU = \varepsilon A - A\varepsilon = O(\varepsilon)$. We are interested in the asymptotic

behaviour of $(Q^TAQ)^{-1}$, $(FAF^T)^{-1}$, and $(FAQ)^{-1}$ as $\varepsilon \to 0$. Notice that:

$$
\begin{aligned}
&F = (Q^TQ)^{-1}Q^TU \quad \text{or}\\
&F = (FQ)^{-1}FU \quad \text{if } (QF)^{-1} \text{ exists, and}\\
&Q = UF^T(FF^T)^{-1} \quad \text{or}\\
&Q = UQ(FQ)^{-1} \quad \text{if } (FQ)^{-1} \text{ exists.}
\end{aligned} \tag{2.15}
$$

Let us consider $(Q^TAQ)^{-1}$ only. The discussion of the remaining inverses is similar. From (2.11) and (2.15) we get

$$
\begin{aligned}
(Q^TAQ)^{-1} &= FA^{-1}F^T(Q^TQFF^T)^{-1} + O(R)\\
&= (Q^TQ)^{-1}Q^TUA^{-1}F^T(Q^TQFF^T)^{-1} + O(R).
\end{aligned}
$$

From (1.2) and (1.3) it follows that

$$
\begin{aligned}
UA^{-1} &= \Big(\sum_{\lambda_j\in\sigma_0} P_j + \varepsilon\Big)\sum_{j=1}^{p}\frac{1}{\lambda_j}\Big[P_j + \sum_{s=1}^{s_j-1}\frac{(-1)^s}{\lambda_j^s}N_j^s\Big]\\
&= \sum_{\lambda_j\in\sigma_0}\frac{1}{\lambda_j}\Big[P_j + \sum_{s=1}^{s_j-1}\frac{(-1)^s}{\lambda_j^s}N_j^s\Big] + O(R).
\end{aligned}
$$

Hence

$$
\begin{aligned}
&(Q^TAQ)^{-1} =\\
&(Q^TQ)^{-1}Q^T\sum_{\lambda_j\in\sigma_0}\frac{1}{\lambda_j}\Big[P_j + \sum_{s=1}^{s_j-1}\frac{(-1)^s}{\lambda_j^s}N_j^s\Big]F^T(Q^TQFF^T)^{-1} + O(\varepsilon),
\end{aligned} \tag{2.16}
$$

i.e. *the principal* component of $(Q^TAQ)^{-1}$ depends on the spectral elements of A related to σ_0. Observe also that, in general, (2.16) is not the spectral decomposition of $(Q^TAQ)^{-1}$.

The matrices Q^TAQ and FAQ will play a very important role in the method presented below. FAF^T plays a similar role to Q^TAQ, but for equations with the transposed matrix A^T.

3. Approximate decomposition of (1.1). Consider now a matrix $U = QF$, where $Q = [q_1 \ \dots \ q_r]$ is a matrix of rank r, $r \le N$, such that $Y = \operatorname{span}\{q_1, \dots, q_r\}$ is a *sufficiently good* approximation of some invariant subspace of A, related to a spectral set σ_0. We are going to decompose the system (1.1) into two parts corresponding to σ_0 and to $\sigma(A)\backslash\sigma_0$. Multiplying (1.1) from the left by U, we get a new system:

$$
UAx = Ub. \tag{3.1}
$$

Since $UA = AU + R$ (with R small), we have $AUx + Rx = Ub$, or

$$AQy + Rx = Ub, \tag{3.2}$$

where $y = Fx$.

Observe that, in general, (3.1) has many solutions (x is one of them), while (3.2), regarded as a system with unknown y and x given, has exactly one solution $y = Fx$. This follows from the fact that the $N \times r$ matrix AQ has maximal possible rank r. If we multiply the solution y of (3.2) by Q we get

$$Qy = QFx = Ux.$$

This is exactly the component of the solution $x = A^{-1}b$ of (1.1) related to U (or, in other words, to the spectral set $\sigma_0 \subset \sigma(A)$).

