SIAM J. MATRIX ANAL. APPL.
Vol. 13, No. 2, pp. 567–593, April 1992

009

REDUCTION TO TRIDIAGONAL FORM AND MINIMAL REALIZATIONS*

BERESFORD N. PARLETT†

Abstract. This paper presents the theoretical background relevant to any method for producing a tridiagonal matrix similar to an arbitrary square matrix. Gragg's work on factoring Hankel matrices and the Kalman–Gilbert structure theorem from systems theory both find a place in the development.

Tridiagonalization is equivalent to the application of the generalized Gram–Schmidt process to a pair of Krylov sequences. In Euclidean space proper normalization allows one to monitor a tight lower bound on the condition number of the transformation. The various possibilities for breakdown find a natural classification by the ranks of certain matrices.

The theory is illustrated by some small examples and some suggestions for restarting are evaluated.

Key words. Lanczos algorithm, linear systems theory, tridiagonal form, minimal realizations

AMS(MOS) subject classifications. 65F15, 93B10, 93C75

1. Introduction and summary.

No one has presented a finite algorithm that is guaranteed to compute a tridiagonal matrix similar to an arbitrary given square complex matrix while avoiding huge intermediate quantities. Section 2 presents a brief sketch of the history of these attempts, along with a parameterization of the possible tridiagonals.

This paper describes theoretical results that are relevant to any method for producing a tridiagonal representation. In particular, the exceptional parameter values, for which the reduction fails, fall into two classes that we call curable and incurable. Cures come with acceptance of a block tridiagonal form. One impetus for this study was the desire to explain the intriguing observation of Taylor, in his dissertation [26], that incurable breakdown is a blessing in disguise because every eigenvalue of the tridiagonal matrix at breakdown is an eigenvalue of the original matrix despite the failure to find any invariant subspace. A satisfactory explanation comes from the canonical structure theorem of linear systems theory and we thank J.W. Demmel for pointing out that we had actually rediscovered that theorem. However the main goal of this essay is to provide the "right" setting for discussing any attempt to produce a stable algorithm for a tridiagonal representation.

Tridiagonal matrices are associated with three-term recurrence relations and with systems of orthogonal polynomials. The literature on these classical topics is vast. Analysts have studied the moment problem, control theorists have studied the sequence of impulse responses for time-invariant linear dynamical systems, and approximation theorists have studied continued fractions. The focus of all these studies is quite different from our goal of reducing a matrix to tridiagonal form but, as indicated above, some of their results help us to answer our questions. To keep this essay to a reasonable length, we have refrained from pointing out connections to the moment problem and to orthogonal polynomials. However, the recent interest in polynomials

* Received by the editors January 29, 1990; accepted for publication (in revised form) October 29, 1990. This paper was completed while the author visited the Numerical Analysis Group of the Oxford University Computing Laboratory, Oxford, U.K. This work was partially supported by Office of Naval Research contract N00014-90-J-1372.

† Department of Mathematics and the Computer Science Division of the Electrical Engineering and Computer Science Department, University of California, Berkeley, California 94720 (parlett@math.berkeley.edu).

orthogonal with respect to an indefinite inner product and to the associated modified moment problem can be thought of as an ascent of the same mountain range that we approach, but by a different route. Matrix factorizations arrived late on the mathematical scene but surely deserve a place alongside the traditional problems mentioned above. A few recent references are [14], [3], [11], and [18]. We do point out some connections with [14] in the final section.

Much of the theory is pure linear algebra and is independent of norms and angles. For this reason, at the risk of seeming pedantic, we make a distinction between $\mathbb{C}^n$ (complex column vectors) and its dual space $\mathbb{C}_*^n$ of linear functionals on $\mathbb{C}^n$ (row vectors). In this setting, the Gram–Schmidt (GS) process is seen as a method for producing a basis in $\mathbb{C}^n$ and a dual basis in $\mathbb{C}_*^n$. This approach is standard in control theory and although numerical analysts seem to favor the term bi-orthogonal over dual, the notion of angle (and hence right angle) is not needed for theoretical purposes. Nevertheless, a plain vector space is an inadequate setting for numerical analysis. A norm is needed to distinguish bad bases from good ones. In §10 we show how to monitor the condition number of the similarity transformation when Euclidean space is an appropriate setting.

The perspective we have reached, after several revisions, is indicated in the following synopsis of the rest of the essay. After a little history, §2 shows the representation of the class of similar tridiagonals by vector pairs and also urges the use of a pair $(\hat{T}, \Omega)$, with $\hat{T}$ symmetric tridiagonal and Ω diagonal, rather than a single matrix $\Omega^{-1}\hat{T}$. Section 3 presents the GS process, GS factorization, and a new extended GS algorithm to overcome breakdown. Section 4 presents the basic ideas of systems theory and the associated Hankel matrices $H^{(1)}$ and $H^{(0)}$. The rank of $H^{(0)}$ and those of related Krylov matrices give a nice characterization of the exceptional vector pairs. That is the end of the preparation. Section 5 discusses triangular factorization of Hankel matrices and the fundamental result for successful reduction: the three pencils $(H_n^{(1)}, H_n^{(0)})$, (B, I_n), and $(\hat{T}_n, \Omega_n)$ are equivalent. Here B is the given $n \times n$ matrix. In different words this result says that tridiagonal reduction is equivalent to the GS process applied to two Krylov sequences. The next two sections concern failure. Section 6 presents Gragg's result on block triangular factorization and mentions the interesting result of Kailath and his coworkers on Schur complements in Hankel matrices. Section 7 explains Taylor's observation and shows that incurable breakdown at step j occurs only when $(\hat{T}_j, \Omega_j)$ is a so-called minimal realization of the transfer function associated with the "initial" vectors. Section 8 is a table that summarizes the theory and §9 presents some small examples. Section 10 shows that, with proper normalization, the matrix Ω reveals the condition number of the transformations. It follows that, in all cases, there is a straightforward algorithm that can force stability and produce a block tridiagonal $\Omega^{-1}\hat{T}$, each of whose eigenvalues is an eigenvalue of B. This stable reduction is mentioned in §11, along with comments of a practical nature that are inspired by the preceding theory. Section 12 puts the results in perspective, makes connections with other approaches, and points to work left undone.

Finally, we point out here that a straightforward attempt to explain breakdown using the Jordan form of B becomes heavily burdened with irrelevant complications. The geometric approach of the Kalman–Gilbert structure theorem provides just the right level of abstraction; the controllable, observable subspace is easy to define and all we need is its dimension. However, a direct attack on the breakdown problem involves finding a basis for this subspace, explicitly or implicitly, and that can be

very complicated.

1.1. Notation. With few exceptions we follow Householder notational conventions: i, j, k, l, m, n for indices; lower case greek for scalars; lower case roman for column vectors; upper case roman for matrices; script upper case for vector spaces.

In addition,

- $\mathbb{C}$ denotes the complex numbers, $\mathbb{C}^{m\times n}$ the space of $m \times n$ complex matrices, $\mathbb{C}^n = \mathbb{C}^{n\times 1}$, $\mathbb{C}^n_* = \mathbb{C}^{1\times n}$.
- If $x \in \mathbb{C}^n$ then x^t denotes its transpose and x^* denotes $(\bar{x})^t$, the conjugate transpose of x.
- The identity matrix $I_n = [e_1, e_2, \cdots, e_n]$, $\tilde{I}_n = [e_n, e_{n-1}, \cdots, e_1]$.
- $\mathbf{0}$ denotes the zero vector in $\mathbb{C}^n$.
- $\mathbb{N}$ denotes the set of natural numbers.
- range (B) denotes the span of B's columns.

2. Tridiagonal form: History and basics.

DEFINITION. A matrix is *tridiagonal* if the (i, j) entry vanishes whenever $|i - j| > 1$.

A consequence of Galois' theory [1] is that there is no finite algebraic procedure for computing the eigenvalues of an $n\times n$ matrix, real or complex, in the general case when $n \geq 5$. Since diagonal and bidiagonal matrices reveal their eigenvalues immediately, the tridiagonal form is the most compact representation that can be expected from a finite process invoking the four basic arithmetic operators and the extraction of roots. It may turn out that this form is too sparse to be achieved in a stable way for the more difficult cases. Nevertheless, there are infinitely many tridiagonal matrices in the similarity class of a given matrix and a parametric representation of them is given later in this section.

Here is a brief history of the search for satisfactory algorithms to reduce a matrix by similarity transformations $B \rightarrow SBS^{-1}$ to tridiagonal form. In [21] Lanczos presented his method of "minimized iterations" that applied to real symmetric matrices and self-adjoint linear operators. However, he also showed the natural generalization to arbitrary square matrices and noted that this general process can break down. Even the symmetric version was sensitive to the effects of roundoff error and the nonsymmetric version received little serious implementation until the 1980s. In 1954, Givens presented a method for reducing a full $n \times n$ real symmetric (or complex Hermitian) matrix to tridiagonal form using plane rotations and fewer than n^3 scalar multiplications. Attempts to generalize this method appeared in [4], [25], and [21]. Demonstrations and explanations of their instabilities appeared in [28] and [23]. However the search lost its sense of importance in the early 1960s with the general acceptance of the Householder/QR method [5], [6] as a fast, stable solution to the nonsymmetric eigenvalue problem for dense matrices. The Householder reduction to upper Hessenberg form (the (i, j) entry vanishes whenever $i - j > 1$) requires only $(5/3)n^3$ multiplications, and the QR phase that diminishes the subdiagonal entries in positions $(i + 1, i)$ requires about $8n^3$ multiplications in practice, though infinitely many in exact arithmetic.

Now consider the fact that the fastest-known way to reduce a full matrix to tridiagonal form (the Lanczos algorithm) requires little more than $2n^3$ multiplications, if it does not break down. So the potential reduction in arithmetic cost from use of tridiagonal rather than Hessenberg form is about 80 percent ($10n^3$ to $2n^3$); in the 1960s a factor of 5 was significant. However, no stable version, one that is guaranteed to avoid huge intermediate quantities, has been found.

Circumstances have changed since the 1960s. Matrices are often larger and sparser. New computer hardware has reduced the dominance of arithmetic operations in assessing speed and cost. Moreover, there are applications for the tridiagonal form that are independent of the eigenvalue problem (see [27]).

Interest in finding better algorithms has revived in the 1980s and is nicely reviewed in [7]. Attention has been concentrated on how to recover from breakdown and this focus is of practical importance. Though breakdown (the vanishing of a denominator) is rare, near breakdowns are not and they provoke instability. Remedies for breakdown lead to remedies for instability.

Next we turn to the degrees of freedom present in the reduction to tridiagonal form.

2.1. Diagonal scaling. If $Q^{-1}BQ = T$, a tridiagonal matrix, and if D is diagonal and invertible, then $(QD)^{-1}B(QD) = D^{-1}TD$ is another tridiagonal matrix similar to B. T and $D^{-1}TD$ are equivalent for theoretical purposes. Note that $T(i,i)$ and $T(i+1,i)T(i,i+1)$, $\quad i = 1, \cdots, n$ are invariant under diagonal scaling.

2.2. Reduced matrices. A tridiagonal matrix is *reduced* if one (or more) of its next-to-diagonal entries vanishes. Observe that an unreduced $n \times n$ tridiagonal T has the property that rank $[T - \xi I]$ can never drop below $n - 1$. It follows that the eigenspace of each eigenvalue is one-dimensional. Such matrices are sometimes called nonderogatory and this property is invariant under similarity transformations. Consequently, a derogatory matrix (such as the identity I) can never be transformed to unreduced tridiagonal form.

