SIAM J. NUMER. ANAL.
Vol. 20, No. 1, February 1983

0036-1429/83/2001-0001 $01.25/0

ORDINARY DIFFERENTIAL EQUATIONS AND THE SYMMETRIC EIGENVALUE PROBLEM*

P. DEIFT,† T. NANDA‡ AND C. TOMEI§

Abstract. In this paper the authors develop a general framework for calculating the eigenvalues of a symmetric matrix using ordinary differential equations. New algorithms are suggested and old algorithms, including *QR*, are interpreted.

1. Introduction. The purpose of this paper is to draw the attention of numerical analysts to a general framework for calculating the eigenvalues of a real symmetric matrix which is suggested by recent work in the dynamical theory of tridiagonal matrices. The idea is as follows:

Consider the flow given by

$$\begin{aligned}&\frac{d}{dt}a_k = 2(b_k^2 - b_{k-1}^2),\\ &\frac{d}{dt}b_k = b_k(a_{k+1}-a_k), \qquad k=1,\cdots,n,\end{aligned} \tag{1}$$

$(b_0 = b_n = 0)$ on the set of real tridiagonal matrices

$$L(a,b)=\begin{pmatrix} a_1 & b_1 & & & \\ b_1 & & & 0 & \\ & & \ddots & & \\ & & & & b_{n-1} \\ 0 & & & b_{n-1} & a_n \end{pmatrix}.$$

The flow has the properties:

(i) The eigenvalues of $L(a(t), b(t))$ are independent of t, i.e., the flow is isospectral,

(ii) $\lim_{t\to\pm\infty} b_k(t) = 0$, $k = 1, \cdots, n-1$.

Consequently, $L(t) \equiv L(a(t), b(t))$ converges as $t \to \pm\infty$ to a diagonal matrix which is necessarily some permutation of the eigenvalues of the initial matrix $L(0)$.

(Equations (1) were introduced by Flaschka [1] in order to study the Toda lattice of statistical mechanics (see § 2) and further analyzed by Moser in [2]. Property (i) is proved in [1] and property (ii) in [2].)

The above scheme turns the problem of calculating the eigenvalues of a tridiagonal matrix (and hence of a general real symmetric matrix by a standard tridiagonalization algorithm) into a problem in the theory of ordinary differential equations. This allows us to introduce ideas from phase space analysis and provides a suggestive picture, as we will see, of the calculational process.

The general framework is as follows: corresponding to each function $G(\lambda)$, real and injective on the spectrum of L, there exists an isospectral flow on the space of

* Received by the editors September 28, 1981, and in revised form February 19, 1982.

† Courant Institute of Mathematical Sciences, New York University, New York, New York 10012. The research of this author was supported by the National Science Foundation under grant MCS-8002561.

‡ Courant Institute of Mathematical Sciences, New York University, New York, New York 10012. The research of this author was supported by the Department of Energy under grant DE-AC02-76ER03077.

§ Courant Institute of Mathematical Sciences, New York University, New York 10012. The research of this author was supported by C.N.Pq. and PUC/RJ, Brazil.

tridiagonal matrices, convergent (exponentially) to a diagonal matrix as $t \to \pm\infty$. Setting $G(\lambda)=\lambda$, we recover the Toda flow (1). Setting $G(\lambda)=\log\lambda$, we obtain a flow (the *QR flow*) whose values at integer times k are equal to the kth iteration of the QR algorithm starting from L. (Setting $G(\lambda)=s\log\lambda$, we recover the QR power method. In particular $s=\frac{1}{2}$ gives LR.) In fact we have the following general result: let $L(k;G)$ be the evaluation at time $t=k$ of the flow corresponding to G with $L(0;G)=L$. Then $e^{G(L(k;G))}$ is equal to the kth iteration of QR starting from $\exp G(L)$. Setting $G(\lambda)=\log\lambda$ we recover the QR result above. Setting $G(\lambda)=\lambda$ we recover a result of Symes [3]. The proof of this general result, which holds for arbitrary, not necessarily tridiagonal, real symmetric matrices, will be given in a forthcoming paper. The paper will also contain a discussion of the nonselfadjoint case, as well as some results on infinite matrices.

Sections 2 and 3 contain the relevant theoretical results. Section 4 describes the calculation of the eigenvalues of some test matrices using (1) and compares the results with those obtained by EISPACK. Section 5 contains further numerical experiments suggested by the phase space picture. The final section is a collection of remarks and suggestions.

It is our hope that the evidence provided by the results of this paper will give an impetus to the further study of the symmetric eigenvalue problem as a dynamical process.

2. The Toda flow. Following Flaschka [1], rewrite (1) in the form

$$\frac{d}{dt}L = BL - LB, \tag{2}$$

where

$$B=\begin{pmatrix} 0 & b_1 & & & & \\ -b_1 & 0 & b_2 & & 0 & \\ & -b_2 & & & & \\ & & \cdot & & & \\ & & & \cdot & & \\ & 0 & & & \cdot & b_{n-1} \\ & & & & -b_{n-1} & 0 \end{pmatrix}.$$

Let $V(t)$ be the solution of the matrix equation

$$\frac{d}{dt}V = BV \quad \text{with } V(0)=1. \tag{3}$$

As $B(t)$ is antisymmetric, $V(t)$ is unitary and one easily checks that

$$L(t)=V(t)L(0)V^{-1}(t). \tag{4}$$

This proves

THEOREM 1. *The flow* (1) *is isospectral.*

Henceforth, we will assume $b_i>0$, $1,\cdots,n-1$ (this property is clearly preserved under the Toda flow). It follows easily (see e.g. [4]) that all the eigenvalues of L are distinct and that the first component of each eigenvector is not zero. We adopt the following convention: The eigenvalues are ordered according to $\lambda_1>\lambda_2>\cdots>\lambda_n$ and the first component u_{1i} of the ith normalized eigenvector $(u_{1i}, \cdots, u_{ni})^T$ is (strictly) positive. In what follows, u will always denote the vector $(u_{11}, \cdots, u_{1n})^T$. Also $\|\cdot\|$ denotes the Euclidean norm in $\mathbb{R}^n$. Let

$$M = \{(\lambda_1, \cdots, \lambda_n : x_1, \cdots, x_n) \in R^{2n} : \lambda_1 > \lambda_2 > \cdots > \lambda_n; \quad x_i > 0, \sum_{i=1}^n x_i^2 = 1\}.$$

Theorem 2 below links tridiagonal matrices L to the top components of their normalized eigenvectors. J. Moser established the result in [2]. A different proof was given in [4] and called the inverse algorithm. It turns out that the connection has been known to some numerical analysts (e.g. Lanczos, Householder, Wilkinson) for nearly 30 years, but in the following more general formulation: if any symmetric matrix A is reduced to tridiagonal form L by an orthogonal similarity S, i.e. $S^TAS = L$, $S^TS = I$, then both L and S are completely determined by Se_1, the first column of S. Now if we take $A = \Lambda = \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$, then indeed $Se_1 = (u_{11}, \cdots, u_{1n})^T$. The proof just invokes the Lanczos algorithm. When $A = \Lambda$ then the algorithm delivers the components of L in terms of $\{\lambda_j, u_{1j}, j = 1, \cdots, n\}$.

We sketch our proof of Theorem 2 for the sake of readers unfamiliar with the Lanczos algorithm. In the body of the paper we refer to Theorem 2 as the inverse algorithm.

THEOREM 2. *The map* $(a_1, \cdots, a_n; b_1, \cdots, b_{n-1}) \to (\lambda_1, \cdots, \lambda_n; u_{11}, \cdots, u_{1n})$ *is a diffeomorphism from the space of real tridiagonal matrices with* $b_i > 0$ *onto* M.

Proof. By the spectral theorem,

$$L = U\Lambda U^T,$$

where $\Lambda = \operatorname{diag}(\lambda_1, \cdots, \lambda_n)$ and $U = (u_{ij})$ is the orthogonal matrix of the eigenvectors. From this it follows that

$$(5) \qquad a_i = L_{ii} = \sum_{j=1}^n \lambda_j u_{ij}^2, \qquad i = 1, 2, \cdots, n,$$

$$(5') \qquad b_i = L_{ii+1} = \sum_{j=1}^n \lambda_j u_{ij} u_{i+1j}, \qquad i = 1, 2, \cdots, n-1.$$

From the eigenvalue equation

$$(6) \qquad b_i u_{i+1j} = (\lambda_j - a_i)u_{ij} - b_{i-1}u_{i-1j}, \qquad 1 \leqq i \leqq n, \quad 1 \leqq j \leqq n \quad (b_0 = b_n = 0)$$

and the orthogonality of U, it follows that

$$(7) \qquad b_i^2 = \sum_{j=1}^n [(\lambda_j - a_i)u_{ij} - b_{i-1}u_{i-1j}]^2, \qquad i = 1, 2, \cdots, n-1,$$

and

$$u_{i+1j} = [(\lambda_j - a_i)u_{ij} - b_{i-1}u_{i-1j}]/b_i, \qquad i = 1, \cdots, n-1.$$

These facts immediately prove that the above map is one-to-one. Indeed, a_1 is determined by $\lambda_1, \cdots, \lambda_n$ and $u_{11}, u_{12}, \cdots, u_{1n}$ through the identity

$$a_1 = \sum_{j=1}^n \lambda_j u_{1j}^2.$$

Then, b_1 and u_{2j} are determined from

$$b_1^2=\sum_{j=1}^{n}[(\lambda_j-a_1)u_{1j}-0]^2, \qquad u_{2j}=[(\lambda_j-a_1)u_{1j}-0]/b_1$$

respectively, and so on. This algorithm also suggests how to prove the onto character of the map: the details can be found, for example, in [4]. □

Remark 1. In § 4 we will use the above theorem to generate tridiagonal matrices with prescribed spectrum and $b_i>0$.

