SIAM J. MATH. ANAL.
Vol. 15, No. 1, January 1984

0036-1410/84/1501-0004 $01.25/0

ON THE GLOBAL CONVERGENCE OF THE TODA LATTICE FOR REAL NORMAL MATRICES AND ITS APPLICATIONS TO THE EIGENVALUE PROBLEM*

MOODY T. CHU†

Abstract. The asymptotic behavior of the Toda lattice, when acting on real normal matrices, is studied. It is shown that the solution flow eventually converges to a diagonal block form where for a real eigenvalue the associated block is of size 1×1 with that eigenvalue as its element and for complex-conjugate pairs of eigenvalues the associated block is of size 2×2 with the real part as its diagonal elements and the (negative) imaginary part as its off-diagonal elements. This result generalizes the well-known asymptotic behavior of Jacobi matrices and is consistent with that from the QR-algorithm.

1. Introduction. Recently the dynamic flow of a special system of differential equations, known as the Toda lattice, has been found to be closely related to the important QR-algorithm [1], [2], [4], [7]. Roughly speaking, the QR-algorithm can be shown to be the time-1 mapping of the solution to the Toda lattice. Specifically, if we consider the following dynamic system for matrices in $\mathbb{R}^{n\times n}$:

$$\dot{X}=[X,\Pi_0 X]=X\cdot\Pi_0 X-\Pi_0 X\cdot X \tag{1.1}$$

where $\Pi_0 X=X^- -X^{-T}$ and X^- is the strictly lower triangular part of X, then the following properties concerning the solution flow $X(t)$ with initial data X_0 at $t=0$ can be derived from the general results presented in the previous paper [1].

LEMMA 1.1. *The solution $X(t)$ is given by*

$$X(t)=Q^*(t)X_0Q(t), \tag{1.2}$$

where $Q(t)$ solves the initial value problem

$$\dot{Q}=Q\cdot\Pi_0 X,\qquad Q(0)=I. \tag{1.3}$$

Indeed $Q(t)$ is exactly the unitary matrix involved in the QR-decomposition [3], [6] of the matrix e^{tX_0}, namely

$$e^{tX_0}=Q(t)R(t) \tag{1.4}$$

where $R(t)$ is an upper triangular matrix with real nonnegative diagonal elements.

LEMMA 1.2. *For $k=0,\pm 1,\pm 2,\cdots$, suppose the matrix $e^{X(k)}$ has the QR-decomposition*

$$e^{X(k)}=Q^{(k)}R^{(k)}. \tag{1.5}$$

Then

$$e^{X(k+1)}=R^{(k)}Q^{(k)}. \tag{1.6}$$

Observe that, by (1.2), the trajectory $X(t)$ is bounded in $\mathbb{R}^{n\times n}$, so its ω-limit set is nonempty, compact and connected. We are interested in finding this set. A special case, when X_0 is a Jacobi matrix (and hence when X_0 is a real symmetric matrix by a standard tridiagonalization algorithm), has been studied extensively by a number of authors [2], [4], [7]. In fact, based on the continuous dependence of the initial data for

*Received by the editors December 7, 1982.
†Department of Mathematics, North Carolina State University, Raleigh, North Carolina 27650.

the system (1.1) and a well-known theorem [5], [6] in the numerical analysis concerning the convergence of the QR-algorithm, we have the following generalization [1].

THEOREM 1. *If the matrix $X_0 \in \mathbb{R}^{n\times n}$ has real distinct eigenvalues $\{\lambda_1 > \lambda_2 > \cdots > \lambda_n\}$, then the Toda flow $X(t)$ converges to an upper triangular matrix with the eigenvalues appearing on the diagonal in the descending order.*

In this paper we want to study the behavior of this flow when complex-conjugate pairs of eigenvalues occur. As is shown in [1], for an arbitrary (nonnormal) 2×2 matrix, the appearance of such a pair of eigenvalues will result in a periodic (in fact, a circular) portrait in the phase plane and thus $X(t)$ has no convergence at all. It is natural, therefore, to restrict ourselves in the study of the normal matrices first.