Derogatory matrices are not "difficult" in any practical sense but the parametric representation of the tridiagonal forms does break down; a uniqueness property is lost. The implication of this is that a large class of methods for reducing an $n \times n$ matrix are liable to terminate early, having found an invariant subspace, range (Q_1), where

$$BQ_1 = Q_1T_1, \quad Q_1 \in \mathbb{C}^{n\times m}, \quad m < n.$$

There are infinitely many ways of adding more columns to obtain an invertible $Q = (Q_1, Q_2)$ and

$$Q^{-1}BQ = \begin{pmatrix} T_1 & * \\ 0 & T_2 \end{pmatrix},$$

where the $*$ may or may not vanish.

For the eigenvalue problem the occurrence of a split in T is an advantage. For a theoretical discussion it is not unreasonable to stop with $BQ_1 = Q_1T_1$ and regard this as benign early termination. The possible continuation of the reduction is just a new problem on a smaller space. This point of view will be taken in what follows.

Note that with a nonderogatory B the tridiagonal representation *may* split, but for a derogatory matrix it *must* split.

2.3. Parametric representation. The following useful result is well known and is not usually attributed to any one person. It is closely related to the implicit Q theorem in [10].

THEOREM 2.1. *If $B \in \mathbb{C}^{n\times n}$ is similar to an unreduced tridiagonal $T \in \mathbb{C}^{n\times n}$, i.e., if*

$$Q^{-1}BQ = T,$$

then Q and T are determined, to within diagonal scaling, by the first (or last) column of Q and the first (or last) row of Q^{-1}.

An equivalent form of this result is less well known despite the fact that it is more elegant theoretically and offers practical advantages as well; see §10.

THEOREM 2.2. *If $B \in \mathbb{C}^{n\times n}$ admits a representation of the form*

$$P^*IQ = \Omega = \operatorname{diag}(\omega_1, \cdots, \omega_n), \quad \omega_i \neq 0, \quad i = 1, \cdots, n,$$

$$P^*BQ = \hat{T} = \text{ unreduced tridiagonal,}$$

with invertible matrices P and Q, then P, Q, Ω, and $\hat{T}$ are determined, to within column scaling, by the first (or last) columns of P and Q.

Remark. By exercising the freedom to scale columns of P and Q, we can arrange that

$$\hat{T}(i, i+1) = \hat{T}(i+1, i) = \omega_{i+1}, \qquad i = 1, \cdots, n-1.$$

Thus $\hat{T}$ is symmetric but not necessarily Hermitian.

Remark. The connection between Theorems 2.1 and 2.2 is that

$$T = \Omega^{-1}\hat{T}.$$

Theorem 2.1 shows that there is a mapping from the projective space, called $\mathbb{CP}^{n-1} \otimes \mathbb{CP}^{n-1}$, of pairs of lines in $\mathbb{C}^n$ into the unreduced tridiagonals[1] in the similarity class of B. However, this mapping is partial; not all pairs yield an image. The emphasis in §§6–9 is on the characterization of the exceptional pairs.

- If B is nonderogatory then the mapping is densely defined. Later sections show that certain determinants depending on the pair must vanish if the pair is exceptional.
- If the pair (q, p) maps to T with $q_1 = q$, $p_1 = p$ then the same pair would map to $\tilde{I}T\tilde{I}$ with $q_n = q, p_n = p$, where $\tilde{I}$ is the reversing matrix, $\tilde{I} = (e_n, e_{n-1}, \cdots, e_2, e_1)$. The use of last columns instead of first columns leads to a map that is isomorphic to the first and will not be mentioned again.
- In the light of this parameterization we may distinguish a special class of methods. The *fixed start* methods never change q_1 and p_1. This class is to be distinguished from those methods that attempt to generalize the Givens method or to reduce bandwidth and so change column 1 of the current Q and P at later steps. Unfortunately, known generalizations of Givens reduction and of bandwidth reduction are also plagued by breakdown and instability. There is more than one member of the fixed start class because there are various possibilities for the way in which the unique Q and P are built up.
- A proof of Theorem 2.2 is given because it is constructive and introduces the Lanczos algorithm [21], the earliest of the *fixed start* methods.

Proof of Theorem 2.2. Let

$$\hat{T} = \operatorname{tridiag}\begin{pmatrix} & \omega_2 & & \omega_3 & & * & & * & & \omega_n & \\ \alpha_1 & & \alpha_2 & & \alpha_3 & & * & & * & & \alpha_n \\ & \omega_2 & & \omega_3 & & * & & * & & \omega_n & \end{pmatrix} = \hat{T}^t.$$

[1] These tridiagonals are normalized by $T(i+1, i) = 1$.

Since Ω is invertible, by hypothesis, the governing equations may be rewritten

$$BQ = Q\Omega^{-1}\hat{T}, \tag{2.1}$$

$$P^*B = \hat{T}\Omega^{-1}P^*. \tag{2.2}$$

Knowledge of p_1 and q_1 yields directly

$$\omega_1 = p_1^* q_1 \neq 0, \qquad \alpha_1 = p_1^* B q_1.$$

To start the construction of Q and P equate column 1 on each side of (2.1) and row 1 on each side of (2.2):

$$\begin{aligned} q_2 &= q_2\left(\frac{\omega_2}{\omega_2}\right) = Bq_1 - q_1\left(\frac{\alpha_1}{\omega_1}\right), \\ p_2^* &= \left(\frac{\omega_2}{\omega_2}\right)p_2^* = p_1^*B - \left(\frac{\alpha_1}{\omega_1}\right)p_1^*, \\ \omega_2 &= p_2^* q_2, \\ \alpha_2 &= p_2^* B q_2, \end{aligned}$$

are each determined by q_1 and p_1.

Next assume that $(q_1, \cdots, q_{j-1})$, $(p_1, \cdots, p_{j-1})$, α_{j-1}, ω_{j-1}, ω_{j-2} are all known. Equate columns $j-1$ on each side of (2.1), rows $j-1$ on each side of (2.2), and rearrange terms to find that

$$\begin{aligned} q_j &= q_j\left(\frac{\omega_j}{\omega_j}\right) = Bq_{j-1} - q_{j-1}\left(\frac{\alpha_{j-1}}{\omega_{j-1}}\right) - q_{j-2}\left(\frac{\omega_{j-1}}{\omega_{j-2}}\right) =: r_j, \\ p_j^* &= \left(\frac{\omega_j}{\omega_j}\right)p_j^* = p_{j-1}^*B - \left(\frac{\alpha_{j-1}}{\omega_{j-1}}\right)p_{j-1}^* - \left(\frac{\omega_{j-1}}{\omega_{j-2}}\right)p_{j-2}^* =: s_j^*, \\ \omega_j &= p_j^* q_j \ \ (\neq 0), \\ \alpha_j &= p_j^* B q_j, \end{aligned}$$

are each determined by previous quantities. Hence, by induction on j, all columns of Q and P are determined by q_1 and p_1.

A similar argument shows that q_n and p_n determine Q, P, Ω and $\hat{T}$. □

The proof shows that (q_1, p_1^*) is an exceptional pair if and only if, for some $l < n$,

$$\omega_{l+1} = p_{l+1}^* q_{l+1} = 0.$$

If $q_{l+1} = 0$ and $p_{l+1}^* \neq 0$ then the algorithm may be continued by redefining q_{l+1} to be *any* vector annihilated by $p_1^*, \cdots, p_l^*$, but not by p_{l+1}^*, and setting $\hat{T}(l+1, l) = 0$. Note that $\hat{T}$ will no longer be symmetric and the dependence of q_i, p_i^*, $i > l$ on q_1 and p_1 has been lost. However, the column space of $[q_1, \cdots, q_l]$ is B-invariant.

If $q_{l+1} \neq 0$, $p_{l+1}^* \neq 0$ but $\omega_{l+1} = p_{l+1}^* q_{l+1} = 0$ then the breakdown is called serious; see [29]. Clearly, there is no pair Q, P with the given q_1 and p_1 that satisfies both equations in Theorem 2.2. Local changes to p_{l+1}^* and q_{l+1} will not suffice.

The important result is that the top unreduced submatrix of every T similar to B is completely determined, to within diagonal scaling, by the pair of directions (q_1, p_1).

3. The two-sided GS algorithm. This section presents a familiar process in an unfamiliar format in order to emphasize the fact that the GS process does not require an inner product. The extended GS procedure (introduced below), though natural, appears to be new.

Given a sequence $\langle u_1, u_2, u_3, \cdots \rangle$ in $\mathbb{C}^n$ and a sequence $\langle v_1^*, v_2^*, v_3^*, \cdots \rangle$ in the dual space $\mathbb{C}_*^n$ of linear functionals on $\mathbb{C}^n$, the GS procedure produces a new pair of sequences that form *dual bases* in the associated subspaces. However, in contrast to the familiar form of the algorithm (when $u_i = v_i$), two-sided GS can break down.

Two-sided GS Algorithm: (GS)
Input: $u_i \in \mathbb{C}^n, u_i \neq \mathbf{0}, i = 1, \cdots, m; v_i^* \in \mathbb{C}_*^n, v_i^* \neq \mathbf{0}^*, i = 1, \cdots, m$
Output:
- $l \in \mathbb{N}$ and $q_i \in \mathbb{C}^n, p_i^* \in \mathbb{C}_*^n, i = 1, \cdots, l, (l \leq m)$ satisfying $p_i^* q_k = 0, i \neq k$, and $p_i^* q_i = \omega_i \neq 0$, $i = 1, \cdots, l$.
- (if $l < m$) p_{l+1}^*, q_{l+1} with $p_{l+1}^* q_{l+1} = 0$. Both, one, or neither of p_{l+1}^* and q_{l+1} may vanish.

Initialization: $l := -1, invar = false$;
repeat
$l := l + 1$;
$q_{l+1} := u_{l+1} - \sum_{i=1}^{l} q_i (p_i^* u_{l+1})/\omega_i$,
$p_{l+1}^* := v_{l+1}^* - \sum_{i=1}^{l} (v_{l+1}^* q_i) p_i^*/\omega_i$,
if $q_{l+1} = \mathbf{0}$ **or** $p_{l+1}^* = \mathbf{0}^*$
then $invar := true; \omega_{l+1} = 0$;
else $\omega_{l+1} := p_{l+1}^* q_{l+1}$
until $l + 1 = m$ **or** $\omega_{l+1} = 0$

Remark 3.1. In practice, when norms are defined on $\mathbb{C}^n$ and $\mathbb{C}_*^n$, it is advisable to normalize the output vectors but this feature is not theoretically necessary.

Remark 3.2. If either q_{l+1} or p_{l+1}^* vanishes (or both), we say that the breakdown is benign, not serious. GS may be continued by choosing as q_{l+1} any vector $\in \text{span}\{u_i\}_{i=1}^m$ annihilated by $p_1^*, \cdots, p_l^*$ but not p_{l+1}^*, but when neither vanishes we have serious breakdown at the end of step l.

The results of the GS will be expressed compactly in terms of the following $n \times l$ matrices:

$$U_l := [u_1, \cdots, u_l], \qquad V_l := [v_1, \cdots, v_l],$$

$$Q_l := [q_1, \cdots, q_l], \qquad P_l := [p_1, \cdots, p_l],$$

together with unit triangular matrices

$$R_l = [r_{ij}] \in \mathbb{C}^{l \times l}, \qquad r_{ij} = \begin{cases} 0, & i > j, \\ 1, & i = j, \\ p_i^* u_j/\omega_j, & i < j, \end{cases}$$

$$L_l = [l_{ij}] \in \mathbb{C}^{l \times l}, \qquad l_{ij} = \begin{cases} v_i^* q_j/\omega_j, & i > j, \\ 1, & i = j, \\ 0, & i < j. \end{cases}$$

These definitions give the GS factorizations

$$U_l = Q_l R_l, \qquad V_l^* = L_l P_l^* \tag{3.1}$$

subject to

$$P_l^* Q_l = \operatorname{diag}(\omega_1, \cdots, \omega_l) = \Omega_l.$$

In practice the modified GS process (MGS) is preferred to GS but MGS is not appropriate in our application in §5.