From this theorem we see that the space of tridiagonal matrices ($b_i>0$) with fixed spectrum $\lambda_1>\lambda_2>\cdots>\lambda_n$ is diffeomorphic to the subset of the unit sphere $\sum_{i=1}^n x_i^2=1$ with $x_i>0$ (and hence diffeomorphic to R^{n-1}). In particular for $n=3$ we have the picture in Fig. 1.

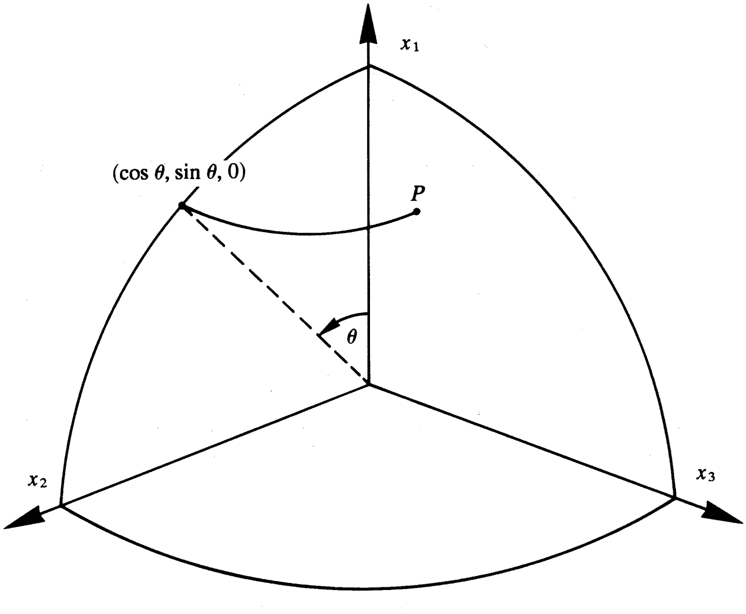

FIG. 1

Suppose $P=(u_{11}, u_{12}, u_{13})$ approaches an edge (but not a vertex): say $P\to(\cos\theta, \sin\theta, 0)$, $0<\theta<\pi/2$. Then the corresponding tridiagonal matrices converge to a matrix of the form

$$\begin{pmatrix} a_1 & b_1 & 0 \\ b_1 & a_2 & 0 \\ 0 & 0 & a_3 \end{pmatrix}$$

with $b_1>0$ and $a_3=\lambda_3$. To see this, consider $P(\varepsilon)=(\alpha_1(\varepsilon), \alpha_2(\varepsilon), \varepsilon)$ where $\varepsilon>0$, $\alpha_1(\varepsilon)^2+\alpha_2(\varepsilon)^2+\varepsilon^2=1$ and $\alpha_1(\varepsilon)\to\cos\theta$, $\alpha_2(\varepsilon)\to\sin\theta$ as $\varepsilon\to 0$. Then, from (6), $u_{2j}=(\lambda_j-a_1)u_{1j}/b_1$, we see that $(u_{21}, u_{22}, u_{23})=(\beta_1(\varepsilon), \beta_2(\varepsilon), \beta_3(\varepsilon))$, where $\beta_3(\varepsilon)=O(\varepsilon)$, so that[1] $(u_{21}, u_{22}, u_{23})\to(\sin\theta, -\cos\theta, 0)$ as $\varepsilon\to 0$. Finally, as (u_{31}, u_{32}, u_{33}) is the cross product of (u_{11}, u_{12}, u_{13}) and (u_{21}, u_{22}, u_{23}), up to a sign, $(u_{31}, u_{32}, u_{33})=(\gamma_1(\varepsilon), \gamma_2(\varepsilon),$

[1] The possibility $(u_{21}, u_{22}, u_{23})\to(-\sin\theta, \cos\theta, 0)$ is excluded because $a_1=\sum_{i=1}^3\lambda_i u_{1i}^2\leqq\lambda_1$ so that $u_{21}=(\lambda_1-a_1)u_{11}/b_1$ is clearly positive.

$\gamma_3(\varepsilon))$, where $\gamma_1(\varepsilon)$ and $\gamma_2(\varepsilon)$ are $O(\varepsilon)$ as $\varepsilon \to 0$. Now, by the inverse algorithm (in particular equation (6)), the tridiagonal matrix corresponding to $P(\varepsilon)$ is

$$L(\varepsilon)=\begin{pmatrix}\alpha_1(\varepsilon) & \alpha_2(\varepsilon) & \varepsilon\\ \beta_1(\varepsilon) & \beta_2(\varepsilon) & \beta_3(\varepsilon)\\ \gamma_1(\varepsilon) & \gamma_2(\varepsilon) & \gamma_3(\varepsilon)\end{pmatrix}\begin{pmatrix}\lambda_1 & 0 & 0\\ 0 & \lambda_2 & 0\\ 0 & 0 & \lambda_3\end{pmatrix}\begin{pmatrix}\alpha_1(\varepsilon) & \beta_1(\varepsilon) & \gamma_1(\varepsilon)\\ \alpha_2(\varepsilon) & \beta_2(\varepsilon) & \gamma_2(\varepsilon)\\ \varepsilon & \beta_3(\varepsilon) & \gamma_3(\varepsilon)\end{pmatrix},$$

which converges as $\varepsilon \to 0$ to

$$L(0)=\begin{pmatrix}\cos\theta & \sin\theta & 0\\ \sin\theta & -\cos\theta & 0\\ 0 & 0 & \pm 1\end{pmatrix}\begin{pmatrix}\lambda_1 & 0 & 0\\ 0 & \lambda_2 & 0\\ 0 & 0 & \lambda_3\end{pmatrix}\begin{pmatrix}\cos\theta & \sin\theta & 0\\ \sin\theta & -\cos\theta & 0\\ 0 & 0 & \pm 1\end{pmatrix}$$

$$=\begin{pmatrix}a_1 & b_1 & 0\\ b_1 & a_2 & 0\\ 0 & 0 & \lambda_3\end{pmatrix}.$$

If $P\to(0,\cos\phi,\sin\phi)$ or $(\cos\psi, 0, \sin\psi)$, $0<\phi$, $\psi<\pi/2$, a similar calculation shows that the corresponding matrices also converge to matrices of the *same* form

$$\begin{pmatrix}a_1 & b_1 & 0\\ b_1 & a_2 & 0\\ 0 & 0 & a_3\end{pmatrix}$$

but with $a_3=\lambda_1$ and λ_2 respectively.

On the other hand, as P converges to the vertex $(1, 0, 0)$, say, the corresponding matrices do not necessarily converge. The element b_1, however, always converges to 0 by (5) and (7),

$$a_1=\sum_{j=1}^{3}\lambda_j u_{1j}^2,\qquad b_1^2=\sum_{j=1}^{3}(\lambda_j-a_1)^2u_{1j}^2,$$

for if $(u_{11}, u_{12}, u_{13})\to(1, 0, 0)$, then $a_1\to\lambda_1$ and hence $b_1\to 0$. Thus in a, b variables, all limit points as $P\to(1, 0, 0)$ are matrices of the form

$$\begin{pmatrix}a_1 & 0 & 0\\ 0 & a_2 & b_2\\ 0 & b_2 & a_3\end{pmatrix}=\begin{pmatrix}1 & 0 & 0\\ 0 & \cos\xi & \sin\xi\\ 0 & \sin\xi & -\cos\xi\end{pmatrix}\begin{pmatrix}\lambda_1 & 0 & 0\\ 0 & \lambda_2 & 0\\ 0 & 0 & \lambda_3\end{pmatrix}\begin{pmatrix}1 & 0 & 0\\ 0 & \cos\xi & \sin\xi\\ 0 & \sin\xi & -\cos\xi\end{pmatrix}$$

with $0\leqq\xi\leqq\pi/2$. If $P=(u_{11}, u_{12}, u_{13})\to(1, 0, 0)$ *and* $u_{12}/u_{13}\to\gamma$, $0\leqq\gamma\leqq\infty$, then, from (6),

$$\frac{u_{22}}{u_{23}}=\left(\frac{\lambda_2-a_1}{\lambda_3-a_1}\right)\frac{u_{12}}{u_{13}},$$

and by orthogonality,

$$u_{21}=-\frac{1}{u_{11}}(u_{22}u_{12}+u_{23}u_{13}),$$

(u_{21}, u_{22}, u_{23}) converges to $(0, \cos\xi, \sin\xi)$, where

$$\cot\xi=\gamma\left(\frac{\lambda_2-\lambda_1}{\lambda_3-\lambda_1}\right),\qquad 0\leqq\xi\leqq\frac{\pi}{2}.$$

Thus, as $P \to (1, 0, 0)$, the corresponding matrices converge if (and only if, as is easy to see) the ratio u_{12}/u_{13} has a limit. Similarly, the vertices $(0, 1, 0)$ and $(0, 0, 1)$ correspond to intervals of matrices of the same form

$$\begin{pmatrix} a_1 & 0 & 0 \\ 0 & a_2 & b_2 \\ 0 & b_2 & a_3 \end{pmatrix},$$

but now $a_1 = \lambda_2$ and λ_3 respectively.

Rephrasing these facts, we see that the space of 3×3 tridiagonal matrices with $b_i > 0$ and fixed spectrum $\lambda_1 > \lambda_2 > \lambda_3$ is homeomorphic to a hexagon with boundary (i.e., a disk with six preferred boundary points): the interior of the space corresponds to matrices with $b_i > 0$ while the boundary consists of pieces with some $b_i = 0$, arranged as in Fig. 2 (cf. van Moerbeke [5]).

