ABSTRACT

PERTURBATIONS OF STABILITY MATRICES WITH APPLICATIONS TO LOTKA-VOLTERRA MODELS OF ECOSYSTEMS

By

Gary William Harrison

A system of differential equations $x' = Ax + g(x)$, where $g(x) = o(\|x\|)$, has an asymptotically stable equilibrium point at the origin if A is a stability matrix, that is, if all the eigenvalues of A have negative real part. In theory this condition has been used to determine the stability of Lotka-Volterra models of ecosystems, but in practice it breaks down because A can only be estimated.

Methods are developed to establish that a matrix is stable when it is only known approximately. Specifically, bounds on ϵ are found for the perturbed matrix $A + \epsilon B$ to be stable when A is stable. When B is a rank one matrix, necessary and sufficient bounds are derived from the Routh-Hurwitz stability criterion. For a general matrix B, sufficient bounds are derived from the Lyapunov equation $A^T S + SA = -Q$. Perturbations ϵB satisfying these latter bounds are shown to form an

open convex set, leading to a method to compute open convex neighborhoods of a stable matrix A which contain only stable matrices. Similar methods yield sufficient conditions for the bordered matrix $\begin{bmatrix} A & b \\ c^T & d \end{bmatrix}$ to be stable when A is stable.

Iterative methods of the form $N^TS^{(i+1)} + S^{(i+1)}N = N^TS^{(i)} + S^{(i)}N + w(A^TS^{(i)}+S^{(i)}A+Q)$ are investigated to solve the Lyapunov equation. They can converge only if the vector scheme $Nx^{(i+1)} = Nx^{(i)} + w(Ax^{(i)}+q)$ converges, and no faster. Sufficient conditions are found for convergence of several methods corresponding to different choices of N, and examples are given showing that a block Seidel type method is usually most efficient and better than previously known methods of solution for large matrices.

The implications for stability of Lotka-Volterra and other ecosystem models are discussed, and conditions are found for stability to be preserved when a new species is added. A domain of attraction for the equilibrium point, which in some cases is the entire positive quadrant, is found using Lyapunov functions. When coefficients in the model are time varying and the derivative of the variations is small enough, solutions are shown to follow close to a moving critical point. If in addition the coefficients are periodic, there is a periodic solution.

PERTURBATIONS OF STABILITY MATRICES WITH APPLICATIONS TO LOTKA-VOLTERRA MODELS OF ECOSYSTEMS

By

Gary William Harrison

A THESIS

Submitted to
Michigan State University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Department of Mathematics

1975

To Andrea, Martha, and Bryce

ACKNOWLEDGMENTS

I wish to thank those who have contributed to this thesis. Dr. Robert Rosen introduced me to biomathematics and continued to give me helpful suggestions and references even after leaving Michigan State University. Dr. J. Sutherland Frame took over as my advisor when Dr. Rosen left and gave me invaluable guidance during the development and writing of the thesis. Dr. P.K. Wong, Dr. Mary Winter, and Dr. M. Tomber completed my doctoral committee. Many others have provided references or contributed through the classes they taught.

Less tangible but no less real has been the support of family and friends, especially my wife, Andrea and my children Martha and Bryce. I might have written a thesis under a different advisor, but without Andrea's cooperation any thesis would have been impossible. I also wish to thank my mother and father, my brother, Evan, and my office mate Mike McGrew.

My research was supported financially by the National Science Foundation and by the Department of Mathematics at Michigan State University.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER I

INTRODUCTION AND BACKGROUND

1.1. Introduction. When is an ecosystem stable? This is one of the most frequently asked questions in ecology today. In fact, the question of stability is one of the central questions in any mathematical analysis of a dynamical system.

Let $x(t)$ be the vector of "state variables", i.e. the variables which characterize the system, and assume

$$x'(t) = f(x,t). \tag{1}$$

Definition 1.1. If f is a function of x only, the system defined by (1) is said to be autonomous.

Definition 1.2. A vector x^E is called an equilibrium point or a critical point of (1) if $f(x^E,t) = 0$ for all t.