It is useful to represent the output quantities Q_l, P_l^* in a way that is independent of the algorithm. There is a unique (oblique) projection Π_k onto $\operatorname{span}(u_1, u_2, \cdots, u_k)$ "along" $\operatorname{span}(v_1^*, v_2^*, \cdots, v_k^*)$, for each $k \leq l$. It is readily verified that a convenient matrix representation is

$$\Pi_k = Q_k (P_k^* Q_k)^{-1} P_k^*.$$

Moreover,

$$q_{k+1} = (I - \Pi_k) u_{k+1}, \qquad p_{k+1}^* = v_{k+1}^* (I - \Pi_k).$$

3.1. Extended GS. It is possible to continue GS despite some serious breakdowns by working with several vectors at each step. This has useful applications. The array dim will hold the dimensions of the diagonal blocks of Ω while b counts the blocks and l counts the columns in P and Q.

Initialization: $b := 0$; $l := 0$; $\dim(0) := 1$; $invariant := false$;
repeat
  $l := l + \dim(b)$; $b := b + 1$;
  $q_b := u_l - \sum_{\nu=1}^{b-1} Q_\nu \Omega_\nu^{-1} P_\nu^* u_l$;
  $p_b^* := v_l^* - \sum_{\nu=1}^{b-1} v_l^* Q_\nu \Omega_\nu^{-1} P_\nu^*$;
  if $q_b = \mathbf{0}$ **or** $p_b^* = \mathbf{0}^*$ **then**
    $invariant := true$; $\dim(b) := 0$;
  else
    $\delta := 1$; $Q_b := (q_b)$; $P_b := (p_b)$;
    while $\Omega_b := P_b^* Q_b$ is singular **and** $l + \delta \leq m$ **do**
      $\hat{u} := u_{l+\delta} - \sum_{\nu=1}^{b-1} Q_\nu \Omega_\nu^{-1} P_\nu^* u_{l+\delta}$;
      $\hat{v}^* := v_{l+\delta}^* - \sum_{\nu=1}^{b-1} v_{l+\delta}^* Q_\nu \Omega_\nu^{-1} P_\nu^*$;
      $Q_b := [Q_b, \hat{u}]$;
      $P_b := [P_b, \hat{v}]$;
      $\delta := \delta + 1$;
    $\dim(b) := \delta$;
until $l + \dim(b) > m$ **or** $invariant$

This algorithm yields decompositions like the one in (3.1) with R_ℓ and L_ℓ triangular, but now the scalar ω_i is replaced by a special Hankel matrix Ω_i whenever there is a breakdown. There is more on this in §6.

4. Linear systems and their Hankel forms. Here we describe those parts of systems theory that are germane to the reduction to tridiagonal form. Associated with any triple (B, q, p^*) with $B \in \mathbb{C}^{n \times n}, q \in \mathbb{C}^n, p^* \in \mathbb{C}_*^n$ is a "dynamical system" that evolves according to the linear laws

$$\dot{x}(\tau) = Bx(\tau) + qv(\tau), \tag{4.1}$$

$$\eta(\tau) = p^* x(\tau), \tag{4.2}$$

where $x(\tau) \in \mathbb{C}^n$ represents the state of the system at time τ, $\dot{x}$ is its time derivative, and $v(\tau)$ (read as *upsilon*, the greek u) represents a scalar control (or input) variable. Without loss of generality, it is assumed that $x(0) = \mathbf{0}$. Normally systems theory is set in $\mathbb{R}^n$, not $\mathbb{C}^n$, but the extra generality comes without penalty. Our system is called a single input, single output, time-invariant system. An excellent introduction to the subject is [18].

It is assumed that the state x is only knowable indirectly through the output function η and the input function v. The connection between the v and the η is most simply expressed via the Laplace transform

$$\tilde{f}(\sigma) := \int_0^\infty e^{-\sigma\tau} f(\tau) d\tau, \tag{4.3}$$

as

$$\tilde{\eta}(\sigma) = \Gamma(\sigma)\tilde{v}(\sigma) \tag{4.4}$$

where Γ is the *transfer function*

$$\Gamma(\sigma) := p^*(\sigma I - B)^{-1} q \tag{4.5}$$

$$= \sum_{k=0}^{\infty} (p^* B^k q)/\sigma^{k+1}. \tag{4.6}$$

The series representation is convergent for $|\sigma|$ large enough and the coefficients $\{p^* B^k q\}_{k=0}^{\infty}$ are called the *Markov parameters* or *impulse responses* of the system. The triple (B, q, p^*) is called a *realization* of Γ. In principle, Γ can be recovered from complete knowledge of just the input and the output functions via (4.4).

Systems theory examines various questions concerning v, x, and η but the one that is most relevant for tridiagonalization is the determination of all the realizations (B, q, p^*) that yield a given rational function Γ or, equivalently, its Markov parameters. This realization problem cuts across our reduction problem (where B is given, not Γ) but the connection is nevertheless illuminating.

The Cauchy–Binet theorem applied to $\sigma I - B$ gives

$$(\sigma I - B) \cdot \operatorname{adj}[\sigma I - B] = \det(\sigma I - B) \cdot I$$

where $\operatorname{adj}[M]$ stands for the classical adjugate matrix made up of $(n-1) \times (n-1)$ cofactors of M. Thus for $\sigma \notin B$'s spectrum,

$$\Gamma(\sigma) = p^* \operatorname{adj}[\sigma I - B] q / \chi_B(\sigma)$$

where $\chi_B(\sigma)$ is the characteristic polynomial of B. Since both numerator and denominator are polynomials in σ, cancellation of common factors is possible and, in

that event, not every eigenvalue of B will be a pole of Γ. Thus it is possible to have a triple (B', q', p'^*), with $B' \in \mathbb{C}^{m\times m}$ and $m < n$, giving rise to the same transfer function Γ as does (B, q, p^*).

DEFINITION. A *minimal realization* (B, q, p^*) of Γ is a realization in which B has minimal order (or dimension).

This minimal order is called the *McMillan degree* of Γ and of the Markov parameter sequence p^*B^iq. Clearly, it is of interest to determine the minimal realizations and one way is to associate with a sequence of Markov parameters the Hankel matrices (and their quadratic forms):

$$\mathbf{H}_l^{(k)} := [p^*B^{k+i+j-2}q] \qquad (i,j = 1,\cdots,l) \tag{4.7}$$

$$= \begin{bmatrix} \mu_k & \mu_{k+1} & \mu_{k+2} & \cdot \\ \mu_{k+1} & \mu_{k+2} & \cdot & \cdot \\ \mu_{k+2} & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix} \qquad (i,j = 1,\cdots,l), \tag{4.8}$$

$$\mathbf{H}^{(k)} = \mathbf{H}_\infty^{(k)}. \tag{4.9}$$

The Hankel property is that the (i,j) entry depends only on $i+j$. It turns out that the rank of $\mathbf{H}^{(0)}$ is the McMillan degree of Γ. For tridiagonalization, an unfortunate choice of $q_1 = q$ and $p_1 = p$ may yield rank $\mathbf{H}^{(0)} < n$ but that is not the sole cause of breakdown in the tridiagonalization process.

The controllable subspace and the observable subspace of the system and the canonical structure theorem will be introduced in §7. We mention in passing that systems theory also considers the *partial* realization problem that arises when only the first m Markov parameters are known. See [13] for an exhaustive treatment of this topic.

5. Hankel factorization; equivalent pencils. The two Hankel matrices $H_n^{(0)}$ and $H_n^{(1)}$, defined in the previous section, play a leading role in the analysis of an attempted reduction of B to tridiagonal form. That is the message of the equivalence theorems in this and later sections. To discuss the factorization of $H_n^{(0)}$ and $H_n^{(1)}$, the following four Krylov matrices are needed:

- $K_m(q,B) := [q, Bq, \cdots, B^{m-1}q] \in \mathbb{C}^{n\times m}$.
- $K(q,B) := K_\infty(q,B)$, the controllability matrix.
- $K_m(p^*,B) := K_m(p,B^*)^*$.
- $K(p^*,B) := K_\infty(p^*,B)$, the observability matrix.

Next we list some elementary but fundamental facts.

LEMMA 5.1. *For any* $j \in \mathbb{N}$,

$$H_m^{(j)} = K_m(p^*,B)B^jK_m(q,B),$$
$$H^{(j)} = K(p^*,B)B^jK(q,B).$$

Proof. The (i,k) entry on each side is $(p^*B^{i-1})B^j(B^{k-1}q)$. □

Note that $H^{(j)}$ is a submatrix of $H^{(0)}$.

COROLLARY 5.2.

$$\text{rank}\,[H^{(j)}] \le \text{rank}\,[H^{(0)}] \le \min\{\text{rank}\,[K(p^*,B)], \text{rank}\,[K(q,B)]\} \le n.$$

LEMMA 5.3. range$[K(q,B)]$ *is the smallest* B*-invariant subspace of* $\mathbb{C}^n$ *that contains* q. range$[K(p^*,B)]$ *is the smallest* B*-invariant subspace of* $\mathbb{C}_*^n$ *that contains* p^*.

Proof. If $\mathcal{S}$ is B-invariant and contains q, then it must contain Bq, $B(Bq)$, etc. Hence, range $[K(q,B)] \subset \mathcal{S}$. Moreover,

$$B \text{ range } [K(q,B)] = \text{range } [Bq, B^2q, \cdots] \subset \text{range } [K(q,B)]$$

and so range $[K(q,B)]$ is invariant. Similarly range $[K(p^*,B)]$ is invariant too. □

Together with n, the following three numbers furnish a complete classification of the various cases in the mapping $(q,p^*) \longrightarrow (\hat{T},\Omega)$.

DEFINITION.

$$\begin{aligned} l &:= \min\{j : H_{j+1}^{(0)} \text{is singular}\}, \\ r &:= \text{rank } [H^{(0)}], \\ m &:= \min\{\text{rank}[K(p^*,B)], \text{rank}[K(q,B)]\}. \end{aligned}$$

The corollary given above and these definitions yield

$$l \le r \le m \le n.$$

It is strict inequalities that lead to early termination in tridiagonalization. Anticipating later sections, we can summarize the situation.

- $l = m < n$ yields benign early termination with an invariant subspace. Tridiagonalization may be continued in infinitely many ways.
- $l < r = m$ yields serious breakdown that can be cured by permitting block tridiagonalization.
- $r < m$ yields incurable breakdown but a minimal realization of the transfer function (see §7).

THEOREM 5.4 (Hankel factorization). *Let* $H^{(0)} = H^{(0)}(B,q,p^*)$. *The two-sided* GS *process applied to the columns of* $K(q,B)$ *and the rows of* $K(p^*,B)$ *yields a whole number* l *and rank* l *matrices* Q_l, P_l^*, *and a unit lower triangular* $l \times l$ *matrix* L_l *such that*

$$K_l(q,B) = Q_l L_l^t, \quad K_l(p^*,B) = L_l P_l^*,$$

and

$$\begin{aligned} H_l^{(0)} &= L_l \Omega_l L_l^t, \\ \Omega_l &:= P_l^* Q_l = \text{diag}(\omega_1, \cdots, \omega_l), \qquad \omega_i \neq 0, i = 1, \cdots, l. \end{aligned}$$

In addition, GS *produces* q_{l+1}, p_{l+1} *satisfying* $\omega_{l+1} = p_{l+1}^* q_{l+1} = 0$.