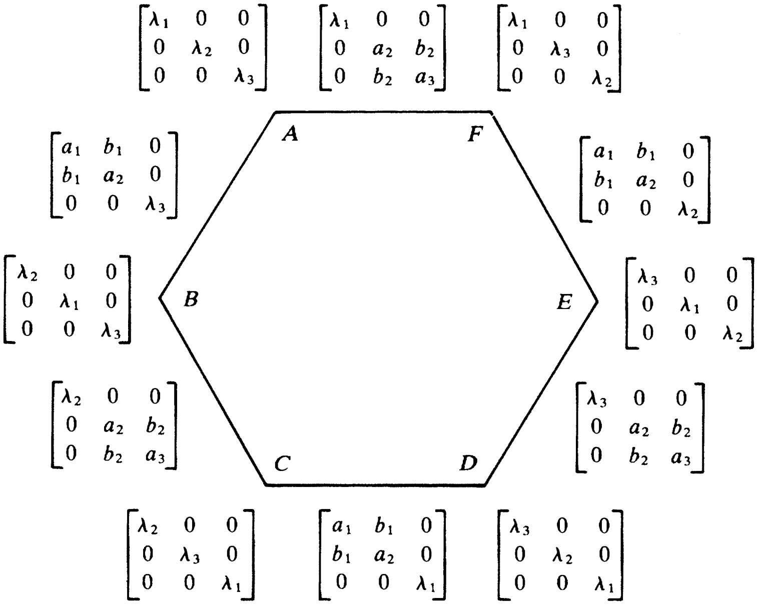
$\begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$
$\begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & a_2 & b_2 \\ 0 & b_2 & a_3 \end{bmatrix}$
$\begin{bmatrix} \lambda_1 & 0 & 0 \\ 0 & \lambda_3 & 0 \\ 0 & 0 & \lambda_2 \end{bmatrix}$
$\begin{bmatrix} a_1 & b_1 & 0 \\ b_1 & a_2 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$
A
F
$\begin{bmatrix} a_1 & b_1 & 0 \\ b_1 & a_2 & 0 \\ 0 & 0 & \lambda_2 \end{bmatrix}$
$\begin{bmatrix} \lambda_2 & 0 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_3 \end{bmatrix}$
B
E
$\begin{bmatrix} \lambda_3 & 0 & 0 \\ 0 & \lambda_1 & 0 \\ 0 & 0 & \lambda_2 \end{bmatrix}$
$\begin{bmatrix} \lambda_2 & 0 & 0 \\ 0 & a_2 & b_2 \\ 0 & b_2 & a_3 \end{bmatrix}$
C
D
$\begin{bmatrix} \lambda_3 & 0 & 0 \\ 0 & a_2 & b_2 \\ 0 & b_2 & a_3 \end{bmatrix}$
$\begin{bmatrix} \lambda_2 & 0 & 0 \\ 0 & \lambda_3 & 0 \\ 0 & 0 & \lambda_1 \end{bmatrix}$
$\begin{bmatrix} a_1 & b_1 & 0 \\ b_1 & a_2 & 0 \\ 0 & 0 & \lambda_1 \end{bmatrix}$
$\begin{bmatrix} \lambda_3 & 0 & 0 \\ 0 & \lambda_2 & 0 \\ 0 & 0 & \lambda_1 \end{bmatrix}$

FIG. 2

AB, CD and EF correspond in u_{1j} variables to the edges $(\cos\theta, \sin\theta, 0)$, $(0, \cos\phi, \sin\phi)$ and $(\cos\psi, 0, \sin\psi)$, $0 \leqq \theta, \phi, \psi \leqq \pi/2$, respectively. FA, BC and DE correspond to the vertices $(1, 0, 0)$, $(0, 1, 0)$ and $(0, 0, 1)$, respectively.

Caveat. The hexagon of Fig. 2 presents only a topological picture and has no metric significance.

Similarly, in the 4×4 case, the space of tridiagonal matrices with $b_i \geqq 0$ is homeomorphic to a truncated octahedron with boundary as shown in Fig. 3, where (i_1, i_2, i_3, i_4) stands for the matrix diag $(\lambda_{i_1}, \lambda_{i_2}, \lambda_{i_3}, \lambda_{i_4})$. The hexagonal faces correspond to matrices with b_1 or $b_3 = 0$ (the 3×3 case), while the square faces (a product of two intervals, one for each 2×2 matrix) corresponds to $b_2 = 0$.

We will use these low-dimensional examples to illustrate the properties of the Toda flow in § 5. It will be clear that the description extends to $n > 4$ as well as to the general flows of § 4.

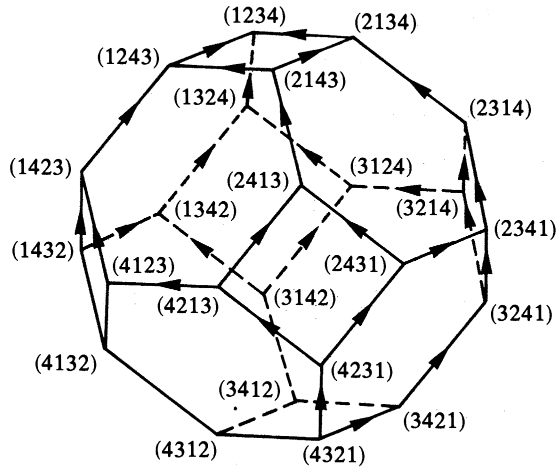

FIG. 3

The following lemma shows that (1) can be solved explicitly in terms of the "inverse" variables (Λ, u) (see also Remark 4).

LEMMA 1 (Moser [2]). *Let $L(t)$ solve* (1). *Then*

$$u(t)=\frac{e^{\Lambda t}u(0)}{\|e^{\Lambda t}u(0)\|}, \qquad -\infty<t<\infty. \tag{8}$$

Proof. From (4), $u_{ij}(t)=\sum_{k=1}^{n} V_{ik}(t)u_{kj}(0)$ and from (3),

$$\frac{d}{dt}u_{ij}(t)=\sum_{k=1}^{n} B_{ik}u_{kj}(t).$$

In particular,

$$\frac{d}{dt}u_{1j}(t)=b_1(t)u_{2j}(t)=(\lambda_j-a_1(t)u_{1j}(t)),$$

i.e.,

$$\frac{d}{dt}u_{1j}(t)=\left(\lambda_j-\sum_{k=1}^{n}\lambda_k u_{1k}^2(t)\right)u_{1j}(t). \tag{9}$$

The lemma follows directly upon substituting (8) into (9). □

Let $\mu=\min_{k=1,\cdots,n-1}(\lambda_k-\lambda_{k+1})$.

THEOREM 3.

$$\left.\begin{aligned}|a_k(t)-\lambda_k|&\leqq c\exp\{-2\mu t\},\\ b_k^2(t)&\leqq c\exp\{-2\mu t\},\end{aligned}\right. \qquad \textit{for } t\geqq 0$$

where c depends on $L(0)$.

Proof. By the explicit solution (8), $\{u_{11}(t), u_{12}(t), \cdots, u_{1n}(t)\}$ converges exponentially to $(1,0,\cdots,0)$. The idea of the proof is to use induction and the inverse algorithm to show that the full matrix $(u_{ij}(t))$ of eigenvectors converges exponentially to the identity matrix.

The proof is in parts; below, the constant c depends only on $L(0)$.

$$\text{(a)} \qquad \begin{array}{l} 1-u_{11}^2(t)\leqq c\exp\{-2\mu t\}, \\ u_{1j}^2(t)\leqq c\exp\{-2\mu t\}, \end{array} \qquad 2\leqq j\leqq n.$$

This follows immediately from Lemma 1.

$$\text{(b)} \qquad |a_1(t)-\lambda_1|\leqq c\exp\{-2\mu t\}.$$

This follows from

$$|a_1(t)-\lambda_1| = \left|\sum_{j=2}^{n}(\lambda_j-\lambda_1)u_{1j}^2(t)\right|.$$

$$\text{(c)} \qquad b_1^2(t)\leqq c\exp\{-2\mu t\}.$$

This follows from (a), (b) and (7),

$$b_1^2(t)=\sum_{j=1}^{n}(\lambda_j-a_1)^2u_{1j}^2(t).$$

$$\text{(d)} \qquad u_{21}^2(t)\leqq c\exp\{-2\mu t\}.$$

By orthogonality, $u_{21}(t)=-(\sum_{j=2}^n u_{1j}(t)u_{2j}(t))/u_{11}(t)$. The inequality now follows from (a) and $\sum_{j=2}^n u_{2j}^2(t)\leqq 1$.

$$\text{(e)} \qquad |1-u_{22}^2(t)|\leqq c\exp\{-2\mu t\}, \qquad u_{2j}^2(t)\leqq c\exp\{-2\mu t\}, \qquad j\geqq 3.$$

From (6),

$$\frac{u_{2j}(t)}{u_{2k}(t)}=\frac{(\lambda_j-a_1)u_{1j}(t)}{(\lambda_k-a_1)u_{1k}(t)}=\frac{(\lambda_j-a_1)}{(\lambda_k-a_1)}\frac{u_{1j}(0)}{u_{1k}(0)}\exp\{(\lambda_i-\lambda_k)t\}.$$

The inequalities now follow from (b), (d) and

$$\sum_{\substack{j=1\\ j\neq 2}}^{n} u_{2j}^2(t)=1-u_{22}^2(t).$$

$$\text{(f)} \qquad |a_2(t)-\lambda_2|\leqq c\exp\{-2\mu t\}.$$

Use $|a_2-\lambda_2|=|\sum_{j=1,j\neq 2}^n(\lambda_j-\lambda_2)u_{2j}^2(t)|$, (d) and (e).

$$\text{(g)} \qquad b_2^2(t)\leqq c\exp\{-2\mu t\}.$$

Use $b_2^2(t)=\sum_{j=1}^n[(\lambda_j-a_2)u_{2j}(t)-b_1u_{1j}(t)]^2$, (c), (e) and (f).

$$\text{(h)} \qquad b_1(t)u_{1s}(t)/u_{2s}(t)\to 0 \quad \text{as } t\to\infty, \quad s\geqq 2.$$

This follows from (6), $b_1(t)u_{1s}(t)/u_{2s}(t)=b_1^2(t)/(\lambda_s-a_1(t))$, (b) and (c).