Assume Q^TAQ to be invertible. Multiplying now (3.2) from the left by Q^T, we obtain a system equivalent to (3.2):

$$Q^TAQy + Q^TRx = Q^TUb. \tag{3.3}$$

Another possibility is to multiply (3.2) from the left by F:

$$FAQy + FRx = FUb. \tag{3.4}$$

If FAQ is invertible (see Section 2) then (3.4) and (3.2) are equivalent. Both systems (3.3) and (3.4) are of dimension $r \times r$ and both satisfy the condition

$$Ux = Qy. \tag{3.5}$$

Algorithms for computing the matrices Q^TAQ, FAQ and the vectors Q^TUb, FUb will be proposed in Section 4.

We may stop here if we only want to have an *approximate* vector y (or Ux), under the assumption that R is *sufficiently small*. In order to compute y we can solve one of the systems

$$Q^TAQv = Q^TUb \tag{3.6}$$

or

$$FAQw = FUb, \tag{3.7}$$

provided that the corresponding $r \times r$ matrix Q^TAQ or FAQ is invertible (see Section 2).

To estimate the errors $\|v - y\|/\|y\|$ and $\|w - y\|/\|y\|$ we can apply the well known inequality given in

(3.8) Lemma. *Let B be an invertible matrix and consider two systems of linear algebraic equations*

$$Bu = d \quad \textit{and} \quad (B + E)(u + \Delta) = d + \delta.$$

Then

$$\frac{\|\Delta\|}{\|u\|} \le \frac{\operatorname{cond}(B)\left(\frac{\|\delta\|}{\|d\|} + \frac{\|E\|}{\|B\|}\right)}{1 - \operatorname{cond}(B)\frac{\|E\|}{\|B\|}}$$

provided that $\operatorname{cond}(B)\|E\|/\|B\| < 1$, *where* $\operatorname{cond}(B) = \|B\|\,\|B^{-1}\|$.

P r o o f. By simple verification. ∎

Applying now (3.8) to (3.6) and (3.7), we get

$$\begin{aligned} &\|v - y\|/\|y\| \le \operatorname{cond}(Q^T AQ)\|Q^T Rx\|/\|Q^T Ub\| = O(R), \\ &\|w - y\|/\|y\| \le \operatorname{cond}(FAQ)\|FRx\|/\|FUb\| = O(R). \end{aligned} \tag{3.9}$$

Observe that if A is ill-conditioned and $\sigma_0 \subset \sigma(A)$ is properly chosen, then we should expect that

$$\operatorname{cond}(Q^T AQ) < \operatorname{cond}(A), \quad \operatorname{cond}(FAQ) < \operatorname{cond}(A).$$

Moreover, sometimes it is possible to get a *full decomposition of* (1.1) corresponding to the decomposition of the spectrum

$$\sigma(A) = \sigma_0 \cup (\sigma(A)\backslash\sigma_0).$$

Assume that, as above, the matrix $U = QF$ is known; then we can also use the matrix $I - U$. If U is a projector of rank r, then $I - U$ is a projector of rank $N - r$. This is the most interesting situation. In general, if U is not a projector, then $I - U$ is a matrix of rank s, where $N - r \le s \le N$. Let

$$I - U = SG, \tag{3.10}$$

where S is an $N \times s$ matrix of rank s $(N - r \le s \le N)$ and G is an $s \times N$ matrix of rank s. If U is a projector then $s = N - r$. Observe that

$$(I - U)A - A(I - U) = -R, \tag{3.11}$$

hence multiplying now (1.1) from the left by U and by $I - U$, and taking into account (3.10) and (3.11), we get

$$AQy + Rx = Ub \quad \text{and} \quad ASz - Rx = (I - U)b,$$

where $y = Fx$ and $z = Gx$ with $x = A^{-1}b$. Since $x = Ux + (I - U)x = Qy + Sz$, we can write

$$\begin{cases} AQy + RQy + RSz = Ub, \\ ASz - RSz - RQy = (I - U)b, \end{cases} \tag{3.12}$$

and, finally, multiplying (3.12) by Q^T and S^T, we obtain

$$\begin{cases} Q^T AQy + Q^T RQy + Q^T RSz = Q^T Ub, \\ S^T ASz - S^T RSz - S^T RQy = S^T(I - U)b. \end{cases} \tag{3.13}$$