We begin in the next section with some preliminary facts. Especially, we point out the differential system which governs the dynamics of the corresponding eigenvectors of the flow $X(t)$. It turns out this system is much easier to handle than the system (1.1) itself. In §3 we discuss how eigenvalues affect eigenvectors and, hence, the entire flow $X(t)$ by the inverse algorithm. Although we only analyze two situations there, they seem to be generic enough to get general conclusions.

2. Preliminary facts. It is obvious, from Lemma 1.1, that normality is preserved along the flow provided that X_0 is a normal matrix. It is also known that there exists a unitary matrix U_0 such that

$$X_0 = U_0^* T U_0, \tag{2.1}$$

where T is a diagonal matrix with eigenvalues as its elements. Without loss of generality we shall assume these elements are arranged in such a way that

$$\mathrm{Re}\lambda_1 \geq \mathrm{Re}\lambda_2 \geq \cdots \geq \mathrm{Re}\lambda_n, \tag{2.2}$$

and that whenever there are complex-conjugate pairs, they are adjacent to each other. By (1.2), it follows that

$$X(t) = U^*(t) T U(t) \tag{2.3}$$

where

$$U(t) = U_0 Q(t). \tag{2.4}$$

Notice that, by (1.3), $U(t)$ satisfies the differential system

$$\dot{U} = U \cdot \Pi_0 X. \tag{2.5}$$

We shall assume X_0 is an upper Hessenberg matrix. Then the following lemma [1] guarantees the preservation of this structure along the entire flow. Recall that this useful property is also enjoyed by the classical QR-algorithm.

LEMMA 2.1. *If X is an upper Hessenberg matrix, so is $\dot{X} = [X, \Pi_0 X]$.*

Let us denote the matrix $U(t)$ in (2.4) by $U(t) = [u_1(t), \cdots, u_n(t)]$ where $u_i(t)$ is the ith column of $U(t)$. Then by (2.3) we have

$$[u_1, \cdots, u_n] \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n} \\ x_{21} & x_{22} & \ddots & \\ & x_{32} & \ddots & \vdots \\ & & \ddots & \\ 0 & & x_{n,n-1} & x_{nn} \end{bmatrix} = T[u_1, \cdots, u_n]. \tag{2.6}$$

So the following equality holds for each $k=1,\cdots,n$.

$$\sum_{i=1}^{k+1} x_{ik}u_i = Tu_k, \tag{2.7}$$

where it is understood that $u_{n+1}=0$. Since all the vectors u_i are mutually orthogonal, we know that for all $1\le i\le n$ and $1\le j\le n$

$$x_{ij}=\langle u_i, Tu_j\rangle, \tag{2.8}$$

where $\langle\cdot,\cdot\rangle$ is the inner product in $\mathbb{C}^n$.

From (2.5), (2.6) and (2.8), it is not hard to see now that

LEMMA 2.2. *For $i=1,\cdots,n$, the vector $u_i(t)$ satisfies the differential system*

$$\dot{u}_i = Tu_i - \sum_{j=1}^{i} \langle u_j, Tu_i\rangle u_j - \langle u_i, Tu_{i-1}\rangle u_{i-1}. \tag{2.9}$$

In particular, the first column $u_1(t)$ of $U(t)$ satisfies the equation

$$\dot{u}_1 = Tu_1 - \langle u_1, Tu_1\rangle u_1. \tag{2.10}$$

Direct substitution also shows that

LEMMA 2.3. *The solution to* (2.9) *is given explicitly by*

$$u_1(t)=\frac{e^{Tt}u_1(0)}{\|e^{Tt}u_1(0)\|_2}. \tag{2.11}$$

We note that the ith component $u_{i1}(t)$ of u_1 is given by

$$u_{il}(t)=\frac{e^{\lambda_i t}u_{i0}}{\left\{\sum_{j=1}^{n}\left|e^{\lambda_j t}u_{j0}\right|^2\right\}^{1/2}} \tag{2.12}$$

where u_{i0} is the complex conjugate of the first component of the ith eigenvector of X_0. The following useful inverse algorithm [5] turns out to be very important.