Now let $j\geqq 3$ and assume by induction that

$$\begin{array}{l} |a_k(t)-\lambda_k|\leqq c\exp\{-2\mu t\}, \qquad k\leqq j-1, \\ b_k^2(t)\leqq c\exp\{-2\mu t\}, \qquad k\leqq j-1, \\ b_{k-1}(t)u_{k-1s}(t)/u_{ks}(t)\to 0 \quad \text{as } t\to\infty, \quad k\leqq j-1, \quad s\geqq k, \\ |1-u_{kk}^2(t)|\leqq c\exp\{-2\mu t\}, \qquad k\leqq j-1, \\ u_{kl}^2(t)\leqq c\exp\{-2\mu t\}, \qquad k\leqq j-1, \quad l\neq k, \\ \dfrac{u_{kr}^2(t)}{u_{ks}^2(t)}\sim c\exp\{2(\lambda_r-\lambda_s)t\} \quad \text{as } t\to\infty, \quad k\leqq j-1, \quad r,s\geqq k. \end{array}$$

We prove that the same relations hold for $k=j$. Let $r, s \geqq j$. Then from (6) we have

$$\frac{u_{jr}(t)}{u_{js}(t)}=\frac{(\lambda_r-a_{j-1})u_{j-1r}(t)-b_{j-2}u_{j-2r}(t)}{(\lambda_s-a_{j-1})u_{j-1s}(t)-b_{j-2}u_{j-2s}(t)}$$

$$=\frac{u_{j-1r}(t)}{u_{j-1s}(t)}\,\frac{(\lambda_r-a_{j-1})-b_{j-2}\dfrac{u_{j-2r}(t)}{u_{j-1r}(t)}}{(\lambda_s-a_{j-1})-b_{j-2}\dfrac{u_{j-2s}(t)}{u_{j-1s}(t)}}$$

$$\sim c\exp\{(\lambda_r-\lambda_s)t\}\left(\frac{\lambda_r-\lambda_{j-1}+o(1)}{\lambda_s-\lambda_{j-1}+o(1)}\right),$$

by the inductive hypothesis.

Again from (6), for $s \geqq j$ we have

$$b_{j-1}=[\lambda_s-a_{j-1}-b_{j-2}u_{j-2s}(t)/u_{j-1s}(t)]\frac{u_{j-1s}(t)}{u_{js}(t)},$$

so that

$$b_{j-1}(t)u_{j-1s}(t)/u_{js}(t)=b_{j-1}^2(t)\left[\lambda_s-a_{j-1}(t)-b_{j-2}\frac{u_{j-2s}(t)}{u_{j-1s}(t)}\right]^{-1}$$

converges to 0 as $t\to\infty$ by the induction assumptions. The proof that $u_{jm}^2(t)\leqq c\exp\{-2\mu t\}$, $m<j$, follows from the inductive assumptions and the orthogonality relations $\sum_{i=1}^n u_{ji}u_{ki}=0$, $k<j$. The inequalities

$$|1-u_{jj}^2(t)|\leqq c\exp\{-2\mu t\},$$

$$u_{jm}^2(t)\leqq c\exp\{-2\mu t\}, \qquad m>j,$$

now follow from $u_{jm}^2(t)/u_{jj}^2(t)\sim c\exp\{2(\lambda_m-\lambda_j)t\}$, $m>j$, and the normalization condition $\sum_{i=1}^n u_{ji}^2(t)=1$, as in (e). Finally the inequalities on $|a_j(t)-\lambda_j|$ and $b_j^2(t)$ follow from

$$|a_j(t)-\lambda_j|=\left|\sum_{\substack{k=1\\k\neq j}}^n(\lambda_k-\lambda_j)u_{jk}^2(t)\right|,$$

$$b_j^2(t)=\sum_{k=1}^n[(\lambda_k-a_j(t))u_{jk}(t)-b_{j-1}(t)u_{j-1k}(t)]^2$$

respectively, as in (b) and (c). This completes the induction and the proof of the theorem. □

Remark 2. The proof that $L(t)$ converges to diag $(\lambda_1,\cdots,\lambda_n)$ appears in Moser [2], but without rates of convergence. However, we note that one can combine Moser's argument with a Gronwall inequality to give a simple proof of Theorem 3. The detailed proof we have given, however, applies to the more general flows on tridiagonal matrices of § 3 below.

Remark 3. There are similar estimates as $t\to-\infty$, but now

$$a_k(t)\to\lambda_{n-k+1}, \qquad k=1,2,\cdots,n.$$

Remark 4. Let

$$H=\frac{1}{2}\sum_{k=1}^n y_k^2+\sum_{k=1}^{n-1}\exp(x_k-x_{k+1}),$$

be the Hamiltonian for n particles on a line interacting pairwise with exponential forces (Toda lattice). The observation of Flaschka [1] was that the Hamiltonian equations of motion for these particles

$$\begin{aligned}
&\frac{d}{dt}x_k = y_k, && 1 \leqq k \leqq n,\\
&\frac{d}{dt}y_1 = \frac{d^2}{dt^2}x_1 = -\exp\{x_1 - x_2\}, &&\\
&\frac{d}{dt}y_k = \frac{d^2}{dt^2}x_k = \exp\{x_{k-1}-x_k\} - \exp\{x_k - x_{k+1}\}, && 1 < k < n,\\
&\frac{d}{dt}y_n = \frac{d^2}{dt^2}x_n = \exp\{x_{n-1}-x_n\} &&
\end{aligned}$$

are equivalent to (1) under the change of variables

$$\begin{aligned}
&a_k = -\frac{y_k}{2}, && k = 1, 2, \cdots, n,\\
&b_k = \tfrac{1}{2}\exp\{(x_k - x_{k+1})/2\}, && k = 1, 2, \cdots, n-1.
\end{aligned}$$

From $|a_k(t) - \lambda_k| \leqq c \exp\{-2\mu t\}$ we see that the eigenvalues of $(-L(0))$ are twice the asymptotic velocities of the particles, while the inequalities $b_k^2(t) \leqq c \exp\{-2\mu t\}$ show that the particles responding to the forces $-(\partial/\partial x_k)\sum_{j=1}^{n-1} e^{x_j - x_{j+1}}$ are asymptotically free. Furthermore by (8), $(\lambda_1, \cdots, \lambda_n)$ and $(\log u_{11}, \cdots, \log u_{1n})$ are (essentially) the action-angle variables for H.

Remark 5. Although these decay rates are of theoretical interest, they are only of limited computational interest because of deflation (see § 3).

Theorem 3 provides us in the 3×3 case in particular, with the phase portrait of the flow shown in Fig. 4.

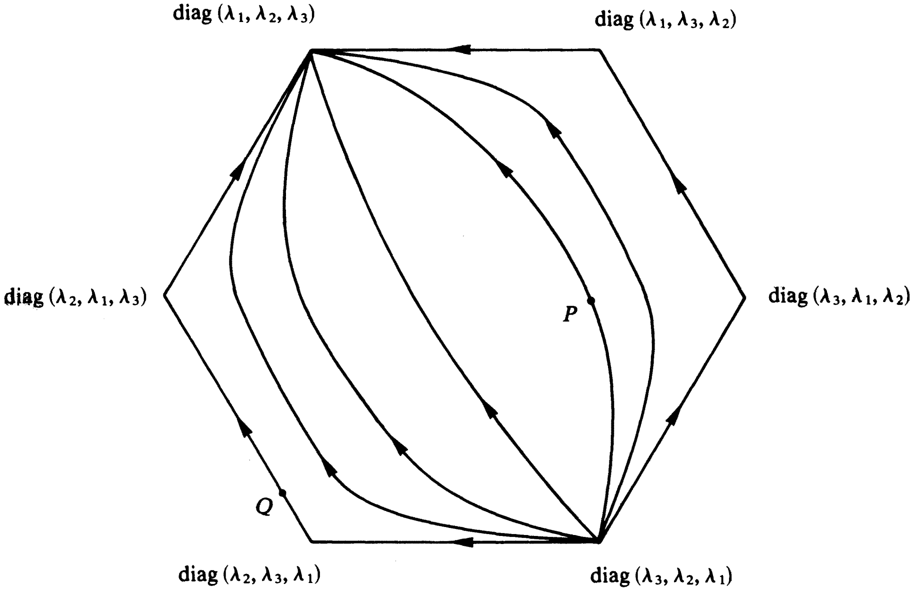
diag $(\lambda_1, \lambda_2, \lambda_3)$
diag $(\lambda_1, \lambda_3, \lambda_2)$
diag $(\lambda_2, \lambda_1, \lambda_3)$
diag $(\lambda_3, \lambda_1, \lambda_2)$
P
Q
diag $(\lambda_2, \lambda_3, \lambda_1)$
diag $(\lambda_3, \lambda_2, \lambda_1)$

FIG. 4

Points P in the interior flow towards diag $(\lambda_1, \lambda_2, \lambda_3)$ in the future and diag $(\lambda_3, \lambda_2, \lambda_1)$ in the past.

Also, from (1), the vertices are (all) the equilibrium points and the edges are invariant under the flow. In particular, the point Q flows towards diag $(\lambda_2, \lambda_1, \lambda_3)$ in the future and diag $(\lambda_2, \lambda_3, \lambda_1)$ in the past.

For general n all the equilibrium points of (1) are of the form diag $(\lambda_{\pi(1)}, \cdots, \lambda_{\pi(n)})$ where π is any permutation of n numbers. From the previous theorem only diag $(\lambda_1, \lambda_2, \cdots, \lambda_n)$ can be stable in the future and only diag $(\lambda_n, \lambda_{n-1}, \cdots, \lambda_1)$, in the past. The following stability theorem is also relevant for the numerical integration of (1).

THEOREM 4. *In the future the equilibrium point* diag $(\lambda_1, \lambda_2, \cdots, \lambda_n)$ *is*:

(i) *stable*[2] *in the class of all tridiagonal matrices with* $b_i > 0$;

(ii) *not asymptotically stable*[2] *in the class of all tridiagonal matrices with* $b_i > 0$;

(iii) *asymptotically stable in the class of all tridiagonal matrices with* $b_i > 0$ *and fixed spectrum* $\lambda_1, \lambda_2, \cdots, \lambda_n$.