Another possibility is to multiply (3.12) by F and G to get

$$\begin{cases} FAQy + FRQy + FRSz = FUb, \\ GASz - GRSz - GRQy = G(I-U)b. \end{cases} \tag{3.14}$$

Both systems (3.13) and (3.14) are of dimension $s+r$, $N \le s+r$. We shall prove:

(3.15) THEOREM. (a) *If Q^TAQ and S^TAS are invertible and $\|R\|$ is small enough, then* (3.13) *and* (1.1) *are equivalent.*

(b) *If FAQ and GAS are invertible and $\|R\|$ is small enough, then* (3.14) *and* (1.1) *are equivalent.*

P r o o f. We shall prove (a). The proof of (b) is analogous. We have to show that:

(i) $\left[\begin{smallmatrix} y \\ z \end{smallmatrix}\right]$ is a solution of (3.13), where $y = Fx$, $z = Gx$, and $x = A^{-1}b$.

(ii) (3.13) has a unique solution.

To verify (i), insert y and z into the first equation of (3.13) to get

$$\begin{aligned} (Q^TAQF + Q^TRQF + Q^TRSG)x - Q^TUb \\ &= Q^T\{[AU + RU + R(I-U)]x - Ub\} \\ &= Q^T\{[AU + R]x - Ub\} = Q^TU[Ax - b] = 0. \end{aligned}$$

Similar calculations show that the second equation of (3.13) is also satisfied.

(ii) will be proved if we show that the matrix of (3.13) is invertible. This matrix has the following block form:

$$\begin{aligned} &\begin{bmatrix} Q^TAQ + Q^TRQ & Q^TRS \\ -S^TRQ & S^TAS - S^TRS \end{bmatrix} \\ &= \begin{bmatrix} Q^TAQ(I + (Q^TAQ)^{-1}Q^TRQ) & Q^TRS \\ -S^TRQ & S^TAS(I - (S^TAS)^{-1}S^TRS) \end{bmatrix}. \end{aligned} \tag{3.16}$$

Observe now that, in general, the matrix

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix}$$

is invertible if A_{11}^{-1} and A_{22}^{-1} exist and $\|A_{12}\|$ and $\|A_{21}\|$ are small enough. In fact, we have to find a matrix

$$\begin{bmatrix} B_{11} & B_{12} \\ B_{21} & B_{22} \end{bmatrix}$$

such that

$$\begin{aligned} A_{11}B_{11} + A_{12}B_{21} = I, \quad A_{11}B_{12} + A_{12}B_{22} = 0, \\ A_{21}B_{11} + A_{22}B_{21} = 0, \quad A_{21}B_{12} + A_{22}B_{22} = I. \end{aligned} \tag{3.17}$$

The first block column of (3.17) yields

$$(I - A_{11}^{-1}A_{12}{A_{22}}^{-1}A_{21})B_{11} = A_{11}^{-1} \quad \text{and} \quad B_{21} = -A_{22}^{-1}A_{21}B_{11}.$$

If $\|A_{12}\|$ and $\|A_{21}\|$ are small enough, the matrix in brackets in the first formula is invertible; hence B_{11} and B_{21} are well defined. The same argument applies to the second block column of (3.17). In particular, for suitable $A_{ij}, i, j = 1, 2$, we obtain (a). ∎