THEOREM 2.1. *Suppose B is an unreduced upper Hessenberg matrix with positive subdiagonal elements and Q is a unitary matrix, then Q and B are uniquely determined by the first column of Q, provided A is given and $B=Q^*AQ$.*

For our application, observe that the subdiagonal elements of $X(t)$ can never change signs along the positive orbit. If we assume, without loss, that X_0 not only is an upper Hessenberg matrix but also is unreduced to begin with, then from (2.6), (2.10) and the above theorem, we know that $X(t)$ and $U(t)$ are completely determined. The detailed analysis is presented in the next section.

3. Convergence of $X(t)$. First of all we should explain the meaning of convergence used in our context. Strictly speaking, convergence would be taken to mean the convergence of the flow $X(t)$ to some limit matrix. In our context, however, we mean convergence under deflations, i.e. we are concerned about the convergence of a submatrix obtained by deflation, as soon as the subdiagonal element is negligible, to another submatrix. The precise meaning will become clear later and indeed, as will be seen also, these two notions of convergence are essentially the same when the Toda lattice is acting on normal matrices.

For the simplicity of discussion, we shall make one more generic assumption, namely $u_{10} \neq 0$ whenever we need it and that X_0 is nonsingular. We shall also use the notation " $\to$ " to mean "converges to."

LEMMA 3.1. *If the eigenvalues in* (2.2) *are such that*

$$\operatorname{Re}\lambda_1 = \lambda_1 > \operatorname{Re}\lambda_2 \geq \cdots \geq \operatorname{Re}\lambda_n, \tag{3.1}$$

then

$$x_{11}(t) \to \lambda_1, \quad x_{21}(t) \to 0 \quad \textit{and} \quad x_{ik}(t) \to 0 \tag{3.2}$$

for every $2 \leq k \leq n$ *as* $t \to \infty$.

Proof. It is clear from (2.12) that as $t \to \infty$,

$$u_{11} \to \frac{u_{10}}{|u_{10}|} \quad \text{and} \quad u_{i1}(t) \to 0 \tag{3.3}$$

for all $i \geq 2$. Let us adopt the following notation in its intuitive sense:

$$\lim_{t \to \infty} u_i(t) = \hat{u}_i. \tag{3.4}$$

Then we have, from (2.8),

$$x_{11}(t) = \langle u_1, Tu_1 \rangle \to \langle \hat{u}_1, T\hat{u}_1 \rangle = \lambda_1 \tag{3.5}$$

and, from (2.7),

$$|x_{21}(t)| = \|Tu_1 - x_{11}u_1\|_2 \to \|T\hat{u}_1 - \lambda_1\hat{u}_1\|_2 = 0. \tag{3.6}$$

Observe that, by (2.8) and (3.6),

$$x_{21}(t) = \langle u_2, Tu_1 \rangle \to \bar{u}_{12}\lambda_1\hat{u}_{11} \to 0 \tag{3.7}$$

implies

$$u_{12}(t) \to 0 \tag{3.8}$$

where $\bar{}$ means the complex conjugate. Therefore,

$$x_{12}(t) = \langle u_1, Tu_2 \rangle = \langle T^*u_1, u_2 \rangle \to \lambda_1\bar{\hat{u}}_{11}u_{12} \to 0. \tag{3.9}$$

Indeed, for every $k > 2$, it is always true that

$$x_{k1}(t) = \langle u_k, Tu_1 \rangle \equiv 0 \to \bar{u}_{1k}\lambda_1\hat{u}_{11} \tag{3.10}$$

implies

$$u_{1k}(t) \to 0. \tag{3.11}$$

Therefore,

$$x_{1k}(t) = \langle u_1, Tu_k \rangle = \langle T^*u_1, u_k \rangle \to \lambda_1\bar{\hat{u}}_{11}u_{1k} \to 0. \tag{3.12}$$

In other words, if condition (3.1) is satisfied, then as $t \to \infty$

$$X(t) \to \begin{bmatrix} \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & x & x & x & x \\ 0 & x & x & x & x \\ 0 & 0 & x & x & x \\ 0 & 0 & 0 & x & x \end{bmatrix}$$

where "x" represents either a nonzero element or an uncertain position.