The same results hold for diag $(\lambda_n, \lambda_{n-1}, \cdots, \lambda_1)$ *in the past.*

Proof. We only consider diag $(\lambda_1, \lambda_2, \cdots, \lambda_n)$; the other equilibrium point is treated similarly. Suppose $\varepsilon > 0$. To prove (i), we must show that there is a $\delta > 0$ such that $\sup_i |a_i(t) - \lambda_i|$, $\sup_i |b_i(t)| < \varepsilon$ for all $t \geqq 0$ if $\sup_i |a_i(0) - \lambda_i|$, $\sup_i |b_i(0)| < \delta$. Again let $\mu \equiv \min_{1 \leqq i \leqq n-1} |\lambda_i - \lambda_{i+1}|$ and choose δ such that $\delta(1 + 2\delta/\mu) < \mu/4$.

Let $t_0 = \inf \{t \geqq 0\colon |a_k(t) - \lambda_k| > \mu/4$ for some $k\}$, where, $t_0 \equiv \infty$ if the set is empty. We will show that, indeed, $t_0 = \infty$. Suppose $t_0 < \infty$. For $t < t_0$, we have, from (1),

$$\frac{d}{dt} b_k^2 \leqq 2b_k^2\left(\lambda_{k+1} - \lambda_k + \frac{\mu}{2}\right) \leqq 2b_k^2\left(-\frac{\mu}{2}\right),$$

so that

$$b_k^2(t) \leqq b_k^2(0)\, e^{-\mu t} \leqq \delta^2 e^{-\mu t}, \qquad 0 \leqq t \leqq t_0.$$

Substituting in $a_k(t) - a_k(0) = 2 \int_0^t [b_k^2(s) - b_{k-1}^2(s)]\, ds$, we obtain

$$|a_k(t) - a_k(0)| \leqq 2\frac{\delta^2}{\mu}, \qquad 0 \leqq t \leqq t_0,$$

and, finally,

$$|a_k(t) - \lambda_k| \leqq \delta\left(1 + 2\frac{\delta}{\mu}\right) < \frac{\mu}{4}, \qquad 0 \leqq t \leqq t_0.$$

But $|a_k(t_0) - \lambda_k| \geqq \mu/4$, which is a contradiction. Hence $|a_k(t) - \lambda_k| < \mu/4$, as well as $b_k^2(t) \leqq \delta^2 e^{-\mu t}$, for all $t \geqq 0$. Finally, by (1),

$$|a_k(t) - \lambda_k| \leqq \delta + 2\int_0^\infty (b_k^2 + b_{k-1}^2)\, dt \leqq \delta + 4\frac{\delta^2}{\mu}, \qquad t \geqq 0.$$

Choosing δ small enough gives the result.

(ii) If the flow were asymptotically stable, then there would exist $\delta > 0$ with the property that

$$\sup_k |a_k(t) - \lambda_k|, \sup_k |b_k(t)| \to 0 \quad \text{as } t \to \infty$$

[2] "Stable" and "asymptotically stable" are defined in the proof.

whenever

$$\sup_k |a_k(0)-\lambda_k|, \sup_k |b_k(0)|<\delta.$$

But in any neighborhood of diag $(\lambda_1, \cdots, \lambda_n)$ there is a tridiagonal matrix with $b_i>0$ and eigenvalues different from $(\lambda_1, \cdots, \lambda_n)$. Statement (ii) now follows from Theorem 3, and the proof of (iii) is clear. □

By Gershgorin's theorem if $\sup b_i<\varepsilon/2$, then $|a_i-\lambda_i|<\varepsilon$. The algorithm of § 4 stops at a time T_ε for which $\sup b_i(T_\varepsilon)<\varepsilon/2$. We will prove a crude bound which is independent of the initial choice of the matrix L (with fixed spectrum, of course).

THEOREM 5.

$$T_\varepsilon<4\frac{n^2}{\varepsilon^2}\max_i |\lambda_i|. \tag{10}$$

Proof. Let $A_\varepsilon=\{t: b_k(t)\geqq\varepsilon \text{ for some } k\}=\bigcup_{k=1}^{n-1}\{t: b_k(t)\geqq\varepsilon\}$. Let t_k be the Lebesgue measure of $\{t: b_k(t)\geqq\varepsilon\}$. Then, from (1),

$$(a_1+\cdots+a_k)(t)-(a_1+\cdots+a_k)(0)=2\int_0^t b_k^2(s)\,ds\geqq 2\varepsilon^2 t_k,$$

for sufficiently large t (note that $t_k<\infty$ by Theorem 3). But, by (5), $|(a_1+\cdots+a_k)(t)|\leqq k\max|\lambda_i|$. Thus $t_k\leqq(k/\varepsilon^2)\max|\lambda_i|\leqq(n/\varepsilon^2)\max|\lambda_i|$. It follows that the measure Σ of A_ε is $\leqq(n^2/\varepsilon^2)\max|\lambda_i|$. But $T_\varepsilon\leqq\Sigma$, and the theorem follows by replacing ε by $\varepsilon/2$. □

A bound of a different nature, which is helpful in understanding some of the qualitative features of the examples in § 4, is

THEOREM 6. *Suppose $u_{1k}^2(0)>\delta^2>0$, $k=1, \cdots, n$. Then, for given $0<\varepsilon<\min(\mu/4, 1)$,*

$$T_\varepsilon\leqq\frac{1}{2\mu}\log\left[\frac{n^2 2^{3n}(\max(1, 2\eta))^{2n-3}}{\varepsilon^2\delta^2(\min(1,\mu))^{2n-4}}\right], \tag{11}$$

where $\eta=\max_i(|\lambda_i|)$ and, again, $\mu=\min_i(\lambda_i-\lambda_{i+1})$.

The proof, which is a straightforward but tedious application of the estimates of Theorem 3, is omitted.

3. Flow for *QR* and other algorithms. In this section we begin by showing that the basic unshifted QR algorithm is an evaluation at integer times of a flow on the space of tridiagonal matrices with fixed spectrum. The remainder of the section describes other isospectral flows of computational interest.

Let L be an $n\times n$ tridiagonal matrix. For convenience we will assume throughout this section that all the eigenvalues λ_i of L are >0. Francis' basic QR algorithm to calculate the eigenvalues of L is as follows. Factorize

$$L_0\equiv L=Q_0R_0$$

where Q_0 is orthogonal and R_0 is upper triangular (the factorization is essentially unique (see e.g. [8])). One then defines by induction a sequence of matrices L_k, Q_k, R_k by

$$L_k\equiv R_{k-1}Q_{k-1}, \qquad k\geqq 1,$$

where L_{k-1} has the factorization

$$L_{k-1}=Q_{k-1}R_{k-1},$$

as above.

As $L_k = Q_{k-1}^T L_{k-1} Q_{k-1}$, the algorithm is isospectral and as Francis [9] showed, L_k converges to diag $(\lambda_1, \cdots, \lambda_n)$ as $k \to \infty$. Moreover, L_k is tridiagonal for all $k > 0$. (To see this, note that the $Q_n R_n$ decomposition is precisely the Gram–Schmidt process applied to the columns of L_k. Thus if L_k is tridiagonal, $(Q_k)_{j+k',j} = 0$ for $k' \geqq 2$. Finally $L_{k+1} = L_{k+1}^T = (R_k Q_k)^T = Q_k^T R_k^T$, from which the tridiagonality is immediate.) One also checks that $(L_k)_{jj+1} > 0$ if $(L_0)_{jj+1} > 0$ for $j = 1, 2, \cdots, n-1$.

Let U_k denote the orthogonal matrix of eigenvectors for L_k,

$$L_k = U_k \Lambda U_k^T,$$

$k \geqq 0$. Set $u_k = U_k^T e_1$, $k \geqq 0$, where $e_1 = (1, 0, \cdots, 0)^T$.

LEMMA 2.

$$u_k = \frac{\Lambda^k u_0}{\|\Lambda^k u_0\|}.$$

Proof. By induction it is enough to show that $u_1 = \Lambda u_0 / \|\Lambda u_0\|$. But

$$u_1 = U_1^T e_1 = U_0^T Q_0 e_1 = U_0^T \left(\frac{L_0 e_1}{\|L_0 e_1\|} \right) = \frac{U_0^T L_0 e_1}{\|U_0^T L_0 e_1\|} = \frac{\Lambda U_0^T e_1}{\|\Lambda U_0^T e_1\|} = \frac{\Lambda u_0}{\|\Lambda u_0\|},$$

and we are done. □

Let A be any real, constant, symmetric $n \times n$ matrix and consider the isospectral flow induced on tridiagonal matrices (recall Theorem 2) by

$$\frac{d\Lambda}{dt} = 0, \qquad \frac{du}{dt} = Au - (Au, u)u, \tag{12}$$

where $\Lambda(t) = \Lambda$ are the eigenvalues of L and $u = (u_{11}, \cdots, u_{1n}) > 0$ are again the first components of its normalized eigenvectors.

When $A = \Lambda$, (12) reduces to (9), the equations of motion for the Toda flow. As in Lemma 1, direct verification shows that

$$u(t) = \frac{e^{At} u(0)}{\|e^{At} u(0)\|}$$

is the (unique) solution of (12) with initial data $u(0)$.