Proof. Algorithm GS (see §3) applied to $K(q,B)$ and $K(p^*,B)$ yields Q_l, P_l so that

$$K_l(q,B) = Q_l R_l, \quad K_l(p^*,B) = L_l P_l^*, \quad P_l^* Q_l = \Omega_l,$$

where R_l is unit upper triangular. By Lemma 5.1

$$\begin{aligned} H_l^{(0)} &= K_l(p^*,B) K_l(q,B) \\ &= L_l P_l^* Q_l R_l \\ &= L_l \Omega_l R_l. \end{aligned}$$

By the LDU theorem L_l, Ω_l, R_l are unique. By symmetry of $H_l^{(0)}$, $R_l = L_l^t$. By definition of l in GS, $\omega_{l+1} = 0, \omega_i \neq 0, i \leq l$. □

LEMMA 5.5 (tridiagonal form). *With the notation of the Hankel factorization theorem, $\hat{T}_l := P_l^* B Q_l$ is symmetric, tridiagonal, and unreduced.*

Proof. The characteristic property of Krylov matrices and the output of the GS process is that, for $j < l$,

$$q_j \in \text{range}\,[K_j(q, B)], \qquad q_j \notin \text{range}\,[K_{j-1}(q, B)].$$

Hence

$$Bq_j \in B \cdot \text{range}\,[K_j(q, B)] \subset \text{range}\,[K_{j+1}(q, B)],$$
$$Bq_j \notin B \cdot \text{range}[K_{j-1}(q, B)] \oplus \text{span}(q) = \text{range}\,[K_j(q, B)].$$

Furthermore, p_i^* annihilates $K_{j+1}(q, B)$ for all $i > j + 1$ and so p_i^* annihilates column j of BQ_l for all $i > j + 1$, and so the (i, j) entry of $P_l^* BQ_l$ vanishes for $i - j > 1$. Similarly, for $i < l$, $p_i^* B \in \text{range}\,[K_{i+1}(p^*, B)]$ whose null space contains q_j for all $j > i + 1$. Hence the (i, j) entry of $P_l^* BQ_l$ vanishes for $j - i > 1$.

It remains to show that $\hat{T}_l$ is unreduced. By the minimality property of l the GS algorithm does not break down at step j for $j < l$. Thus Bq_j has a nonzero component γ in q_{j+1} when expanded in terms of $(q_1, q_2, \cdots, q_{j+1})$. Since p_{j+1}^* annihilates $q_1, q_2, \cdots, q_j$ but not q_{j+1}, it follows that

$$p_{j+1}^* Bq_j = p_{j+1}^* q_{j+1} \gamma = \omega_{j+1} \gamma \neq 0, \qquad j < l.$$

By similar arguments,

$$p_j^* Bq_{j+1} = \delta p_{j+1}^* q_{j+1} = \delta \omega_{j+1} \neq 0, \qquad j < l. \qquad \square$$

First we consider the generic case when $K(q, B)$ and $K(p^*, B)$ have full rank n and the associated system (B, q, p^*) is said to be controllable and observable.

THEOREM 5.6 (equivalence theorem, version 1). *If invertible $H_n^{(0)}(B, q, p^*)$ permits triangular factorization*

$$H_n^{(0)} = L_n \Omega_n L_n^t,$$

then the following pencils are equivalent:

$$(H_n^{(1)}, H_n^{(0)}), \quad (B, I), \quad (\hat{T}_n, \Omega_n).$$

Here $\hat{T}_n$ is the symmetric, unreduced, tridiagonal matrix

$$\hat{T}_n = L_n^{-1} H_n^{(1)} L_n^{-t} = P_n^* B Q_n.$$

The first and third pencils are symmetric but not necessarily Hermitian.

In systems theory these generic $H_n^{(0)}$ are called strongly regular.

Proof. By Lemma 5.1, $H_n^{(0)} = K_n(p^*, B) K_n(q, B)$ and if either Krylov matrix were rank deficient, then so would be their product. Since $H_n^{(0)}$ permits triangular factorization,

$$l = r = m = n.$$

Thus all the transformation matrices that appear below are invertible:

$$\begin{aligned} H_n^{(1)} - \lambda H_n^{(0)} &= K_n(p^*, B)(B - \lambda I)K_n(q, B) \quad \text{(Lemma 5.1)} \\ &= L_n P_n^*(B - \lambda I)Q_n L_n^t \quad \text{(GS factorization)} \\ &= L_n(\hat{T}_n - \lambda \Omega_n)L_n^t. \end{aligned}$$

Theorem 5.4 yields $P_n^* Q_n = \Omega_n$ =diagonal, Lemma 5.5 yields $P_n^* B Q_n = \hat{T}_n$ = tridiagonal, unreduced. Thus $H_n^{(1)} = L_n \hat{T}_n L_n^t$ and $\hat{T}$ must be symmetric. □

Since the pencils (B, I) and $(\hat{T}_n, \Omega_n)$ are equivalent, $\Omega_n^{-1}\hat{T}_n$ has the same spectrum as B.

To complete the circle of ideas it is necessary to identify the matrix Q_n coming from the GS factorization $K_n(q, B) = Q_n L_n^t$ with the matrix Q generated in the proof of Theorem 2.2 solely from the property of producing a tridiagonal form $BQ = Q(\Omega^{-1}\hat{T})$. This will show that the result in Theorem 5.6 is categorical; that is, Hankel factorization, explicit or implicit, is the only mechanism for producing the desired $(\hat{T}, \Omega)$ representation. In other words,

$$\text{tridiagonal reduction} \equiv \text{Krylov matrices} + \text{GS}.$$

The identification is an immediate consequence of the fact that $Qe_{j+1} = \phi_j(B)q_1$ where ϕ_j is a monic polynomial of degree j. To see this, note that $BQ = Q(\Omega^{-1}\hat{T})$ implies that

$$Bq_j = q_{j+1} + q_j(\hat{T}(j, j)/\omega_j) + q_{j-1}(\omega_j/\omega_{j-1}), \qquad j = 2, \cdots, n-1.$$

Thus, if $q_j = \phi_{j-1}(B)q_1$, $q_{j-1} = \phi_{j-2}(B)q_1$, then $q_{j+1} = \phi_j(B)q_1$, for $j = 2, \cdots, n-1$. But $q_1 = \phi_0(B)q_1$, and $q_2 = Bq_1 - q_1(\hat{T}(1,1)/\omega_1) = \phi_1(B)q_1$, and the principle of induction establishes the polynomial representation. In matrix terms $Q = K_n(q_1, B)R^{-1}$ for some unit upper triangular R. Similarly, $P^* := \Omega Q^{-1}$ satisfies $P^* = L^{-1}K_n(p_1^*, B)$ for some unit lower triangular L. Finally

$$\Omega = P^*Q = L^{-1}K_n(p_1^*, B)K_n(q_1, B)R^{-1} = L^{-1}H_n^{(0)}R^{-1}$$

and the uniqueness of triangular factorization shows that R must be the matrix L_n^t from Theorem 5.6.

Before proceeding to cases of failure to produce a tridiagonal form $(\hat{T}_n, \Omega_n)$, we repeat that the case

$$l = r = m < n$$

is essentially like the one above. Either q_{l+1} or p_{l+1}^* generated by GS vanishes and it is only necessary to replace the zero vectors by suitably chosen nonzero vectors to ensure continuation of GS until step n. However, $\hat{T}_n$ will be reduced. It may be preferable to stop with $(\hat{T}_l, \Omega_l)$. Although this pencil is not equivalent to (B, I), nevertheless, when $q_{l+1} = 0$, it follows that

$$\text{range } K_l(q, B) = \text{range } K(q, B),$$

an invariant subspace. It follows that $(\hat{T}_l, \Omega_l)$ is equivalent to $(\hat{B}, I_l)$ where $\hat{B}$ is the restriction of B to the invariant subspace. Not only is every eigenvalue λ of $\Omega_l^{-1}\hat{T}_l$ an eigenvalue of B, but right eigenvectors are explicitly given by $Q_l v$ where $\hat{T}_l v = \Omega_l v\lambda$.

We will not consider this case in any more detail.

6. Block tridiagonal form. This section examines those exceptional cases when $H_n^{(0)}$ has full rank n but does not permit triangular factorization. In the terminology of the previous section: $l < r = m = n$. This is serious breakdown and there is no tridiagonal form $(\hat{T}, \Omega)$ with the given parameter values q and p^*. In the language of system theory, $H_n^{(0)}$ is said to be regular but not strongly regular.

A natural way to persist with the starting vectors and obtain a sparse representation of B is to accept a block tridiagonal representation. The smaller the blocks, the better, and it is of interest to describe the most refined representation that is possible. This was done in [12] and the result may be found more easily in the definitive paper [13], where the authors give ample references to related earlier work. Here is our description of Gragg's result.

DEFINITION. Let $\nu(1), \nu(2), \cdots, \nu(k)$ be the sequence of index values j such that $H_j^{(0)}$ is invertible. These are the *degree indices* of $H^{(0)}$. Let $\nu(0) = 0$.

Under our assumption on $H^{(0)}$, $\nu(k) = n$ for some $k \leq n$.

THEOREM 6.1 (block triangular factorization). *The most refined block triangular factorization of* $H^{(0)}$ *is*

$$H^{(0)} = L\Omega L^t$$

where $\Omega = \Omega_1 \oplus \Omega_2 \oplus \cdots \oplus \Omega_k \oplus O_\infty$, *and* Ω_j *is a nonsingular right lower triangular Hankel matrix of order* $\nu(j) - \nu(j-1)$. *Here* $\nu(k) = \operatorname{rank} H^{(0)} = n$. *Moreover, columns* $\nu(j)+1$ *to* $\nu(j+1)$ *of* L^{-1} *have unit lower triangular Toeplitz structure, for* $j = 0, \cdots, k-1$. *However, the entries of* Ω_i *below the secondary diagonal are not uniquely determined by* $H^{(0)}$.

An Ω_j of order 3 has the form

$$\begin{bmatrix} 0 & 0 & \pi_3 \\ 0 & \pi_3 & \pi_4 \\ \pi_3 & \pi_4 & \pi_5 \end{bmatrix}, \quad \pi_3 \neq 0, \quad (\pi_4, \pi_5 \text{ not unique}).$$

In [14] the irregular orthogonal polynomials are chosen so that each Ω_i is a multiple of $\tilde{I}$, the reversal matrix, i.e., $\pi_4 = \pi_5 = 0$ in the matrix above. In general, the Schur complement of $H_j^{(0)}$ in $H^{(0)}$ is not a Hankel matrix so the only surprising feature of Theorem 6.1 is the Hankel structure of the Ω_j and the structure of L^{-1}. The extended GS algorithm of §3 is intimately connected with the factorization.

We will not present a proof of Theorem 6.1. Instead, we offer some further discussion. The reduced matrix that remains to be processed after j steps of triangular factorization is called the *Schur complement* of the leading principal $j \times j$ submatrix. Surprisingly, neither [13] nor [16] discuss Schur complements in Hankel matrices. One reason may be that the Schur complement of $H_j^{(0)}$ in $H^{(0)}$ is not a Hankel matrix although it does possess an interesting structure. Kailath and his coworkers have studied these Schur complements in the course of their work on *displacement rank*.

THEOREM 6.2 (Schur complements). *If* $H_j^{(0)}$ *is invertible then its Schur complement* $H_{\langle j \rangle}^{(0)}$ *in* $H^{(0)}$ *exists and is triangularly congruent to a Hankel matrix*

$$H_{\langle j \rangle}^{(0)} = LHL^t$$

where L *is unit lower triangular and Toeplitz.*

The reader is referred to [22] for proofs that use the bilinear forms associated with matrices. Matrices with the structure shown in Theorem 6.2 are called quasi-Hankel by Kailath.