Observe that the matrix R can be expressed by means of A, Q, S, F, G. To make use of the decomposition (3.13) or (3.14) of the system (1.1) one can apply the following iterative procedure: we start with an arbitrary pair of vectors $y_0 \in \mathbb{R}^r$ and $z_0 \in \mathbb{R}^s$; then the consecutive vectors y_{k+1} and z_{k+1} are computed from the system of algebraic equations

$$\begin{cases} Q^TAQy_{k+1} + Q^TRQy_k + Q^TRSz_k = Q^TUb, \\ S^TASz_{k+1} - S^TRSz_k - S^TRQy_k = S^T(I-U)b \end{cases} \tag{3.18}$$

(for (3.13)), or

$$\begin{cases} FAQy_{k+1} + FRQy_k + FRSz_k = FUb, \\ GASz_{k+1} - GRSz_k - GRQy_k = G(I-U)b \end{cases} \tag{3.19}$$

(for (3.14)).

Observe that each iteration step using (3.18) or (3.19) consists in solving two independent systems of dimension r and s with matrices Q^TAQ and S^TAS, or FAQ and GAS respectively. If the decomposition is done properly, then the *conditioning* of these systems should be much better than that of the original system (1.1). Moreover, the two systems admit *parallel* solution.

Now let us transform slightly the systems (3.18) and (3.19) to obtain new systems, more convenient for computations. If we express R in terms of Q, F, and S, G, then we easily obtain a new form of (3.18):

$$\begin{cases} Q^TAQv_{k+1} = Q^TUr_k, \\ S^TASw_{k+1} = S^T(I-U)r_k, \end{cases} \tag{3.20}$$

where

$$x_k = Qy_k + Sz_k, \qquad v_{k+1} = y_{k+1} - Fx_k,$$
$$r_k = b - Ax_k, \qquad w_{k+1} = z_{k+1} - Gx_k.$$

An analogous transformation applied to (3.19) gives

$$\begin{cases} FAQv_{k+1} = FUr_k, \\ GASw_{k+1} = G(I-U)r_k, \end{cases} \tag{3.21}$$

with the same definitions of x_k, r_k, v_{k+1}, and w_{k+1}. In the next section we discuss the algorithms of computation of the entries in (3.20) and (3.21). We also present possible simplifications of (3.20) and (3.21). The following theorem makes use of equations (3.13) or (3.14), which are more satisfactory for theoretical investigations than (3.20) and (3.21) respectively.

(3.22) THEOREM. (a) *If Q^TAQ and S^TAS are invertible, and*

$$\operatorname{cond}(Q^TAQ)\frac{\|R\|}{\|Q^TAQ\|} + \operatorname{cond}(S^TAS)\frac{\|R\|}{\|S^TAS\|}$$

is small enough, then for any y_0 and z_0 the process (3.20) *converges to the solution $x = A^{-1}b$ of* (1.1).

(b) *If FAQ and GAS are invertible, and*

$$\operatorname{cond}(FAQ)\frac{\|R\|}{\|FAQ\|} + \operatorname{cond}(GAS)\frac{\|R\|}{\|GAS\|}$$

is small enough, then for any y_0 and z_0 the process (3.21) *converges to the solution $x = A^{-1}b$ of* (1.1).

Proof. We prove (a) only. The proof of (b) is analogous. Observe that the iterative process under consideration is of the general form

$$\begin{bmatrix} \widetilde{A} & 0 \\ 0 & \widetilde{B} \end{bmatrix} \begin{bmatrix} y_{k+1} \\ z_{k+1} \end{bmatrix} + \begin{bmatrix} \Delta_A & C \\ D & \Delta_B \end{bmatrix} \begin{bmatrix} y_k \\ z_k \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},$$

while the system of equations we are going to solve (see (3.15)) is

$$\begin{bmatrix} \widetilde{A} + \Delta_A & C \\ D & \widetilde{B} + \Delta_B \end{bmatrix} \begin{bmatrix} y \\ z \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \end{bmatrix},$$

with some matrices $\widetilde{A}$, $\widetilde{B}$, Δ_A, Δ_B, C, D, and vectors b_1 and b_2.