Set $A = \log \Lambda$. Then $u(t) = \Lambda^t u(0) / \|\Lambda^t u(0)\|$ and in particular, $u(n) = \Lambda^n u(0) / \|\Lambda^n u(0)\|$. By Lemma 2 and the inverse algorithm (Theorem 2), we have proved

THEOREM 7 (*QR flow*).

$$L_n = L(n),$$

where L_n is the nth iterate of QR with $L_0 = L$, and $L(n)$ is the evaluation at $t = n$ of the flow (12) *with $A = \log \Lambda$ and $L(0) = L$.*

Upon replacing λ_k with $\log \lambda_k$, all the propositions of § 2 remain valid. In particular, we obtain a new proof that L_k converges exponentially to diag $(\lambda_1, \cdots, \lambda_n)$ (in that order!),

$$|a_j(t) - \lambda_j| \leqq c\, e^{-2\mu t}, \qquad b_j^2(t) \leqq c\, e^{-2\mu t},$$

with

$$\begin{aligned} \mu = \mu_{QR} &= \min_{1 \leqq j \leqq n-1} (\log \lambda_j - \log \lambda_{j+1}) \\ &= \min_{1 \leqq j \leqq n-1} \log (\lambda_j / \lambda_{j+1}) = \log \min_{1 \leqq j \leqq n-1} (\lambda_j / \lambda_{j+1}). \end{aligned}$$

Moreover, the qualitative phase space phenomena (e.g., absorption, bending) of § 5 below carry over to the QR flow.

The quantity μ_{QR} is essential in understanding shifting and deflation, as we will see below.

We consider now more general flows on tridiagonal matrices. Fix L and let $G(\lambda)$ be a real valued, one-to-one function defined on the spectrum of L. Set $B = (G(L))_+ - (G(L))_-$ where $(G(L))_\pm$ denote the (strictly) upper/lower triangular parts of $G(L)$ respectively. In particular, B is antisymmetric. If $G(\lambda) = \lambda^k$, $k > 0$, a straightforward calculation shows that $B_k L - LB_k$ is tridiagonal,[3] where $B_k = (L^k)_+ - (L^k)_-$, $k > 0$. If we take linear combinations, the same is then true for general G. It follows now, as in Theorem 1, that

$$\frac{dL}{dt} = BL - LB, \qquad B = (G(L)_+ - (G(L))_-,$$

is an isospectral flow $L(0) \to L(t)$ on the space of tridiagonal matrices. As in the proof of Lemma 1,

$$\frac{d}{dt} u_{1j}(t) = \sum_{i=1}^{n} B_{1i} u_{ij}(t) = \sum_{i=2}^{n} (G(L))_{1i} u_{ij}(t).$$

But

$$\sum_{i=1}^{n} (G(L))_{1i} u_{ij}(t) = G(\lambda_j) u_{1j}(t).$$

Thus,

$$\begin{aligned} \frac{d}{dt} u_{1j}(t) &= (G(\lambda_j) - (G(L))_{11}) u_{1j}(t) \\ &= \left(G(\lambda_j) - \sum_{i=1}^{n} G(\lambda_i) u_{1i}^2(t) \right) u_{1j}(t), \end{aligned}$$

or

$$\frac{du}{dt} = G(\Lambda)u - (G(\Lambda)u, u)u,$$

and again all the proofs of convergence and rates of convergence of § 2 go through for the general $G(\lambda)$.

Setting $G(\lambda) = \lambda$, we obtain the Toda flow (2) (or $A \equiv \Lambda$ in (12)); setting $G(\lambda) = \log \lambda$, we obtain the QR flow ($A \equiv \log \Lambda$ in (12)). For $G(\lambda) = s \log \lambda$, we obtain the QR power method [8], but we do not give details, the proofs being very similar to the QR case $s = 1$. We mention, however, that the $s = \frac{1}{2}$ flow corresponds to the Cholesky LR algorithm. In other words, if we evaluate the QR flow $A = \log \Lambda$ at integer times, we obtain the iterates of the QR algorithm; on the other hand, if we evaluate the same flow at half-integer times, we obtain the iterates of the Cholesky LR algorithm. This result is to be anticipated, of course, from the well known relationship between QR and LR for positive definite matrices (see [8, p. 545]).

[3] This is a particular case of the following general fact. Let L be tridiagonal and symmetric (or more generally band symmetric). If G is symmetric, commutes with L and $B = G_+ - G_-$, then $BL - LB$ is tridiagonal (respectively band symmetric of same bandwidth as L).

Equation (12) with $A = \log \Lambda$ expresses the QR flow in λ, u variables. As an example, we display B for QR explicitly in the 2×2 case:

$$B = \begin{pmatrix} 0 & b_1 \int_{-\infty}^{0} \frac{d\lambda}{\lambda^2 - (a_1+a_2)\lambda + (a_1a_2 - b_1^2)} \\ -b_1 \int_{-\infty}^{0} \frac{d\lambda}{\lambda^2 - (a_1+a_2)\lambda + (a_1a_2 - b_1^2)} & 0 \end{pmatrix}.$$

The $n \times n$ case is similar and involves increasingly complicated expressions for B. As opposed to the Toda flow, the QR flow is "nonlocal" in k in the sense that da_k/dt and db_k/dt now depend on *all* the variables $a_j(t)$, $b_j(t)$, $1 \leqq j \leqq n$, at time t. Clearly it could be extremely difficult to integrate this flow numerically: it is indeed remarkable that the QR algorithm gives such a simple solution to this complicated flow at integer times.

As mentioned in the introduction, we have the following result: $e^{L(k;G)}$ is equal to the kth iterate of QR starting from L, where $L(k; G)$ is the evaluation at time $t = k$ of the flow corresponding to G, with $L(0, G) = L$. The proof of this result, which we present in a forthcoming paper, depends on the following theorem[4] which is of independent interest: Let $L(t, G)$ be as above and let $e^{tG(L)} = Q(t)R(t)$ be the (essentially unique) QR factorization of $e^{tG(L)}$. Then

$$L(t, G) = Q(t)^T L Q(t).$$

We end this section with a brief description of the concepts of *shifting* and *deflation*. The basic QR algorithm as described above is slow and never used. Indeed, consider the case where the convergence exponent $\mu_{QR} = \log \min_{1\leqq j \leqq n-1} (\lambda_j/\lambda_{j+1})$, for a tridiagonal matrix L with entries $\{a_i, b_j\}$, is approximately 1. Thus $\lambda_1/\lambda_n \sim e^{n-1}$ and if $b_i \sim 1$, then $a_n/a_1 \sim e^{n-1}$: in other words, the matrices for which the basic QR algorithm can work are unlikely to occur. Practical QR algorithms rely on the following ingenious observation (see e.g. [8]); although μ_{QR} governs the rate of convergence of the entire matrix L to diagonal form, it is clear from the proof of Theorem 3 (the situation with b_1 and b_{n-1} is symmetrical) that $b_{n-1} \to 0$ with a convergence exponent $\log(\lambda_{n-1}/\lambda_n)$. Suppose we *shift* the matrix $L \to L - \sigma$ in such a way that $\lambda_n - \sigma$ is small (for example $\sigma \equiv a_n$). Then $\log[(\lambda_{n-1} - \sigma)/(\lambda_n - \sigma)]$ is large and $b_{n-1} \to 0$ with accelerated convergence (if $\sigma \equiv a_n$, the convergence is in fact cubic; see e.g. [8]). Once b_{n-1} is very small it is possible to *deflate* the matrix, i.e., set $b_{n-1} \equiv 0$, and a_n gives an eigenvalue with great accuracy. The calculation proceeds with matrices of decreasing size.

The situation with Toda, however, is very different. If $\mu_{\text{Toda}} = \min_{1 \leqq k \leqq n-1} (\lambda_k - \lambda_{k+1}) \sim 1$, and $b_i \sim 1$, then $a_n - a_1$ is of order n, which is perfectly reasonable. The concept of shift is connected with the functional form of the convergence exponent for QR, and has no relevance to Toda as $\lambda_{k-1} - \lambda_k$ is invariant under shift. The appropriate notion is scaling, $L \to \sigma L$, so that $\mu_{\text{Toda}}(\sigma L) = \sigma \mu_{\text{Toda}}(L)$; in other words, the question of scaling is precisely the problem of choosing an effective time step strategy. On the other hand, the flows defined by $G(\lambda) = \lambda^m$, with convergence exponent $\min_{1 \leqq k \leqq n-1} (\lambda_k^m - \lambda_{k+1}^m)$, $m > 1$, benefit greatly by shifting (and

[4] Bill Symes has kindly pointed out to us that this theorem occurs in [11, p. 358], where it is proved using very different methods.

scaling). The benefit, however, should be balanced against the increasing algebraic complexity of the equations; for example, for $G(\lambda)=\lambda^2$ the equations are

$$\frac{da_k}{dt}=2[b_k^2(a_{k+1}+a_k)-b_{k-1}^2(a_k+a_{k-1})],$$

$$\frac{db_k}{dt}=b_k[a_{k+1}^2-a_k^2+b_{k+1}^2-b_{k-1}^2].$$

Finally we note that, unlike shifting, deflation is intrinsic to the eigenvalue problem and Toda and the higher order algorithms can only improve under deflation.

4. Numerical examples. Our objective in the calculations which follow is only to present evidence that the Toda method is competitive with implemented *QR* algorithms and not to present a detailed, exhaustive comparison of the techniques. The implemented version of *QR* which we used was RATQR in EISPACK; other routines, for example TQL1 in EISPACK, are faster but seldom by more than a factor of 2.

All computations were done on a CDC 6600. It is clear, however, that the Toda method is eminently suited to vector processing, and this implies a gain of a factor n in machine time. Rather than integrating equations (2) directly, we have found it useful to work with the equivalent system

$$\begin{aligned}
&\frac{d}{dt}S_k=C_k, && k=1,\cdots,n-1,\\
(1')\qquad &\frac{d}{dt}S_n=0,\\
&\frac{d}{dt}C_k=2C_k(S_{k+1}-2S_k+S_{k-1}), && k=1,\cdots,n-1,
\end{aligned}$$

where

$$\begin{aligned}
&S_0=0,\\
&S_k=\sum_{j=1}^{k}a_j, && k=1,\cdots,n,\\
&C_k=2b_k^2, && k=1,\cdots,n-1.
\end{aligned}$$

Equations (1′) were integrated with a fourth-order Runge–Kutta scheme. The initial time step was chosen equal to $\frac{1}{50}[\max_i\,(|a_i(0)|,|b_i(0)|)]^{-1}$ and was doubled approximately every $n/2$ iterations. No doubt this naive integration scheme could be considerably improved. Also the algorithm was implemented without deflation.