Applying these results to Gragg's result (Theorem 6.1) shows that a leading principal submatrix of these Schur complements is actually of Hankel form. In fact, we can give a realization of this Hankel part in terms of the so-called Lanczos vectors, which are the columns of Q and P in the GS factorizations

$$K(q,B) = QL^t, \qquad K(p^*,B) = LP^*.$$

Referring to the diagonal blocks Ω_i in Theorem 6.1, we have the following useful result:

- Let $d_{j+1} = \nu(j+1) - \nu(j)$. The leading principal submatrix of order d_{j+1} of $H^{(0)}_{\langle\nu(j)\rangle}$ is

$$\Omega_{j+1} := H^{(0)}_{d_{j+1}}(q_{j+1}, p^*_{j+1}, B).$$

This is the first invertible submatrix of the Schur complement. Note that $q_{j+1} = (I - \Pi_{\nu(j)})B^{\nu(j)}q_1$. Here Π is the GS projector defined in §3.

The hypotheses of Theorem 5.6 may now be weakened.

THEOREM 6.3 (equivalence theorem, version 2). *Let* $H^{(0)} = H^{(0)}(B,q,p^*)$. *If* rank $[H^{(0)}]$ = *order of* $B = n$, *then the following pencils are equivalent:*

$$(H^{(1)}_n, H^{(0)}_n), \quad (B, I_n), \quad (\hat{T}, \Omega),$$

where Ω *is block diagonal as given above and*

$$\hat{T} = L^{-1}H^{(1)}_n L^{-t} = P^*BQ$$

is symmetric block tridiagonal with structure conformable to Ω. *The block sizes in* Ω *are minimal. The off-diagonal blocks of* $\hat{T}$ *are null except in the lower right entries.*

The proof of this theorem is analogous to the proof of Lemma 5.5 and will be omitted. The only difference is that in extending Lemma 5.5 the phrase "$Bq_{j-1} \in$ range $[K_j]$" must be replaced by "range $(Bq_{j-1}) \subset$ range $[K_j]$" since q_{j-1} is now a matrix. The structure of Ω and $\hat{T}$ is shown at the end of this section.

Although the theorems extend readily to block form, the elementary sequential process of §2 (the Lanczos algorithm) for computing $\hat{T}$ and Ω will not suffice alone when the block sizes vary. For example, let the jth (block) row of $\hat{T}$ be

$$(0 \cdots 0\ B_j\ A_j\ B^t_{j+1}\ 0 \cdots 0), \quad A_j \in \mathbb{C}^{d_j \times d_j}, \quad B_j \in \mathbb{C}^{d_j \times d_{j-1}},$$

where

$$d_j := \nu(j) - \nu(j-1)$$

and let

$$\hat{Q} = (Q_1, Q_2, \cdots, Q_k), \qquad Q_j \in \mathbb{C}^{n \times d_j}.$$

Then equate the jth (block) column on each side of

$$B\hat{Q} = \hat{Q}\Omega^{-1}\hat{T}, \qquad \hat{P}^*B = \hat{T}\Omega^{-1}\hat{P}^*$$

to find

$$Q_{j+1}\Omega_{j+1}^{-1}B_{j+1} = R_{j+1} := BQ_j - Q_j\Omega_j^{-1}A_j - Q_{j-1}\Omega_{j-1}^{-1}B_j^t \in \mathbb{C}^{n\times d_j}$$

and

$$B_{j+1}^t\Omega_{j+1}^{-1}P_{j+1}^* = S_{j+1}^* := P_j^*B - A_j\Omega_j^{-1}P_j^* - B_j\Omega_{j-1}^{-1}P_{j-1}^* \in \mathbb{C}^{d_j\times n}.$$

At the jth step of the process the right sides are known but $Q_{j+1}, \Omega_{j+1}, B_{j+1}$, and P_{j+1}^* are not. Moreover, if $d_{j+1} > d_j$ these right sides are not sufficient by themselves to determine any of the unknowns. This is a manifestation of the lack of uniqueness in Theorem 6.1. More work must be done with the matrix B to continue the computation. Nevertheless, the extended GS process (§3) can determine the blocks $Q_1, Q_2, \cdots$, and $P_1^*, P_2^*, \cdots$ and these, in turn, determine $\hat{T}$ and Ω. At each step the block sizes are minimal subject to the constraint that $P_j^*Q_j$ be invertible.

These ideas lead to the look-ahead Lanczos algorithm that was presented in [25]. The practical defect of this method is that the sizes of the blocks are not known in advance and so dynamic allocation of storage is desirable for the blocks A_j and B_j of $\hat{T}$. However, FORTRAN does not permit such an arrangement. We say more about these ideas in §11.

This section has shown that breakdown corresponding to $l < r = n$ can be cured by accepting block tridiagonal representations.

6.1. The structure of $\hat{T}$ and Ω.

$$Q := [q_1, Bq_1, B^2q_1, q_2, Bq_2, q_3, Bq_3, B^2q_3, B^3q_3],$$
$$P := [p_1, B^*p_1, B^{*2}p_1, p_2, B^*p_2, p_3, B^*p_3, B^{*2}p_3, B^{*3}p_3],$$
$$\mu_k^j := p_j^*B^kq_j,$$
$$\Omega = \Omega_1 \oplus \Omega_2 \oplus \Omega_3$$
$$= \begin{bmatrix} 0 & 0 & \mu_2^1 \\ 0 & \mu_2^1 & \mu_3^1 \\ \mu_2^1 & \mu_3^1 & \mu_4^1 \end{bmatrix} \oplus \begin{bmatrix} 0 & \mu_1^2 \\ \mu_1^2 & \mu_2^2 \end{bmatrix} \oplus \begin{bmatrix} 0 & 0 & 0 & \mu_3^3 \\ 0 & 0 & \mu_3^3 & \mu_4^3 \\ 0 & \mu_3^3 & \mu_4^3 & \mu_5^3 \\ \mu_3^3 & \mu_4^3 & \mu_5^3 & \mu_6^3 \end{bmatrix},$$

$$\hat{T} = \begin{bmatrix} 0 & \mu_2^1 & \mu_3^1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \mu_2^1 & \mu_3^1 & \mu_4^1 & 0 & 0 & 0 & 0 & 0 & 0 \\ \mu_3^1 & \mu_4^1 & \mu_5^1 & 0 & \mu_1^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \mu_1^2 & \mu_2^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & \mu_1^2 & \mu_2^2 & \mu_3^2 & 0 & 0 & 0 & \mu_3^3 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \mu_3^3 & \mu_4^3 \\ 0 & 0 & 0 & 0 & 0 & 0 & \mu_3^3 & \mu_4^3 & \mu_5^3 \\ 0 & 0 & 0 & 0 & 0 & \mu_3^3 & \mu_4^3 & \mu_5^3 & \mu_6^3 \\ 0 & 0 & 0 & 0 & \mu_3^3 & \mu_4^3 & \mu_5^3 & \mu_6^3 & \mu_7^3 \end{bmatrix}.$$

The diagonal blocks of $\hat{T}$ are Hessenberg Hankel matrices. The off-diagonal blocks each have a single nonzero entry in the lower right position whose value is equal to the other nonzero entries in its antidiagonal. By reversing the order of the rows in each block, we see that for each $\sigma \in \mathbb{C}$, $\hat{T} - \sigma\Omega$ is a Hessenberg matrix. Thus its determinant and the derivatives with respect to σ may be evaluated rapidly and stably by Hyman's recurrence. See [24].

7. Incurable breakdown and minimal realizations. In this section we consider the general case and face up to the situation when $H^{(0)} = H^{(0)}(B, q, p^*)$ has rank $r < n (= \textit{order of } B)$. Recall that r is the McMillan degree of the transfer function $p^*(\sigma I - B)^{-1}q$. A classical result of Kronecker (see [16]) states:

$$\text{if rank } [H^{(0)}] = r \quad \text{ then } H_r^{(0)} \text{ is invertible.}$$

It follows that a (block) triangular factorization of $H_n^{(0)}$ must stop when Ω has r rows. With $r < n$ the output pencil $(\hat{T}, \Omega)$ is not equivalent to (B, I). Moreover, (B, I) need not be equivalent to $(H_n^{(1)}, H_n^{(0)})$; the breakdown is incurable.

It is of interest to express r in terms of geometric quantities that are directly related to B, q, and p^*. This expression is a byproduct of the canonical structure theorem of linear systems theory. See [9] and [19]. The rank of $H^{(0)}(B, q, p^*)$ is the dimension of any controllable, observable subspace of the system (B, q, p^*).

THEOREM 7.1 (minimal realization theorem). *Given B, q, and p^* there is a maximally refined block tridiagonal-diagonal pair $(\hat{T}, \Omega)$, not unique, such that $(\Omega^{-1}\hat{T}, e_1, e_1^*)$ is a minimal realization;*

$$p^*(\sigma I - B)^{-1}q = e_1^*(\sigma I - \Omega^{-1}\hat{T})^{-1}e_1 \quad \textit{for all } \sigma \notin B\textit{'s spectrum.}$$

Every eigenvalue of $\Omega^{-1}\hat{T}$ is an eigenvalue of B.

At first glance it is surprising, and pleasing, that incurable breakdown yields such a rich harvest of eigenvalues despite $(\hat{T}, \Omega)$ not being equivalent to the restriction of (B, I) to any B-invariant subspace of $\mathbb{C}^n$. At second glance it is annoying, but interesting, that from the matrices $Q \in \mathbb{C}^{n\times r}$ and $P \in \mathbb{C}^{n\times r}$ that yield $\Omega = P^*Q$, $\hat{T} = P^*BQ$ there is no direct way of computing eigenvectors of B that belong to eigenvalues of $\Omega^{-1}\hat{T}$. This is because there are no invariant subspaces of B in the range of Q and P. The recent paper [2] shows how to append columns to P and Q to obtain bases for $\mathcal{K}(q, B)$ and $\mathcal{K}(p, B^*)$.

Next, we give a brief summary of the canonical structure theorem. See [18] for a full account.

Recall from §4 that the linear system under consideration is

$$\dot{x} = Bx + qv, \quad x(0) = 0,$$

$$\eta = p^*x, \qquad v(0) \neq 0.$$

There are four special subspaces of $\mathbb{C}^n$ associated with the system.

- $\mathcal{K}(q, B) := \textit{range of } K(q, B) = S_c$,
  the *controllable* subspace. It is the smallest B-invariant subspace containing q.
- $\mathcal{N}(p^*, B) := \textit{null space of } K(p^*, B) = S_{\bar{o}}$,
  the *unobservable* subspace, the largest B-invariant subspace annihilated by p^*.

The analysis to follow holds for any complements in $\mathbb{C}^n$ of these two subspaces. However, there is an obvious choice of complement in each case.

- $\mathcal{N}(q^*, B^*) := \textit{null space of } K(q^*, B^*) = S_{\bar{c}}$,
  the *unobservable* subspace for the dual system (B^*, p, q^*), the largest B^*-invariant subspace annihilated by q^*.
- $\mathcal{K}(p, B^*) := \textit{range of } K(p, B^*) = S_o$,
  the *controllable* subspace for the dual system, the smallest B^*-invariant subspace containing p.

Remark 7.1. The controllable subspace should be called the reachable subspace because it consists of all the states x that can be obtained at a given time by suitable choice of the scalar function $v(t)$. On the other hand, $\mathcal{N}(p^*, B)$ is well named since it consists of the states that yield $\eta = 0$.