Apparently when this convergence phenomenon happens, one is tempted to perform the deflation and to proceed the computation on the submatrix. We would like to point out, however, that those uncertain positions are really not entirely uncertain (they are uncertain simply because we don't care to include the analysis in Lemma 3.1). As a matter of fact, from (2.9), we know that for each $k \geq 2$, the eigenvector u_k is governed by

$$u_k = Tu_k = \sum_{i=1}^{k} \langle u_i, Tu_k \rangle u_i - \langle u_k, Tu_{k-1} \rangle u_{k-1}, \tag{3.13}$$

whereas, from (3.8), (3.9), (3.11) and (3.12), we see that the vector $\tilde{u}_k \in \mathbb{C}^{n-1}$, governed by

$$\dot{\tilde{u}}_k = \tilde{T}\tilde{u}_k - \sum_{i=2}^{k} \langle \tilde{u}_i, \tilde{T}\tilde{u}_k \rangle \tilde{u}_i - \langle \tilde{u}_k, \tilde{T}\tilde{u}_{k-1} \rangle \tilde{u}_{k-1},$$

where $\tilde{T}$ is obtained from T by deleting the first row and column, would describe the behavior of u_k as well when t is large enough. Therefore, those uncertain positions are actually converging according to either Lemma 3.1, with λ_1 being replaced by λ_2, or the next lemma, with λ_1 and λ_2 being replaced by λ_2 and λ_3. It is in this sense that we mean convergence.

LEMMA 3.2. *If the eigenvalues in* (2.2) *are such that*

$$\mathrm{Re}\lambda_1 = \mathrm{Re}\lambda_2 > \mathrm{Re}\lambda_3 \geq \cdots \geq \mathrm{Re}\lambda_n \tag{3.14}$$

and if $\lambda_1 = a + ib$ *with* $b \neq 0$, *then as* $t \to \infty$, *we have*

$$\begin{aligned} &x_{11}(t) \to a, \quad x_{22}(t) \to a, \quad x_{32}(t) \to 0, \\ &x_{21}(t) \to (\operatorname{sgn} x_{21}(0))|b|, \qquad x_{12}(t) \to -(\operatorname{sgn} x_{21}(0))|b|, \end{aligned} \tag{3.15}$$

and for all $k \geq 3$

$$x_{1k}(t) \to 0, \qquad x_{2k}(t) \to 0. \tag{3.16}$$

Proof. It is clear again from (2.12) that as $t \to \infty$,

$$u_{11}(t) \to \frac{e^{ibt}u_{10}}{\{|u_{10}|^2 + |u_{20}|^2\}^{1/2}}, \qquad u_{21}(t) \to \frac{e^{-ibt}u_{20}}{\{|u_{10}|^2 + |u_{20}|^2\}^{1/2}} \tag{3.17}$$

and for all $i \geq 3$,

$$u_{i1}(t) \to 0. \tag{3.18}$$

Notice that $u_{11}(t)$ and $u_{22}(t)$ do not converge at all. But we still use the notation (3.17) to indicate how they behave when t becomes large. Since X_0 is a real matrix, it must be that $u_{10} = \bar{u}_{20}$. Therefore

$$x_{11}(t) = \langle u_1, Tu_1 \rangle = \sum_{i=1}^{n} \lambda_1 |u_{i1}|^2 \to (a+ib)|\hat{u}_{11}|^2 + (a-ib)|\hat{u}_{21}|^2 = a. \tag{3.19}$$

Thus

$$|x_{21}(t)| = \|Tu_1 - x_{11}u_1\|_2 \to \|T\hat{u}_1 - a\hat{u}_1\| = b \tag{3.20}$$

implies that

$$x_{21}(t)\to\pm b \tag{3.21}$$

where the sign of this limit is the same as that of $x_{21}(0)$ since $x_{21}(t)$ can never change signs. Since $b\neq 0$, it follows, assuming $x_{21}(t)\to b$, from the fact

$$u_2=\frac{Tu_1-x_{11}u_1}{x_{21}} \tag{3.22}$$

that

$$u_{12}(t)\to iu_{11}(t),\quad u_{22}(t)\to iu_{21}(t),\quad u_{i2}(t)\to 0 \tag{3.23}$$