As mentioned in § 2, the algorithm stops when all the b_i are less than $\varepsilon/2$. By second order perturbation theory (see e.g. [7, Chap. XII]) λ_k-a_k is of order $(\sum_{i=1}^{n-1}b_i^2)/\mu$. In particular the method is quadratically convergent and when $\mu\sim1$, the eigenvalues are computed to an accuracy ε^2. If μ is small however, all we can say is that λ_k-a_k is computed to an accuracy ε, by Gershgorin.

Example 1. Here we calculate the eigenvalues of tridiagonal symmetric matrices with

$$\begin{aligned}
&L_{ii}=2, && 1\leqq i<n,\\
&L_{nn}=1,\\
&L_{ii+1}=1, && 1\leqq i\leqq n-1.
\end{aligned}$$

The eigenvalues are $4\cos^2(j\pi/2n+1)$, $j=1, \cdots, n$, $\mu=\min(\lambda_j-\lambda_{j+1})\sim 8\pi^2/(2n+1)^2$ for large n.

Table 1 compares the machine time in seconds for the *QR* method and the ODE method. The termination criterion for the ODE method was $\sup_i b_i<10^{-5}/2$ and for *QR* the variable EPS in RATQR was set equal to 10^{-5}.

TABLE 1

Order of matrix	*QR*	Toda
100	1.79	8.36
300	17.57	44.29
500	51.02	125.51

The numbers in Table 1 represent the worst case in our experiments with varying the time step. For example, if we use an initial time step $(\frac{1}{10})(\max_i(|a_i|, |b_i|))^{-1}$, and double it every $n/3$ iterations, the ODE method determines the eigenvalues to an accuracy $\leqq 10^{-5}$ in 0.70 of the time required for *QR*.

The *phase space* time for the matrices in Table 1 agreed within a factor of 5 with the formula

$$T'\equiv\frac{1}{2\mu}\log\frac{1}{\varepsilon^2\delta^2}.$$

The factors of n inside the logarithm of (11) give a gross overestimate for the matrices at hand.

Example 2. Let L be given by $L_{kk}=n-k+1$, $1\leqq k\leqq n$, $L_{kk+1}=1$, $1\leqq k\leqq n-1$, for which

$$\mu_{\text{Toda}}=\min_i(\lambda_i-\lambda_{i+1})\sim 1.$$

The entries in Table 2 give the machine time in seconds for *QR* and the Toda method respectively.

TABLE 2
EPS $=10^{-4}$, $\sup_i b_i<10^{-4}$.

Order of matrix	*QR*	Toda
100	1.903	7.17
300	19.99	30.73
500	57.50	77.87

As in Example 1, we remark that if we use a different time step, starting at $\frac{1}{10}[\max_i|a_i|, |b_i|]^{-1}$ and doubling every $n/3$ iterations, we achieve an accuracy of 10^{-4} in the 300×300 case in a third of the time of *QR*.

Again, the *phase space* time for the matrices in Table 2 agreed within a factor of 5 with the formula

$$T'_\varepsilon=\frac{1}{2\mu}\log\frac{1}{\varepsilon^2\delta^2}.$$

Example 3 (an experiment). In this example we considered matrices of size 40×40 with fixed spectrum $\lambda_k=k$, $k=1,\cdots,40$. We then obtained 100 vectors $(u_{11},\cdots,u_{140})$ from a random generator (RANF (N)) and used the inverse algorithm (Theorem 2) to calculate the corresponding matrices $L(a,b)$. Finally we used (1) to calculate back the eigenvalues. The eigenvalues were computed to an accuracy of 10^{-6}, with initial time step 10^{-6}, and the step was doubled every 75 iterations.

The results are:

$$\text{average machine time} = 2.44 \text{ sec.},$$

$$\text{standard deviation} = 0.03 \text{ sec.}$$

From this experiment we learn that matrices chosen at random are far from the boundary of phase space (see § 5).

5. Isospectral analysis: Absorption and bending. In this section we analyze a number of 3- and 4-dimensional examples in order to illustrate some interesting phase space phenomena for the Toda flow. The other flows of § 3, of course, have similar properties.

We begin with the 19 distinct 3×3 matrices on the isospectral manifold

$$\lambda_1=8,\quad \lambda_2=4,\quad \lambda_3=2$$

constructed according to the inverse algorithm of Theorem 2 from vectors (u_{11}, u_{12}, u_{13}) of the form $(v_1, v_2, v_3)/(v_1^2+v_2^2+v_3^2)^{1/2}$, where $v_i\in\{1, 10^{-5}, 10^{-10}\}$. Then the eigenvalues are calculated back again using (1) as in Example 3, § 4, with an accuracy of 10^{-6}. The extreme values 10^{-5}, 10^{-10} were chosen for v_i in order to place the initial matrices close to edges and vertices where, as we will see, the interesting phenomena occur.

Schematically, the matrices are positioned with x's in phase space shown in Fig. 5.

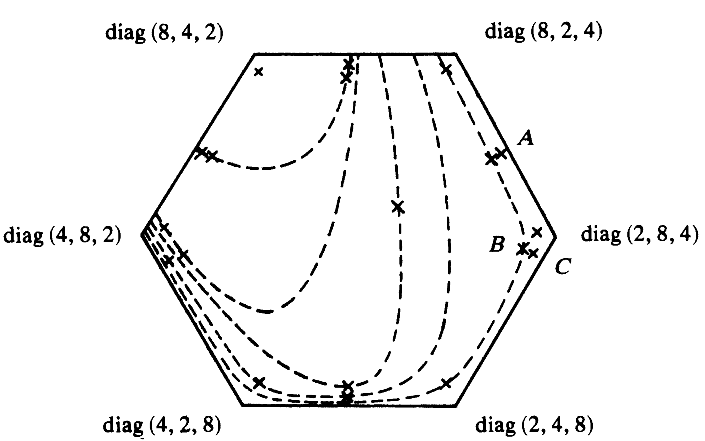

FIG. 5

In Fig. 5, the dotted lines are isochronal, i.e., all matrices on the same dotted line take the same time t to achieve the accuracy 10^{-6}.

Consider matrix A with trajectory given in Table 3.

TABLE 3

Time	a_1	a_2	a_3	b_1	b_2
0	5.0000E+00	5.0000E+00	4.0000E+00	3.0000E+00	1.8856E−10
1.0000E+00	8.0000E+00	2.0000E+00	4.0000E+00	1.4872E−02	9.8021E−10
2.0000E+00	8.0000E+00	2.0000E+00	4.0000E+00	3.6865E−05	7.2790E−09
2.6050E+00	8.0000E+00	2.0000E+00	4.0000E+00	9.7753E−07	2.4413E−08

This matrix takes a dramatically shorter time to calculate the λ's than matrices "close" to it. What happens is that the trajectory through A approaches close enough to diag (8, 2, 4) to terminate the algorithm. This is the phenomenon of *absorption.*

Consider the matrices B and C with trajectories given in Tables 4 and 5. Schematically, the orbit of C is as shown in Fig. 6.

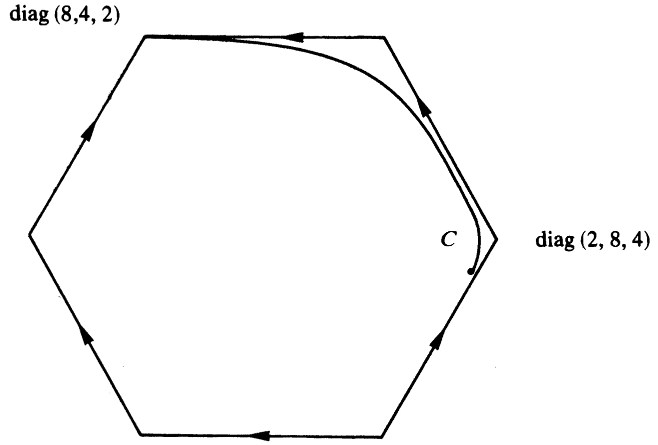

FIG. 6

TABLE 4
Matrix B.

Time	a_1	a_2	a_3	b_1	b_2
0	2.0000E+00	7.6000E+00	4.4000E+00	6.3246E−05	1.2000E+00
1.0000E+00	2.0001E+00	7.9998E+00	4.0001E+00	2.4206E−02	2.4420E−02
2.0000E+00	6.3557E+00	3.6443E+00	4.0000E+00	2.6762E+00	8.5441E−04
3.0000E+00	8.0000E+00	2.0000E+00	4.0000E+00	9.1380E−03	5.3790E−03
4.0000E+00	8.0000E+00	2.0000E+00	3.9992E+00	2.2655E−05	3.9730E−02
5.0000E+00	8.0000E+00	2.0442E+00	3.9578E+00	5.6748E−00	2.8719E−01
6.0000E+00	8.0000E+00	3.0814E+00	2.9186E+00	2.0536E−10	9.9668E−01
7.0000E+00	8.0000E+00	3.9694E+00	2.0306E+00	2.7872E−12	2.4564E−01
8.0000E+00	8.0000E+00	3.9994E+00	2.0006E+00	5.0664E−14	3.3751E−02
9.0000E+00	8.0000E+00	4.0000E+00	2.0000E+00	9.2781E−16	4.5690E−03
1.0000E+01	8.0000E+00	4.0000E+00	2.0000E+00	1.6993E−17	6.1885E−04
1.1000E+01	8.0000E+00	4.0000E+00	2.0000E+00	3.1125E−19	8.3684E−05
1.2000E+01	8.0000E+00	4.0000E+00	2.0000E+00	5.7007E−21	1.1325E−05
1.3000E+00	8.0000E+00	4.0000E+00	2.0000E+00	1.0441E−22	1.5327E−06
1.3215E+01	8.0000E+00	4.0000E+00	2.0000E+00	4.4183E−23	9.9705E−07

TABLE 5
Matrix C.