Let $e_{y_k} = y - y_k$ and $e_{z_k} = z - z_k$; then

$$\begin{bmatrix} e_{y_{k+1}} \\ e_{z_{k+1}} \end{bmatrix} = -\begin{bmatrix} \widetilde{A}^{-1}\Delta_A & \widetilde{A}^{-1}C \\ \widetilde{B}^{-1}D & \widetilde{B}^{-1}\Delta_B \end{bmatrix} \begin{bmatrix} e_{y_k} \\ e_{z_k} \end{bmatrix} = T \begin{bmatrix} e_{y_k} \\ e_{z_k} \end{bmatrix}.$$

The norm of the iteration matrix T can be easily estimated in the standard way:

$$\|T\| \le \|\widetilde{A}^{-1}\Delta_A\| + \|\widetilde{A}^{-1}C\| + \|\widetilde{B}^{-1}D\| + \|\widetilde{B}^{-1}\Delta_B\|.$$

Since in our case $\widetilde{A} = Q^TAQ$, $\widetilde{B} = S^TAS$, $\Delta_A = Q^TRQ$, $\Delta_B = -S^TRS$, $C = Q^TRS$, $D = -S^TRQ$, we immediately get a sufficient condition for convergence of the process:

$$\|T\| \le K\left[\operatorname{cond}(Q^TAQ)\frac{\|R\|}{\|Q^TAQ\|} + \operatorname{cond}(S^TAS)\frac{\|R\|}{\|S^TAS\|}\right] < 1$$

with some constant K depending on Q and S. This completes the proof of (a). ∎

4. Algorithms. In this section we propose two algorithms giving entries for the iterative processes (3.20) and (3.21). As before let U be a matrix related to some spectral set $\sigma_0 \subset \sigma(A)$, *approximately* commuting with A. Certain suggestions on construction of such a matrix U are discussed in Section 5.

Gram–Schmidt Process. The Gram–Schmidt Process is suitable for equation (3.20). We consider two applications of this process.

A. We try to express the matrix U (of rank $r < N$) in the form $U = QF$, where Q is an $N \times r$ matrix with orthonormal columns:

$$Q^T Q = I_r. \tag{4.1}$$

The splitting $U = QF$ *is exactly the Gram–Schmidt Process*, applied to the consecutive columns of U. In this case, F is an upper triangular matrix of the *coefficients of the Gram–Schmidt Process.* The same algorithm should be applied to $I - U$: $I - U = SG$ and $S^T S = I_s$. The equations obtained are a little simpler than (3.20), namely,

$$\begin{cases} Q^T A Q v_{k+1} = F r_k, \\ S^T A S w_{k+1} = G r_k, \end{cases} \tag{4.2}$$

with x_k, r_k, r_{k+1}, and w_{k+1} as defined after (3.20) (because $Q^T U = Q^T Q F = F$, and $S^T(I - U) = S^T S G = G$).

B. We modify the Gram–Schmidt Process, applied also to the consecutive columns of $U = QF$. Now instead of the orthogonality assumption (4.2), we impose the condition

$$Q^T A Q = I_r.$$

The matrix F is again upper triangular. As before, the same algorithm should be applied to $I - U$: $I - U = SG$ and $S^T A S = I_s$. If we succeed in this operation (this is always possible when A is positive definite), the iterative process (3.20) will be *explicit*:

$$\begin{cases} y_{k+1} = F x_k + Q^T U r_k, \\ z_{k+1} = G x_k + S^T (I - U) r_k, \end{cases} \tag{4.3}$$

with $x_k = Q y_k + S z_k$, $r_k = b - A x_k$ or, equivalently,

$$x_{k+1} = x_k + [Q Q^T U + S S^T (I - U)] r_k \tag{4.4}$$

with $x_k \to x$ as $k \to \infty$ under the assumptions of Theorem (3.22).