Remark 7.2. From the abstract point of view, the introduction of the last two subspaces is the only occasion when we invoke the antilinear mapping $v^* \Leftrightarrow v$ to transform a linear functional in $\mathbb{C}_*^n$ to a vector in $\mathbb{C}^n$. This device puts all the important spaces into $\mathbb{C}^n$.

By taking the intersection of these four subspaces in an appropriate order a canonical block structure for B is revealed. The subscripts $\bar{c}$ and $\bar{o}$ signify uncontrollable and unobservable, respectively.

$$\begin{aligned}
\mathcal{S}_{c\bar{o}} &:= \mathcal{N}(p^*, B) \cap \mathcal{K}(q, B), \\
\mathcal{S}_{co} &:= \mathcal{K}(q, B) \cap \mathcal{K}(p, B^*), \\
\mathcal{S}_{\bar{c}\bar{o}} &:= \mathcal{N}(p^*, B) \cap \mathcal{N}(q^*, B^*), \\
\mathcal{S}_{\bar{c}o} &:= \mathcal{N}(q^*, B^*) \cap \mathcal{K}(p, B^*).
\end{aligned}$$

Clearly,

$$\mathbb{C}^n = \mathcal{S}_{c\bar{o}} \oplus \mathcal{S}_{co} \oplus \mathcal{S}_{\bar{c}\bar{o}} \oplus \mathcal{S}_{\bar{c}o}.$$

Also,

$$\begin{aligned}
&\mathcal{S}_{c\bar{o}} \text{ and } \mathcal{S}_c = \mathcal{S}_{c\bar{o}} \oplus \mathcal{S}_{co} \text{ are } B\text{-invariant}, \\
&\mathcal{S}_{\bar{c}o} \text{ and } \mathcal{S}_o = \mathcal{S}_{co} \oplus \mathcal{S}_{\bar{c}o} \text{ are } B^*\text{-invariant}.
\end{aligned}$$

The controllable, observable subspace $\mathcal{S}_{co}$ is not invariant under B nor under B^*.

By taking, in order, any bases for the four subspaces listed above, a new representation $\tilde{B}$ for B is obtained that is block upper triangular.

$$\tilde{B} = \begin{bmatrix} \tilde{B}_{11} & \tilde{B}_{12} & \tilde{B}_{13} & \tilde{B}_{14} \\ 0 & \tilde{B}_{22} & 0 & \tilde{B}_{24} \\ 0 & 0 & \tilde{B}_{33} & \tilde{B}_{34} \\ 0 & 0 & 0 & \tilde{B}_{44} \end{bmatrix} = F^{-1}BF$$

and the starting vectors have the following representation:

$$\tilde{q} = \begin{pmatrix} \tilde{q}^1 \\ \tilde{q}^2 \\ 0 \\ 0 \end{pmatrix} = F^{-1}q, \qquad \tilde{p} = \begin{pmatrix} 0 \\ \tilde{p}^2 \\ 0 \\ \tilde{p}^4 \end{pmatrix} = F^*p.$$

Remark 7.3. It appears to be traditional in systems theory to invert the order of $\mathcal{S}_{\bar{c}\bar{o}}$ and $\mathcal{S}_{\bar{c}o}$. This has the advantage of putting the bad subspace $\mathcal{S}_{\bar{c}\bar{o}}$ in final position, but the disadvantage of forgoing the block triangular form dear to the hearts of matrix theorists.

The form of $\tilde{q}$ and $\tilde{p}$ reveals the final result

$$\Gamma(\sigma) = p^*(\sigma I - B)^{-1}q = (\tilde{p}^2)^*(\sigma I - \tilde{B}_{22})^{-1}\tilde{q}^2.$$

The system $(\tilde{B}_{22}, \tilde{q}^2, (\tilde{p}^2)^*)$ is a *minimal realization* of the transfer function Γ. Moreover,

$$H^{(0)}(q, p^*, B) = H^{(0)}(\tilde{q}^2, (\tilde{p}^2)^*, \tilde{B}_{22})$$

and

$$\text{rank}[H^{(0)}] = \dim\ \mathcal{S}_{co}.$$

In systems theory we start from the Markov parameters and seek a minimal realization. To obtain a tridiagonal representation, we start with B and seek q and p^* to ensure that (B, q, p^*) is minimal.

We now give the interpretation of the canonical structure theorem for the reduction of (B, I) to a more condensed form. Consider a method that builds up Q and P one column at a time. Suppose that serious breakdown occurs with

$$\omega_{j+1} := p_{j+1}^* q_{j+1} = 0, \quad q_{j+1} \neq 0, \quad p_{j+1}^* \neq 0^*.$$

Rank $[H^{(0)}]$ is not known but the following dichotomy holds. Either

$$p_{j+1}^* B^\nu q_{j+1} = 0, \qquad \nu = 0, 1, \cdots, n-j-1, \tag{7.1}$$

in which case $p_{j+1}^* B^\nu q_{j+1} = 0$ for all ν and rank $[H^{(0)}] = j$. The current matrix pencil $(\hat{T}_j, \Omega_j)$ is a minimal realization of the transfer function and every eigenvalue of $\Omega_j^{-1}\hat{T}_j$ is an eigenvalue of B, or

$$\delta = \min\{\nu\ :\ p_{j+1}^* B^{\nu-1} q_{j+1} \neq 0\} \leq n - j \tag{7.2}$$

and another step of (block) tridiagonalization may be taken with block size δ. No eigenvalue of $\Omega_j^{-1}\hat{T}_j$ is an eigenvalue of B, although some could be close.

This dichotomy is the content of the mismatch theorem in Taylor's dissertation [26]. Incurable breakdown occurs only when a minimal realization has been found. The word mismatch indicates that the eigenvalues associated with q and the eigenvalues associated with p^* are not the same sets. In other words, $\mathcal{S}_{co}$ is neither $\mathcal{K}(q, B)$ nor $\mathcal{K}(p, B^*)$, but a proper subset.

8. Summary table. Given $B \in \mathbb{C}^{n\times n}$, $q \in \mathbb{C}^n$, $p^* \in \mathbb{C}_*^n$,

$$\begin{aligned} H^{(0)} &:= [p^* B^{i+j-2} q], \qquad i, j = 1, 2, \cdots, \\ l &:= \min\{\nu : H_{\nu+1}^{(0)} \text{is singular}\}, \\ r &:= \text{rank}\,[H^{(0)}], \\ m &:= \min\{\text{rank}[K(q, B)], \text{rank}[K(p^*, B)]\}. \end{aligned}$$

There is a unique pair $(\hat{T}, \Omega)$ with

$$\begin{aligned} K_l(q, B) &= Q_l L^t, \qquad K_l(p^*, B) = L P_l^* \quad \text{(by GS)}. \\ \Omega &= \text{diag}\,(\omega_1, \cdots, \omega_l) = P_l^* Q_l, \\ H_l^{(0)} &= L\Omega L^t \quad \text{(triangular factorization)}, \\ H_l^{(1)} &= L\hat{T}L^t, \quad \hat{T}(i, i+1) = \hat{T}(i+1, i) = \omega_{i+1}, \quad i = 1, \cdots, l-1. \end{aligned}$$

The eigenvalues θ of $(\hat{T}, \Omega)$ will be called Ritz values. In Table 8.1 the phrase "invariant subspace" is an abbreviation for the assertion that either the column space of $K_l(q, B)$ or the row space of $K_l(p^*, B)$ is invariant under B. Recall from §5 that

$$l \leq r \leq m \leq n.$$

TABLE 8.1

Description	Case	Invariant subspace	Ritz values
Benign early termination	$l = r = m < n$	Yes	Each $\theta \in$ spec(B)
Incurable breakdown	$l = r < m$	No	Each $\theta \in$ spec(B)
Curable breakdown	$l < r = m$	No	No $\theta \in$ spec(B)

Generically equality holds throughout and $(\hat{T}, \Omega)$ is equivalent to (B, I). Recall that spec(B) denotes the spectrum of B.

When $l < r$ there is a (block) tridiagonal/diagonal pair $(\hat{T}, \Omega)$ of order r such that each $\theta \in$ spec (B). If and only if $r = m$ then at least one of range $K_r(q, B)$, range $K_r(p^*, B)$ is B-invariant.

9. Examples. To benefit from the examples it is preferable to work out both the computed quantities $(Q, P^*, \cdots)$ and the associated theoretical ones $(H^{(0)}, K, \cdots)$. This has been done in Example 9.1, but only the bare essentials are given in Example 9.2. All the relevant inequalities for l, r, and m are covered.

Example 9.1. Tridiagonalization with initial vectors (q, p^*) breaks down incurably at the end of step 1. Nevertheless, the 1×1 reduced pencil $(5, 1)$ delivers an eigenvalue of B.

$$B = \begin{bmatrix} 1 & 2 & 3 & 4 \\ 0 & 5 & 0 & 6 \\ 0 & 0 & 7 & 8 \\ 0 & 0 & 0 & 9 \end{bmatrix}, \quad q = \begin{bmatrix} 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}, \quad p = \begin{bmatrix} 0 \\ 1 \\ 0 \\ -1 \end{bmatrix},$$

$$K(q, B) = \begin{bmatrix} 1 & 3 & 13 & \cdot & \cdot \\ 1 & 5 & 25 & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \cdot \end{bmatrix}, \qquad Q_2 = \begin{bmatrix} 1 & -2 \\ 1 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},$$

$$K(p^*, B) = \begin{bmatrix} 0 & 1 & 0 & -1 \\ 0 & 5 & 0 & -3 \\ 0 & 25 & 0 & 3 \\ \cdot & \cdot & \cdot & \cdot \end{bmatrix}, \qquad P_2^* = \begin{bmatrix} 0 & 1 & 0 & -1 \\ 0 & 0 & 0 & 2 \end{bmatrix},$$

$$H^{(0)} = \begin{bmatrix} 1 & 5 & 25 & \cdot \\ 5 & 25 & \cdot & \cdot \\ 25 & \cdot & \cdot & \\ \cdot & \cdot & & \\ \cdot & & & \end{bmatrix} = \begin{pmatrix} 1 \\ 5 \\ 25 \\ \cdot \\ \cdot \end{pmatrix} (1, 5, 25, \cdot, \cdot),$$

$$\mathcal{S}_{c\bar{o}} = \text{span}(e_1), \qquad \mathcal{S}_{co} = \text{span}(e_2),$$
$$\mathcal{S}_{\bar{c}\bar{o}} = \text{span}(e_3), \qquad \mathcal{S}_{\bar{c}o} = \text{span}(e_4),$$
$$\hat{T} = [5], \qquad \Omega = [1].$$

TABLE 9.1

Case	Starting vectors	Markov parameters	$l\ r\ m$
(i)	$q = e_1, p = e_1$	$1\,0\,0\,0\,0\cdots$	1 1 1
(ii)	$q = e_2, p = e_2$	$1\,0\,0\,0\,0\cdots$	1 1 2
(iii)	$q = e_3, p = e_2$	$0\,1\,0\,0\,0\cdots$	0 2 3
(iv)	$q = e_3, p = e_1$	$0\,0\,1\,0\,0\cdots$	0 3 3
(v)	$q = e_4, p = e_1$	$0\,0\,0\,1\,0\cdots$	0 4 4
(vi)	$q = p = e_1 + e_2 + e_3 + e_4$	$4\,3\,2\,1\,0\cdots$	2 4 4

Example 9.2.

$$B = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 \end{bmatrix}.$$

- In Table 9.1, case (i) yields benign early termination since q_1 is an eigenvector.
- Case (ii) yields incurable breakdown after one step with $T = [0]$ and this reveals an eigenvalue but no eigenvector.
- In cases (iii), (iv), and (v) the reduction cannot begin and $H^{(0)}$ does not permit triangular factorization. The extended algorithm breaks down incurably after one block step revealing 2, 3, or 4 eigenvalues, respectively.
- Case (vi) suffers curable breakdown at step 2. The extended algorithm yields

$$\hat{T}_4 = \begin{pmatrix} 3 & -1/4 & 0 & 0 \\ -1/4 & -5/16 & 0 & 1 \\ 0 & 0 & 1 & -2 \\ 0 & 1 & -2 & 1 \end{pmatrix}, \qquad \Omega_4 = \begin{pmatrix} 4 & 0 & 0 & 0 \\ 0 & -1/4 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & -2 \end{pmatrix}.$$

The pair $(\hat{T}_2, \Omega_2)$ gives no useful information about B. The Markov parameters of the system (B, q_3, p_3^*) are $(0, 1, -2, 1, 0, \cdots)$. Although $(\hat{T}_4, \Omega_4)$ is equivalent to (B, I), it is less informative than $(\tilde{I}B, \tilde{I})$ but does have smaller bandwidth! Note that the Ω_4 is not unique. The one shown here is produced by the extended Gram–Schmidt algorithm of §3 but another good choice puts the (4,4) entries of T and Ω to -3 and 0, respectively.