Time	a_1	a_2	a_3	b_1	b_2
0	2.0000E+00	7.6000E+00	4.4000E+00	4.3346E−10	1.2000E+00
1.0000E+00	2.0000E+00	7.9999E+00	4.0001E+00	2.4206E−07	2.4420E−02
2.0000E+00	2.0000E+00	3.0000E+00	4.0000E+00	5.7653E−05	4.4728E−04
3.0000E+00	2.0003E+00	7.9997E+00	4.0000E+00	3.9394E−02	8.1925E−06
4.0000E+00	7.2516E+00	2.7484E+00	4.0000E+00	1.9825E+00	4.2484E−07
5.0000E+00	8.0000E+00	2.0000E+00	4.0000E+00	5.6146E−03	2.9369E−00
6.0000E+00	8.0000E+00	2.0000E+00	4.0000E+00	1.3917E−05	2.1701E−05
7.0000E+00	8.0000E+00	2.0000E+00	4.0000E+00	3.4497E−00	1.6035E−03
8.0000E+00	8.0000E+00	2.0000E+00	4.0000E+00	8.5510E−11	1.1848E−03
9.0000E+00	8.0000E+00	2.0000E+00	4.0000E+00	2.1196E−13	8.7515E−01
1.0000E+01	8.0000E+00	2.0021E+00	3.9979E+00	5.2567E−16	6.4621E−02
1.1000E+01	8.0000E+00	2.1081E+00	3.8919E+00	1.3390E−18	4.5216E−01
1.2000E+01	8.0000E+00	3.5144E+00	2.4856E+00	6.5512E−21	8.5756E−01
1.3000E+01	8.0000E+00	3.9883E+00	2.0117E+00	1.0172E−22	1.5238E−01
1.4000E+01	8.0000E+00	3.9998E+00	2.0002E+00	1.9125E−24	2.0741E−02
1.5000E+01	8.0000E+00	4.0000E+00	2.0000E+00	3.5026E−26	2.8073E−03
1.6000E+01	8.0000E+00	4.0000E+00	2.0000E+00	6.4153E−28	3.7993E−04
1.7000E+01	8.0000E+00	4.0000E+00	2.0000E+00	1.1250E−29	5.1417E−05
1.8000E+01	8.0000E+00	4.0000E+00	2.0000E+00	2.1521E−31	6.9586E−06
1.8970E+01	8.0000E+00	4.0000E+00	2.0000E+00	4.4442E−33	9.9998E−07

Matrix B requires 13 units of phase space time while C requires 19 units. The discrepancy in the times occurs because the trajectory through C is close enough to be slowed down by the equilibrium point diag $(2, 8, 4)$, but not close enough to be absorbed. This is the phenomenon of *bending*.

These phenomena occur with increasing combinatorial complexity in all dimensions n. For example, when $n = 4$, the following (schematic) trajectories can occur (see Fig. 7):

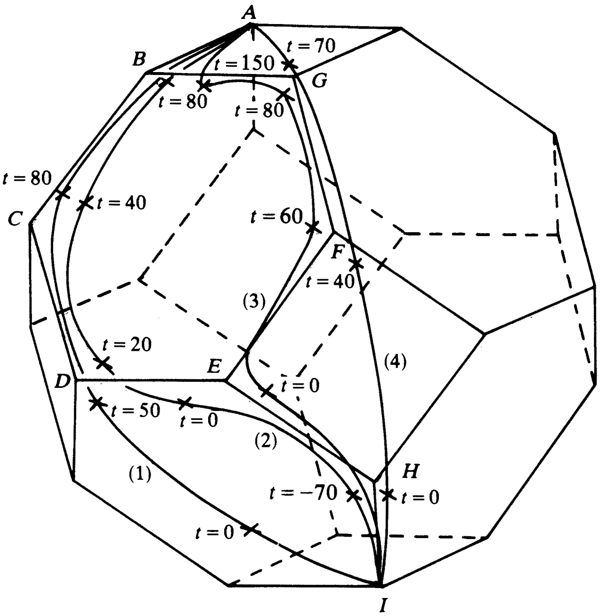

FIG. 7

$$A \equiv \mathrm{diag}\,(10.4, 10.3, 10.2, 10.1), \qquad B \equiv \mathrm{diag}\,(10.4, 10.3, 10.1, 10.2),$$
$$C \equiv \mathrm{diag}\,(10.4, 10.1, 10.3, 10.2), \qquad D \equiv \mathrm{diag}\,(10.1, 10.4, 10.3, 10.2),$$
$$E \equiv \mathrm{diag}\,(10.1, 10.3, 10.4, 10.2), \qquad F \equiv \mathrm{diag}\,(10.3, 10.1, 10.4, 10.2),$$
$$G \equiv \mathrm{diag}\,(10.3, 10.4, 10.1, 10.2), \qquad H \equiv \mathrm{diag}\,(10.1, 10.3, 10.2, 10.4),$$
$$I \equiv \mathrm{diag}\,(10.1, 10.2, 10.3, 10.4).$$

At time $t=0$, the orbits 1), 2), 3) and 4) pass through the isospectral matrices obtained from the inverse algorithm (Theorem 2) using

$$\lambda_1 = 10.4, \quad \lambda_2 = 10.3, \quad \lambda_3 = 10.2, \quad \lambda_4 = 10.1$$

and

$$(u_{11}, u_{12}, u_{13}, u_{14}) = (10^{-4}, 10^{-4}, 10^{-4}, 1)/(3\times(10^{-4})^2 + 1^2)^{1/2},$$
$$(u_{11}, u_{12}, u_{13}, u_{14}) = (10^{-8}, 10^{-8}, 10^{-4}, 1)/(2\times(10^{-8})^2 + (10^{-4})^2 + 1^2)^{1/2},$$
$$(u_{11}, u_{12}, u_{13}, u_{14}) = (10^{-4}, 10^{-8}, 10^{-4}, 1)/(2\times(10^{-4})^2 + (10^{-8})^2 + 1^2)^{1/2},$$
$$(u_{11}, u_{12}, u_{13}, u_{14}) = (10^{-4}, 10^{-8}, 10^{-8}, 1)/(2\times(10^{-8})^2 + (10^{-4})^2 + 1^2)^{1/2},$$

respectively.

These trajectories show in particular how orbits can split apart and recombine when $n>3$.

6. Concluding remarks. 1. The flows

$$\frac{dL}{dt} = [B, L], \qquad B = (G(L))_+ - (G(L))_-,$$

can be applied directly to symmetric band matrices of any bandwidth and again $L(t)$ converges to a diagonal matrix (see Nanda [6]), so it is not necessary to first tridiagonalize the matrix of interest. But we have not yet experimented with these flows at the practical level.

2. A block version of the Toda flow is also of interest. Here

$$L = \begin{pmatrix} A_1 & B_1 & & & \\ B_1^T & A_2 & B_2 & & 0 \\ & & & \ddots & \\ & 0 & & & B_{n-1} \\ & & & B_{n-1}^T & A_n \end{pmatrix}$$

where the matrices A_i are real and symmetric (and not necessarily of the same order). The flow is

$$\frac{dL}{dt} = [B, L],$$

where

$$B = \begin{pmatrix} 0 & B_1 & & & \\ -B_1^T & 0 & B_2 & & 0 \\ & & & \ddots & \\ & 0 & & & B_{n-1} \\ & & & -B_{n-1}^T & 0 \end{pmatrix}.$$

Again B_i converges to 0 and L converges to a block diagonal matrix.

The flow above clearly raises a number of interesting possibilities. For example, consider the partition

$$L=\begin{pmatrix} a_1 & B_1 \\ B_1^T & A_2 \end{pmatrix}$$

where a_1 is a scalar. Under the flow, L converges to a matrix of the form

$$\begin{pmatrix} \lambda & 0 \\ 0 & A_2' \end{pmatrix}.$$

Thus the flow deflates the matrix.

In a forthcoming paper we show how to apply the Toda and related flows to nonsymmetric matrices (see also [6]) and infinite matrices.

Acknowledgment. We are pleased to thank Eli Isaacson for many valuable conversations, Eugene Isaacson and Olof Widlund for encouragement and many helpful suggestions, Jürgen Moser for bringing the paper of Symes [3] to our attention and Don Marchesin for raising the question of block matrices. Finally we would like to thank the referees for helping us to make the work presentable to the matrix eigenvalue community.

REFERENCES

[1] H. FLASCHKA, *The Toda lattice*, I, Phys. Rev. B, 9 (1974), pp. 1924–1925.

[2] J. MOSER, *Finitely many mass points on the line under the influence of an exponential potential—An integrable system*, in Dynamic Systems Theory and Applications, J. Moser, ed., Springer-Verlag, New York, Berlin, Heidelberg, 1975, pp. 467–497.

[3] W. W. SYMES, *The QR Algorithm and Scattering for the Finite Nonperiodic Toda Lattice*, Michigan State Univ., preprint, 1980.

[4] P. DEIFT, F. LUND AND E. TRUBOWITZ, *Nonlinear wave equations and constrained harmonic motion*, Comm. Math. Phys., 74 (1980), pp. 141–188.

[5] P. VAN MOERBEKE, *The spectrum of Jacobi matrices*, Invent. Math., 37 (1976), pp. 45–81.

[6] T. NANDA, Ph.D. Thesis, New York Univ., New York, 1982.

[7] M. REED AND B. SIMON, *Methods of Mathematical Physics*, Vol. 4, Academic Press, New York, 1978.

[8] J. WILKINSON, *The Algebraic Eigenvalue Problem*, Oxford Univ. Press, London, 1965.

[9] J. FRANCIS, *The QR transformation, a unitary analogue to the LR transformation*, I, Comput. J., 4 (1961), pp. 265–271.

[10] B. N. PARLETT, *The Symmetric Eigenvalue Problem*, Prentice-Hall, Englewood Cliffs, NJ, 1980.

[11] W. W. SYMES, *Hamiltonian group actions and integrable systems*, Physica, 1D (1980), pp. 339–374.