Lanczos Process. To define the entries for the iterative process (3.21) we can apply the Lanczos Process to the matrix UA: we find an $N \times r$ matrix Q with orthonormal columns and an $r \times r$ lower Hessenberg matrix T (quasi-triangular, i.e. triangular with one additional diagonal, nearest to the main diagonal) such that

$$UAQ = QT^T, \quad Q^T Q = I_r.$$

In such a way we obtain an orthonormal basis of the space

$$Y = U\mathbb{R}^N = \operatorname{span}\{q_1, \dots, q_r\},$$

where $Q = [q_1 \ \dots \ q_r]$. Hence $U = QF$ for some matrix F. Moreover, $QFAQ = QT^T$ and the orthogonality condition $Q^TQ = I_r$ gives

$$FAQ = T^T.$$

In other words, the Lanczos Process defines directly the matrices Q and $FAQ = T^T$; then FAQ is an upper Hessenberg matrix. If A is symmetric, then $T^T = FAQ$ is symmetric tridiagonal. The analogous procedure applied to the matrix $(I - U)A = SGA$ gives

$$(I - U)AS = SZ^T$$

with a lower Hessenberg matrix Z, and S satisfying $S^TS = I_s$. Hence $GAS = Z^T$ is also an upper Hessenberg matrix. In this way we have determined the principal entries for the process (3.21).

Now we only need to transform the right hand side of (3.21). Observe that $F = Q^TQF = Q^TU$ and $G = S^TSG = S^T(I - U)$. Hence (3.21) becomes

$$\begin{cases} T^Tv_{k+1} = Q^TU^2r_k, \\ Z^Tw_{k+1} = S^T(I - U)^2r_k. \end{cases} \tag{4.5}$$

Both subsystems of (4.5) are Hessenberg (tridiagonal if A is symmetric).

There is another known version of the Lanczos Process wich results in tridiagonal T and Z for any matrix A. However, this version is considered to be less stable.

We are looking for $N \times r$ matrices Q_1 and Q_2 such that

$$\begin{cases} UAQ_1 = Q_1T_1^T, \\ A^TU^TQ_2 = Q_2T_2^T, \end{cases} \tag{4.6}$$

and

$$Q_2^TQ_1 = Q_1^TQ_2 = I_r. \tag{4.7}$$

Since $U = Q_1F$, (4.6) and (4.7) imply

$$Q_2^TUAQ_1 = T_1^T = T_2 = FAQ_1.$$

Since both T_1 and T_2 are lower Hessenberg matrices, it follows that

$$\begin{cases} UAQ_1 = Q_1T^T, \\ A^TU^TQ_2 = Q_2T, \end{cases} \tag{4.8}$$

with $Q_2^TQ_1 = Q_1^TQ_2 = I_r$ and $T = Q_1^TA^TU^TQ_2$ tridiagonal. From the condition $U = Q_1F$ we get $F = Q_2^TU$.

Applying the similar procedure to $I - U$ with S_1, S_2, and Z in place of Q_1, Q_2, and T, we obtain the *tridiagonal version* of (3.21):

$$\begin{cases} T^Tv_{k+1} = Q_2^TU^2r_k, \\ Z^Tw_{k+1} = S^T(I - U)^2r_k, \end{cases} \tag{4.9}$$

with x_k, y_k, z_k, v_k, w_k defined as before.

5. The matrix U. Certainly there are many possibilities of constructing a suitable matrix U. Here we give some remarks on applications of polynomials of the matrix A. It seems that nice results can be obtained in the case of A diagonalizable and with *real spectrum* $\sigma(A)$, using *Bernstein-like polynomials* approximating step functions (see [3]).