10. Monitoring the condition number. For some applications it is meaningful to enrich $\mathbb{C}^n$ with the Euclidean inner product $\langle v, w\rangle := w^* v$, $w \in \mathbb{C}^n, v \in \mathbb{C}^n$. In this setting it is appropriate to normalize all the auxiliary vectors $\{q_i\}$ and $\{p_i\}$ so that

$$\|q_i\|^2 := \langle q_i, q_i\rangle = \langle p_i, p_i\rangle = \|p_i\|^2 = 1, \quad \text{all } i.$$

Let $\sigma_{\min}(X)$ denote the smallest singular value of X. Let

$$\|X\| = \max_{u \neq 0} \|Xu\|/\|u\|,$$

the spectral norm.

THEOREM 10.1. *If the columns of Q_j and P_j are unit vectors and if $P_j^* Q_j = \hat{\Omega} = \operatorname{diag}(\Omega_1, \Omega_2, \cdots, \Omega_k)$, then*

$$\sigma_{\min}(Q_j) \geq \min_{1 \leq i \leq k} \sigma_{\min}(\Omega_i)/\sqrt{j},$$
$$\sigma_{\min}(P_j) \geq \min_{1 \leq i \leq k} \sigma_{\min}(\Omega_i)/\sqrt{j}.$$

Proof. There exist unit vectors $u \in \mathbb{C}^n$, $v \in \mathbb{C}^j$ such that

$$Q_j v = u \sigma_{\min}(Q_j).$$

Then

$$\begin{aligned} \min_{1 \leq i \leq k} \sigma_{\min}(\Omega_i) &\leq \|\hat{\Omega} v\| \\ &= \|P_j^* Q_j v\| \\ &\leq \|P_j^*\| \|Q_j v\| \\ &= \|P_j^*\| \|u \sigma_{\min}(Q_j)\| \\ &\leq \sqrt{j} \cdot \sigma_{\min}(Q_j). \end{aligned}$$

The first inequality uses

$$\|\hat{\Omega} v\|^2 = \sum_{i=1}^{k} \|\Omega_i v^{(i)}\|^2 \geq \min_{1 \leq i \leq k} \sigma_{\min}^2(\Omega_i) \sum_{i=1}^{k} \|v^{(i)}\|^2$$

and $v^{(i)}$ denotes the ith set of entries in v. The last inequality makes use of the relations

$$\|P_j^*\|^2 = \|P_j^* P_j\| = \lambda_{\max}(P_j^* P_j) \leq \operatorname{trace}(P_j^* P_j) = j. \qquad \square$$

When Ω_i is 1×1, then $\sigma_{\min}(\Omega_i) = |\Omega_i|$. In any case the quantities $\|\Omega_i^{-1}\|^{-1}$ are readily computed during or after reduction. When the matrices Q_j and P_j are built up column by column and normalized, it is necessary to compute the quantities $\|r_i\|, \|s_i^*\|$, $i = 1, \cdots, j$. The vectors r_i and s_i^* are defined in the proof of Theorem 2.2. Despite the extra storage required we advocate keeping $\omega_i, \|r_i\|, \|s_i^*\|$, and $\alpha_i = p_i^* B q_i$. The resulting tridiagonal matrix $\hat{T}$ is then defined by

$$\hat{T}(i,i) = \alpha_i, \quad \hat{T}(i,i-1) = \|r_i\| \omega_i, \quad \hat{T}(i-1,i) = \|s_i^*\| \omega_i,$$
$$\Omega = \operatorname{diag}(\omega_1, \cdots, \omega_l).$$

The point is that all the stored quantities give information directly relevant to the stability of the transformation. It is valuable to know whether a small value $\hat{T}(i, i-1)$ comes from a small ω_i or a small $\|r_i\|$.

10.1. A stable reduction. In the block reduction algorithm that would produce the pair $(\hat{T}, \hat{\Omega})$ of §6 there is no need to insist on the lower triangular Hankel form for the blocks

$$\Omega_i = \begin{bmatrix} 0 & 0 & * \\ 0 & * & * \\ * & * & * \end{bmatrix}.$$

When the setting of Euclidean space is appropriate, we can select a tolerance tol and require that

$$\sigma_{\min}(\Omega_i) \geq \text{tol}, \quad \text{all } i.$$

Under this restriction the algorithm will produce in all cases the most refined pair $(\hat{T}_r, \hat{\Omega}_r)$ that satisfies the following conditions:

1. $(\hat{T}_r, \hat{\Omega}_r)$ is a minimal realization.
2. $\sigma_{\min}(Q_r) \geq \text{tol}/\sqrt{r}, \sigma_{\min}(P_r^*) \geq \text{tol}/\sqrt{r}$.

This is a satisfactory method insofar as it extracts as much useful information as possible from the choice (q, p^*). It is not satisfactory insofar as this most compact structure is not known in advance and examples can be constructed in which $\hat{T}_r$ and $\hat{\Omega}_r$ will be of maximal bandwidth. See case (v) in Table 9.1.

11. Comments on implementation.

11.1. Preconditioning for fixed start methods. We know of no way to choose starting pairs (q, p^*) that are guaranteed not to be exceptional. In some applications, e.g., when a sequence of close matrices must be analyzed, good choices for the pair may be known. In the absence of such additional information our theory suggests that q and p^* should be chosen at random from a uniform distribution.

In practice there is merit in having $\hat{T}$ a graded matrix with large entries near the top of the matrix. When $\hat{T}$ has such a structure it is easier to find the eigenvalues of $(\hat{T}, \Omega)$ in a roughly monotonic order (by decreasing absolute value). These considerations suggest that

$$q = B^\nu(\text{ random}), \quad p^* = (\text{random })^* B^\nu, \quad \nu \geq 1,$$

should be preferable to random vectors. On the other hand, each application of B devoted to a starting vector can be regarded as a waste of a step in tridiagonal reduction. Yet an attractive feature of choosing $\nu \geq 1$ is that it forces $\mathcal{K}(q, B) \subset \mathcal{R}(B)$ and $\mathcal{K}(p, B^*) \subset \mathcal{R}(B^*)$. This is essential when B is not just a matrix but an operator with unwanted infinite eigenvalues, as can occur in generalized eigenvalue problems.

We have used $\nu = 1$ in the symmetric case and advocate the same policy here. It is easy to implement and seems to keep breakdowns at bay but we have no theoretical justification for it.

11.2. Stable reduction to block tridiagonal form. When the Euclidean inner product is appropriate, then the idea (in §10) of controlling the condition number of Q and P may be combined with the block reduction of §6. Thus we suppose that $\|q_i\| = \|p_i^*\| = 1$, all i, and there will be diagonal positive definite scaling matrices D_q and D_p. The new matrices Q_j and P_j will satisfy

$$\text{new } \Omega_j = P_j^* Q_j = D_p \Omega_j D_q$$

where Ω_j is the Hankel matrix discussed in §6.

A suitable lower bound tol on $\min_i \sigma_{\min}(\Omega_i)$ may be selected. If ϵ denotes the roundoff unit, then a value such as $\epsilon^{1/2}$ is a natural choice. Then the Lanczos algorithm may be applied in the normal way. However, if, at the end of step j, the new Lanczos vectors q_{j+1} and p_{j+1} satisfy $\omega_{j+1} := p_{j+1}^* q_{j+1} <$tol the algorithm uses $B^\nu q_{j+1}$ for $q_{j+\nu+1}$ and $p_{j+1}^* B^\nu$ for $p_{j+\nu+1}^*$ until

$$\sigma_{\min}[H_\nu^{(0)}(B, q_{j+1}, p_{j+1}^*)] \geq \text{tol}.$$

The $\hat{T}$ and Ω resulting from this strategy differ from the convenient "symmetric Hessenberg" form by matrices of small norm. This should facilitate the solution of the auxiliary problem $(\hat{T} - \lambda\Omega)u = \mathbf{0}$.

For determining eigenvalues alone it is the ratios $|\omega_{i+1}/\omega_i|^{1/2}$ that matter. While these quantities and the $|\alpha_i/\omega_i|$ remain bounded by a small multiple of $\|B\|$ the reduced pair (T_j, Ω_j) is completely satisfactory.

11.3. Local change to the starting pair. If serious breakdown occurs late in the reduction to tridiagonal form, then a restart with a new (random) starting pair amounts to a complete write-off of the expenses of the first attempt. Geist, Lu, and Wachspress, in [8], have studied a compromise in which the cost of reduction with the new pair is very small compared with the initial reduction. However, the theory developed in this essay shows that there are limitations on this technique. Recall the four indices that characterize the realization: l, r, m, and n.

The limitation is that, although l may be increased (which is good), both r and m either remain the same or decrease.

To justify these comments we describe briefly the way that Geist, Lu, and Wachspress carry out the reduction. They perform a sequence of explicit similarity transformations on B; at step j, row j and column j of the current matrix are put into tridiagonal form. Let T_j be the $j \times j$ tridiagonal obtained at the end of step j and suppose that serious breakdown is detected with $B(j+1, j) = 0$, $B(j, j+1) \neq 0$. Their remedy is to apply an elementary similarity transformation on the first two rows and columns of the current array. The transformation matrix is of the form

$$\left[\begin{pmatrix} 1 & 0 \\ \xi & 1 \end{pmatrix} \oplus I_{n-2}\right] \quad [B] \left[\begin{pmatrix} 1 & 0 \\ -\xi & 1 \end{pmatrix} \oplus I_{n-2}\right]$$

and brings nonzero values into position $(3, 1)$. This bulge in the tridiagonal form is chased down the matrix, from $(3, 1)$ to $(4, 2)$ to $(5, 3)$ $\cdots$, in a way that is familiar to those who have studied the symmetric tridiagonal QL algorithm. The cost of this chasing procedure is a small multiple of j and the result is a new value in position $(j + 1, j)$. If the new value is also tiny, then the whole procedure may be tried again.

We claim:

1. The recovering procedure is equivalent to replacing q_1 by $q_1 - \xi q_2 = \theta q_1 + \eta B q_1$, while leaving p_1 unchanged.
2. Let $\mathcal{K}(v) = \text{span}(v, Bv, B^2v, \cdots)$. Then

$$\begin{aligned} \mathcal{K}((\theta I + \eta B)q_1) &= \{(\theta I + \eta B)\phi(B)q_1 : \phi \text{ ranges over all polynomials}\} \\ &\subset \mathcal{K}(q_1). \end{aligned}$$

   Thus

$$(\mathcal{K}(p_1^*)) \cap \mathcal{K}((\theta I + \eta B)q_1\xi) \subset (\mathcal{K}(p_1^*) \cap \mathcal{K}(q_1)) = \mathcal{S}_{co},$$

   and the new values of m and r (see §8) cannot be increased and may decrease.
3. $H_{l+1}^{(0)}(\theta q + \eta B q, p^*, B) = \theta H_{l+1}^{(0)}(q, p^*, B) + \eta H_{l+1}^{(1)}(q, p^*, B)$.
   Thus the new $H_{l+1}^{(0)}$ need not be singular.