We note that an equilibrium point x^E is a solution of (1). In this thesis we will use the following definition of stability introduced by Lyapunov [25]:

Definition 1.3. An equilibrium point x^E is stable if for every $\epsilon > 0$ and every $t_0 \geq 0$, there is a δ such that any solution of (1) with $\|x(t_0) - x^E\| < \delta$ satisfies $\|x(t) - x^E\| < \epsilon$ for all $t \geq t_0$.

Definition 1.4. An equilibrium point x^E is asymptotically stable if it is stable and for every $t_0 \geq 0$ there is a δ such that $\|x(t_0) - x^E\| < \delta$ implies $x(t)$ converges to x^E as t goes to infinity.

The stability of a general solution $\bar{x}(t)$ of (1) is that of the equilibrium point 0 under the change of variables $y(t) = x(t) - \bar{x}(t)$.

One of the standard techniques to prove stability is given by the following theorem, found in most advanced texts on differential equations. (See for example, Hahn [12], Hale [13], or Rosen [33].)

Definition 1.5. A function $g(x)$ is said to be $o(\|x\|)$, written $g(x) = o(\|x\|)$, if for every $\epsilon > 0$ there is a δ such that $\|x\| < \delta$ implies $\|g(x)\| < \epsilon\|x\|$.

Theorem 1.6 (Lyapunov, 1893). Let

$$x' = Ax + g(x,t), \tag{2}$$

where A is a matrix, and $g(x,t) = o(\|x\|)$. Then if the 0 vector is asymptotically stable for

$$x' = Ax \tag{3}$$

the 0 vector is asymptotically stable for (2).

Theorem 1.6 can be generalized to have nonconstant matrices $A(t)$ in (2) and (3) by introducing the concept of uniform asymptotic stability, but we will not need this generality.

Theorem 1.7. The system (3) is asymptotically stable if and only if all the eigenvalues of A have negative real part.

Theorem 1.7 is our motivation to study matrices whose eigenvalues have negative real part. To avoid repetition, we make the following definition:

Definition 1.8. A is a stability matrix (or a matrix A is said to be stable) if all the eigenvalues of A have negative real part.

Lyapunov gave an alternative method to prove stability, known as his second or direct method. The function V in Theorem 1.9 is called a Lyapunov function. For a proof see Hahn [12] or Yoshizawa [44].

Theorem 1.9. Let 0 be an equilibrium point of (1) and assume there is a differentiable function $V(x,t)$ defined on a neighborhood of 0 in the domain of f, satisfying

(i) $V(0,t) \equiv 0$,

(ii) $a(x) \leq V(x,t)$, for some continuous positive definite function $a(x)$,

(iii) $V'(x(t),t) \leq 0$.

Then 0 is a stable equilibrium point of (1).

If $f(x,t)$ is bounded when x belongs to a compact set and (iii) is replaced by

(iv) $V'(x(t),t) \leq -c(x)$, for some continuous positive definite function $c(x)$,

then 0 is asymptotically stable equilibrium point of (1).

1.2. Lotka-Volterra Model of an Ecosystem. To study the stability of an ecosystem, we will use the Lotka-Volterra model introduced for a two species system by Lotka in 1925 [24], and independently for the general n-species system by Volterra in 1926 [40]. (See also Volterra [41].) Let $p_i(t)$ be the population (or, as is often preferred, the population density) of the ith species at time t in an ecosystem with n species. Then the growth rate $p_i'(t)$ is given by

$$p_i' = p_i\left(r_i + \sum_{j=1}^{n} \alpha_{ij}p_j\right) \tag{4}$$

where r_i is the intrinsic growth rate and α_{ij} is the interaction coefficient measuring the effect of species j on the growth of species i. If species j preys on species i, α_{ij} is negative and α_{ji} positive. If species i and j compete for the same resource α_{ij}

and α_{ji} are both negative. The terms α_{ii} represent self-interaction or self-competition. They were omitted by both Lotka and Volterra in their original formulation, but there is good biological reason to include them and assume that they are less than or equal to zero. We will discuss this system and the significance of the coefficients more fully in Chapter 5. We also refer the reader to Rosen [33], who discusses the relationship of (4) to the "law of mass action" of chemistry.