Let W_n be a polynomial of degree n. Then, for $A = \sum_{j=1}^{p}(\lambda_j P_j + N_j)$, we have

$$W_n(A) = \sum_{j=1}^{p}\left[P_j W_n(\lambda_j) + \sum_{s=1}^{n+1}\frac{W_n^{(s)}(\lambda_j)}{s!}N_j^s\right]. \tag{5.1}$$

Let $\Omega \subset \mathbb{R}$ be an interval containing $\sigma(A)$; let Ω_1 be a subset of Ω and χ_{Ω_1} the characteristic function of Ω_1. Assume that $W_n^{(k)}(z) \to \chi_{\Omega_1}^{(k)}(z)$ as $n \to \infty$ at any point $z \in \mathbb{R}$ of continuity of χ_{Ω_1}, and that $\sigma(A)\cap\partial\Omega_1\cap\Omega = \emptyset$. Then $W_n(\lambda_j) \to 1$ as $n \to \infty$ for $\lambda_j \in \sigma(A)\cap\Omega_1$, $W_n(\lambda_j) \to 0$ as $n \to \infty$ for $\lambda_j \in \sigma(A) \cap (\Omega\backslash\Omega_1)$, while $W_n^{(s)}(\lambda_j) \to 0$ as $n \to \infty$ for $\lambda_j \in \sigma(A)$, $s > 0$. In other words, $U_n = W_n(A) \to \sum_{\lambda_j\in\Omega_1} P_j$ as $n \to \infty$.

As an example, consider the following sequence of *Bernstein-like polynomials* $B_n(x_1, x_2, \lambda)$ (see also [3]), which can be applied when A *has real spectrum* in $[-1, 1]$. The function $B_n(x_1, x_2, \lambda)$ of three real variables: x_1, x_2, $-1 < x_1 < x_2 < 1$, and λ, is a polynomial of degree n with respect to λ:

$$B_n(x_1, x_2, \lambda) = \sum_{j_n(x_1)\le j\le j_n(x_2)} \binom{n}{j}\left(\frac{1+\lambda}{2}\right)^j\left(\frac{1-\lambda}{2}\right)^{n-j}$$

where $j_n(x) = \frac{1}{2}n(1 + x)$, $|x| < 1$.

The graph of $B_{31}(-0.5, 0.5, \lambda)$ is shown in Figure 1.

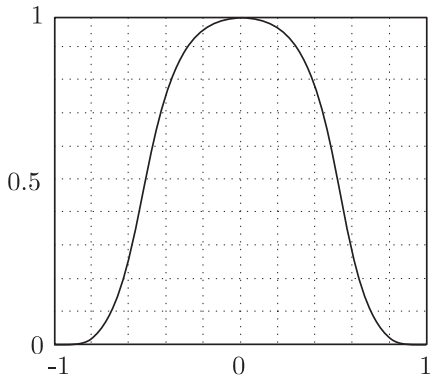

Fig. 1. The graph of $B_{31}(-0.5, 0.5, \lambda)$

Define $U = B_n(x_1, x_2, A)$ with fixed x_1, x_2 and n. This matrix corresponds to the spectral set $\sigma_0 = [x_1, x_2] \cap \sigma(A)$ and *approximately cuts off* the part of the spectrum $\sigma(A)$ contained in $[-1, x_1] \cup [x_2, 1]$.

A good method to compute values of the polynomial $B_n(x_1, x_2, \lambda)$ is to apply the so-called Newton formula:

$$B_n(x_1, x_2, \lambda) = b_0^n(x) + b_1^n(x)(\lambda - \lambda_1) + b_2^n(x)(\lambda - \lambda_1)(\lambda - \lambda_2) + \ldots + b_n^n(x)(\lambda - \lambda_1)\ldots(\lambda - \lambda_n).$$

The coefficients $b_j^n(x)$ are divided differences of $B_n(x, \cdot)$. The suitable choice of the knots $\lambda_1, \ldots, \lambda_n$ was discussed for example in [2] and [3].