12. Conclusion. The last few sections may have deflected the reader's attention away from the big picture so we take the opportunity to recapitulate the main points. Although the canonical structure theorem supplies the right decomposition for understanding incurable breakdown, yet linear systems theory is not itself relevant. The

matrix B is given and, once the vector q and linear functional p^* are chosen, then the "moment" matrix $H^{(0)}$ is fixed and it determines the numbers l and r while B, q, and p^* determine m. See §5 for definitions.

In the generic case $l = r = m = n$ and tridiagonalization succeeds although it need not be stable. However, §10 shows a natural way to monitor the condition number of the transformation and this greatly improves the situation because some of the algorithms proposed for reduction hide the onset of instability by using drastic but hidden scaling of the columns of P and Q to force ${P_j}^* Q_j = I_j$.

In the nongeneric case, there is always a block triangular factorization of $H^{(0)}$ with minimal sizes for the diagonal blocks. We have, in effect, lifted Gragg's block form back to the GS procedure to produce our extended GS algorithm, which shows how to continue the algorithm by working with several vectors simultaneously.

This is not the only way to overcome a curable breakdown. Gutknecht's approach in [14] may be described briefly in the following way. Recall that before breakdown occurs $K_k(q) = Q_k {L_k}^t$ and, with $L_k^{-1} := (\lambda_{ij})$, for $j < k$,

$$q_{j+1} = K_k {L_k}^{-t} e_{j+1} = \sum_{i=1}^{j+1} (B^{i-1} q_1) \lambda_{j+1,i} = \phi_j(B) q_1$$

where

$$\phi_j(t) = \sum_{i=1}^{j+1} \lambda_{j+1,i}\, t^{i-1}, \qquad \lambda_{j+1,j+1} = 1,$$

is a monic polynomial of degree j and is sometimes called the jth (generalized) Lanczos polynomial. The $\{\phi_j\}_1^k$ is called a sequence of **formal** orthogonal polynomials. The proper L^2 inner product function $\langle f, g \rangle$ is replaced by an appropriate linear functional on the pointwise product $f g$. The three-term recurrence connecting standard orthogonal polynomials extends to this more general setting—in the absence of breakdown. Gutknecht shows how to define polynomials ϕ_{k+1}, $\phi_{k+2}, \cdots$ (and hence rows of L^{-1}) after each curable breakdown in such a way as to preserve as much as possible of the standard recurrence relations. As mentioned in §6 the result is equivalent to choosing L, whenever there is freedom, to make the diagonal blocks Ω_i in Theorem 6.1 antidiagonal, i.e., all zeros except along the secondary (NE–SW) diagonal.

By accepting blocks that are larger than minimal size given in §6, it is straightforward to set an upper bound on the condition number of Q and P and then to compute the most refined block tridiagonal form $(\hat{T}, \Omega)$ consistent with this bound. Nevertheless triangular factorization cannot proceed beyond the effective rank of $H^{(0)}$ because the Schur complement of ${H_r}^{(0)}$ vanishes. It is at this stage that the Kalman–Gilbert theorem can be invoked to show that $(\hat{T}_r, \Omega_r)$ is a minimal realization of B, q, p^*. In other words, for some conformable r-vectors $\tilde{q}$ and $\tilde{p}$, and for all σ in the resolvent set,

$$\tilde{p}^* (\sigma\, \Omega_r - \hat{T}_r)^{-1}\, \tilde{q} = p^* (\sigma\, I - B)^{-1} q.$$

It is this transfer function approach that shows immediately that each eigenvalue of $\hat{T}$, Ω is an eigenvalue of B.

For specialists in large eigenvalue problems, it is unnerving to have a solution λ to $Bz = z\lambda$ without knowledge of an appropriate invariant subspace. To algebraists who invoke the characteristic polynomial, this absence of z is a natural state of affairs.

Our approach stops with a minimal realization but there are further questions to be asked. If $r < m$ then how is it possible to append further columns q_{r+1}, $q_{r+2}, \cdots$ or further functionals $p^*{}_{r+1}$, $p^*{}_{r+2}, \cdots$ to obtain bases for the controllable and observable subspaces? This problem is addressed in [2]. Their idea is to modify the Lanczos algorithm a little.

After incurable breakdown at step r the remaining pair q_{r+1}, p^*_{r+1} is used to generate more vectors according to

$$q_{r+i+1} = (I - \Pi_r)Bq_{r+i}, \qquad p^*_{r+i+1} = p^*_{r+i}B(I - \Pi_r),$$

until linear independence is lost. The new q's span $\mathcal{S}_{c\bar{o}}$ and the new p's span $\mathcal{S}_{\bar{c}o}$. The final space $\mathcal{S}_{\bar{c}\bar{o}}$ must be obtained, if needed, by more primitive means. It is the null space of the matrix

$$(p_1, \cdots, p_r, p_{r+1}, \cdots, p_m, q_{r+1}, \cdots, q_l)^*$$

obtained by appending the new p's and q's to the columns of P_r. The extra coefficients generated by these modifications permit the calculation of the canonical form in §7 and with it the remaining eigenvalues and all the eigenvectors.

Although the moment matrix $H^{(0)}$ is an essential theoretical tool it is not available in practice nor indeed are the Krylov matrices $K(q, B)$ and $K(p^*, B)$. From the practical point of view the essential, impressive insight of Cornelius Lanczos was to recognize that Bq_j is preferable to $B^j q_1$, and $p_j^* B$ to $p_1^* B^j$, for the purpose of computing Q, P, $\hat{T}$, and Ω.

Plenty of questions remain unanswered. Is it advisable to put more effort into the choice of q_1 and p_1^* ? Is it better to restart after an early serious breakdown or to continue with a block tridiagonal output? What are good ways to compute, or update, the partial eigensolution of the reduced problem $(\hat{T} - \lambda\Omega)s = 0$? Can the residual error bounds presented in [17] be used for terminating the reduction when only a few eigenpairs of B are wanted? How should we compensate for the effect of roundoff error?

This paper sought to provide a good framework for looking at reduction to tridiagonal form.

REFERENCES

[1] G. Birkhoff and S. Maclane, *A Survey of Modern Algebra*, Chapter XV, Macmillan, New York, 1953.

[2] D. Boley and G. H. Golub, *The nonsymmetric Lanczos algorithm and controllability*, Tech. Report NA-90-06, Computer Science Department, Stanford University, Stanford, CA, May 1990.

[3] A. Draux, *Polynômes Orthogonaux Formels-Applications*, Lecture Notes in Mathematics, Vol. 974, Springer-Verlag, Berlin, Heidelberg, New York, 1983.

[4] V. N. Faddeeva, *Computational Methods of Linear Algebra*, Dover, New York, 1959.

[5] J. G. F. Francis, *The* QR *transformation, Part* I, Comput. J., 4 (1961), pp. 265–271.

[6] ———, *The* QR *transformation, Part* II, Comput. J., 4 (1962), pp. 332–345.

[7] G. A. Geist, *Reduction of a general matrix tridiagonal form*, Tech. Report, ORNL/TM-10991, Oak Ridge National Laboratory, Oak Ridge, TN, March 1989.

[8] G. A. Geist, A. Lu and E. L. Wachspress, *Stabilized Gaussian reduction of an arbitrary matrix to tridiagonal form*, Tech. Report ORNL/TM-11089, Oak Ridge National Laboratory, Oak Ridge, TN, June 1989.

[9] E. Gilbert, *Controllability and observability in multivariable control systems*, SIAM J. Control, 1 (1963), pp. 128–151.

[10] G. H. Golub and C. Van Loan, *Matrix Computations*, Second Edition, The Johns Hopkins University Press, Baltimore, MD, 1989.

[11] G. H. GOLUB AND M. H. GUTKNECHT, *Modified moments for indefinite weight functions*, Interdisciplinary Project for Supercomputing Research Report No. 89-04, ETH-Zentrum, CH-8092 Zurich, 1989.
[12] W. B. GRAGG, *Matrix interpretations and applications of the continued fraction algorithm*, Rocky Mountain J. Math., 4 (1974), pp. 213–225.
[13] W. B. GRAGG AND A. LINDQUIST, *On the partial realization problem*, Linear Algebra Appl. 50 (1985), pp. 277–319.
[14] M. H. GUTKNECHT, *A completed theory of the unsymmetric Lanczos process and related algorithms: Part* I, SIAM J. Matrix Anal. Appl., this issue, pp. 594–639.
[15] A. S. HOUSEHOLDER, *The Theory of Matrices in Numerical Analysis*, Blaisdell, New York, 1964.
[16] I. S. IOHVIDOV, *Hankel and Toeplitz Matrices and Forms*, G.P.A. Thijsse, trans., Birkhäuser-Verlag, Basel, 1982.
[17] W. KAHAN, B. N. PARLETT, AND E. JIANG, *Residual error bounds on approximate eigensystems of nonnormal matrices*, SIAM J. Numer. Anal., 19 (1982), pp. 470–484.
[18] T. KAILATH, *Linear Systems*, Prentice-Hall, Englewood Cliffs, NJ, 1980.
[19] R. E. KALMAN, *Mathematical description of linear systems*, SIAM J. Control, 1 (1963), pp. 152–192.
[20] ———, *On partial realizations, transfer functions and canonical forms*, Acta Polytech. Scand. Math. Comput. Sci. Ser., MA31 (1979), pp. 9–32.
[21] C. D. LABUDDE, *The reduction of an arbitrary real sparse matrix to tridiagonal form using similarity transformations*, Math. Comp., 17 (1963), pp. 443–447.
[22] C. LANCZOS, *An iteration method for the solution of the eigenvalue problem of linear differential and integral operators*, J. Res. Nat. Bur. Standards, 45, (1950), pp. 255–282 (see pp. 266–270).
[23] H. LEV-ARI AND T. KAILATH, *Triangular factorization of structured Hermitian matrices*, Operator Theory: Adv. Appl., 18 (1986), pp. 301–324.
[24] B. N. PARLETT, *A note on LaBudde's algorithm*, Math. Comp., 19 (1964), pp. 505–506.
[25] B. N. PARLETT, D. R. TAYLOR AND Z.-S. LIU, *A look-ahead Lanczos algorithm for unsymmetric matrices*, Math. Comp., 44 (1985), pp. 105–124.
[26] C. STRACHEY AND J. G. F. FRANCIS, *The reduction of a matrix to codiagonal form by elimination*, Comput. J., 4 (1961), pp. 168–176.
[27] D. R. TAYLOR, *Analysis of the look ahead Lanczos algorithm*, Ph.D. thesis, Center for Pure and Applied Mathematics, University of California, Berkeley, CA, 1982. Also as Tech. Report CPAM-108, Center for Pure and Applied Mathematics, University of California, Berkeley, CA.
[28] E. L. WACHSPRESS, ADI *solution of Lyapunov equations*, presented at MSI Workshop on Practical Iterative Methods for Large-Scale Computations, Minneapolis, MN, October 1988.
[29] J. H. WILKINSON, *Instability of the elimination method of reducing a matrix to tridiagonal form*, Comput. J., 5 (1962), pp. 61–70.
[30] ———, *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.