for all $i\geq 3$. So by (2.8), we know

$$x_{22}(t)=\langle u_2, Tu_2\rangle\to a \tag{3.24}$$

and

$$x_{12}(t)=\langle u_1, Tu_2\rangle\to -b. \tag{3.25}$$

By (2.7), simple calculation also shows

$$|x_{32}(t)|=\|Tu_2-x_{12}u_1-x_{22}u_2\|_2\to\|T\hat{u}_2+b\hat{u}_1-a\hat{u}_2\|=0. \tag{3.26}$$

We now claim for all $k\geq 3$, as $t\to\infty$

$$u_{1k}(t)\to 0,\qquad u_{2k}(t)\to 0. \tag{3.27}$$

Indeed this fact follows from solving the following system of equations

$$\langle u_k, Tu_1\rangle=0,\qquad \langle u_k, Tu_2\rangle=0, \tag{3.28}$$

or equivalently

$$\begin{aligned}&\bar{\hat{u}}_{1k}(a+ib)\hat{u}_{11}+\bar{\hat{u}}_{2k}(a-ib)\hat{u}_{21}=0,\\&\bar{\hat{u}}_{1k}(a+ib)i\hat{u}_{11}-\bar{\hat{u}}_{2k}(a-ib)i\hat{u}_{21}=0.\end{aligned} \tag{3.29}$$

Therefore, for all $k\geq 3$,

$$x_{1k}(t)=\langle u_1, Tu_k\rangle=\langle T^*u_1, u_k\rangle\to\langle T^*\hat{u}_1, \hat{u}_k\rangle=0, \tag{3.30}$$

$$x_{2k}(t)=\langle u_2, Tu_k\rangle=\langle T^*u_2, u_k\rangle\to\langle T^*u_2, u_k\rangle=0. \tag{3.31}$$

In summary, this lemma states that if condition (3.14) holds, then

$$X(t)\to\begin{bmatrix}a & -b & 0 & 0 & 0\\ b & a & 0 & 0 & 0\\ 0 & 0 & x & x & x\\ 0 & 0 & x & x & x\\ 0 & 0 & 0 & x & x\end{bmatrix}$$

where again "x" represents uncertain positions.

Finally we note that for the case $b=0$ (multiple eigenvalues), similar results (a 2×2 diagonal block) still can be obtained. Even for the nongeneric case when $\mathrm{Re}\lambda_1=\mathrm{Re}\lambda_2=\mathrm{Re}\lambda_3=\mathrm{Re}\lambda_4$, an argument analogous to Lemma 3.2 can still show the convergence. It is interesting to see the asymptotic behavior of the general flow [1]

$$\dot{X}=[X,\Pi_0(G(x))] \tag{3.32}$$

where $G(z)$ is an analytic function defined on an open set containing the spectrum on X_0. The analysis, nevertheless, is much harder than (1.1) since we don't have a system as nice as (2.9) and we are still working on it.

REFERENCES

[1] M. T. CHU, *The generalized Toda lattice, the QR-algorithm and the centre manifold theory*, SIAM J. Alg. Discrete Meth., 5 (1984), to appear.

[2] P. DEIFT, T. NANDA, AND C. TOMEI, *Differential equations for the symmetric eigenvalue problem*, SIAM J. Numer. Anal., 20 (1983), pp. 1–22.

[3] J. G. F. FRANCIS, *The QR transformation, a unitary analogue to the LR transformation*, Comput. J., 4 (1961), pp. 265–281.

[4] J. MOSER, *Finitely many mass points on the line under the influence of an exponential potential—an integrable system*, Dynamical Systems, Theory and Applications, J. Moser, ed., Lecture Notes in Physics 38, Springer-Verlag, Berlin, 1975, pp. 467–497.

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

[6] A. RALSTON AND P. RABINOWITZ, *A First Course in Numerical Analysis*, McGraw-Hill, New York, 1978.

[7] W. W. SYMES, *The QR-algorithm and scattering for the finite nonperiodic Toda lattice*, Physica, 40 (1982), pp. 275–280.