Below we give tables of Chebyshev knots and coefficients b_j^{31}, $j = 0, 1, 2, \ldots, 31$, for $B_{31}(-0.5, 0.5, \lambda)$. In this case, only 15 (even) coefficients do not vanish.

Chebyshev knots for $n = 31$	**The coefficients** b_j^{31} **of** $B_{31}(-0.5, 0.5, \lambda)$
$\lambda_1 = .998795456205172$	$b_0^{31} = 0.0$
$\lambda_3 = .949067674327418$	$b_2^{31} = -1.0$
$\lambda_5 = .740951125354959$	$b_4^{31} = 1.667$
$\lambda_7 = .671558954847018$	$b_6^{31} = -0.5881$
$\lambda_9 = .941544065183021$	$b_8^{31} = -2.2957$
$\lambda_{11} = .336889853392220$	$b_{10}^{31} = 11.4375$
$\lambda_{13} = .903989293123443$	$b_{12}^{31} = -4.3486$
$\lambda_{15} = .427555093430282$	$b_{14}^{31} = -34.0687$
$\lambda_{17} = .989176509964781$	$b_{16}^{31} = 29.9361$
$\lambda_{19} = .148730474455362$	$b_{18}^{31} = 33.4487$
$\lambda_{21} = .803207531480645$	$b_{20}^{31} = -36.0957$
$\lambda_{23} = .595699304492433$	$b_{22}^{31} = -18.8784$
$\lambda_{25} = .970031253194544$	$b_{24}^{31} = 20.4324$
$\lambda_{27} = .242980179903264$	$b_{26}^{31} = 9.2076$
$\lambda_{29} = .857728610000272$	$b_{28}^{31} = 1.1829$
$\lambda_{31} = .514102744193222$	$b_{30}^{31} = 0.3135$
$*\quad *$	The b_j^{31} for j odd all vanish
$\lambda_{2j} = -\lambda_{2j+1}$	

6. Final remarks. The crucial point of the method discussed above is the decomposition $U = QF$ and $I - U = GS$, when U is given. All processes presented here generate a kind of orthogonal basis of the space $Y = U\mathbb{R}^N$ (the columns of the matrix Q). When dealing with the imple-

mentation of the numerical process of splitting $U = QF$, it is important to incorporate *some criteria* in this implementation. These criteria should enable us to decide which elements of the basis under construction are nearly linearly dependent on those already accepted. Such elements have to be rejected. Only after accepting or rejecting consecutive elements of the basis, the matrix U is *really defined by means of its factors* Q *and* F. At this point a *non-polynomial intervention* occurs. It seems that both processes of orthogonalization (Gram–Schmidt and Lanczos) are suitable to this end.

Let us remark at the end that the choice of U with $U^2 = U$ (a projector) is favorable. This condition implies that $s = N - r$; hence the decomposed systems (3.20) and (3.21) are both of dimension N. Possibilities of construction of projectors U will be discussed elsewhere.

References

[1] G. G. Lorentz, *Bernstein Polynomials*, University of Toronto Press, 1953.

[2] V. I. Lebedev and S. A. Finogenov, *On the order of choosing parameters in Chebychev cyclic iterative process*, Zh. Vychisl. Mat. i Mat. Fiz. 11 (1971), 425–438 (in Russian).

[3] K. Moszyński, *Bernstein polynomials and eigenvalue problems*, report of KBN Grant No 211689191, Department of Mathematics, Computer Science and Mechanics, University of Warsaw, 1992.

KRZYSZTOF MOSZYŃSKI
DEPARTMENT OF MATHEMATICS, COMPUTER SCIENCE AND MECHANICS
UNIVERSITY OF WARSAW
BANACHA 2
02-097 WARSZAWA, POLAND
E-mail: KMOSZYNS@MIMUW.EDU.PL

Received on 21.2.1994