We assume that the coefficients in (4) are such that there is an equilibrium point p^E satisfying

$$r_i + \sum \alpha_{ij} p_j^E = 0 \qquad i = 1,2,\ldots,n \tag{5}$$

with $p_j^E > 0$ for all j. Clearly, negative values for p_i have no meaning. There are other equilibrium points where $p_i = 0$ for some i and (5) is satisfied for the other values of i, but a zero value for p_i corresponds to extinction for that species. To investigate the stability of p^E we make the change of co-ordinates

$$x = p - p^E. \tag{6a}$$

The stability properties of p^E are the same as those of the zero vector in

$$x_i' = \sum_{j=1}^{n} p_i^E \alpha_{ij} x_j + \sum_{j=1}^{n} x_i \alpha_{ij} x_j, \qquad i = 1,2,\ldots,n. \tag{6b}$$

The zero vector is asymptotically stable in (6) if the matrix

$$A = [a_{ij}], \quad a_{ij} = p_i^E \alpha_{ij} \tag{7}$$

is a stability matrix. Several authors (May [28], Rosen [33], and Strobeck [36]) have discussed the stability of p^E by examining the eigenvalues of A.

It would appear that for an actual ecosystem the mathematical analysis of stability would be rather straightforward, since several numerical techniques for finding the eigenvalues of a matrix are well developed. But this presupposes an exact knowledge of the coefficients r_i and α_{ij}, whereas the biologist will know the signs of the α_{ij} but will only be able to give rough estimates for their magnitudes. This research started with the question: If the entries of A are known only approximately, can we still guarantee that A is a stability matrix?

1.3. Description of the Problem. To put the problem more precisely, if A is a given stability matrix and B is a given matrix, we want to know for what values of ε, $A + \varepsilon B$ is a stability matrix. More generally, let E_{ij} denote the matrix with 1 in the (i,j) position and zeroes elsewhere. We would like to find an open region about the origin in n^2-space, such that if the n^2-vector (ε_{ij}) is in the region,

$A + \sum_{i,j} \epsilon_{ij} E_{ij}$ is a stability matrix. That is, we would like to be able to perturb the entries of A independently, not just in a prescribed ratio as is the case when the perturbation has the form ϵB. While most perturbation studies only deal with changes that are small compared with the original values, we will want to investigate perturbations that are of the same magnitude as the original matrix. In fact, we seek bounds on ϵ that are as large as possible.

In Chapter II, we find sufficient conditions for the perturbed matrices to be stable in both of the above cases, by using the criterion of Lyapunov for a stability matrix. Part of the significance of these results is that the bounds on ϵ and on ϵ_{ij} are not just theoretical, but methods are developed to compute them. We also find sufficient conditions for the bordered matrix $\begin{bmatrix} A & b \\ c^T & d \end{bmatrix}$ to be stable when A is stable.

The methods of Chapter II depend on the solution S of the Lyapunov matrix equation

$$A^T S + SA = -Q \tag{8}$$

for a given symmetric, positive definite matrix Q. In Chapter III, we develop iterative procedures to solve (8), which appear to be superior to existing procedures if the matrix A is large.

In Chapter IV, we use the Routh-Hurwitz criterion for stability matrices to find computable, necessary and sufficient bounds on ϵ for $A + \epsilon B$ to be a stability matrix, if B is a rank one matrix.

In Chapter V, we return to an examination of the stability of the Lotka-Volterra system (4). Besides answering our original questions from the methods of Chapters II and IV, we use the criterion for a bordered matrix to be stable to find conditions for an ecosystem to remain stable when a new species is added, and use Lyapunov functions based on the solution S of equation (8) to establish a domain of attraction for the equilibrium point p^E, that is, a neighborhood of p^E such that any solution starting in this neighborhood tends to p^E. Finally, we consider not just constant, but also time varying perturbations $B(t)$ and show that if the derivative of $B(t)$ is small enough, the solutions stay close to a certain time-varying critical vector $c(t)$. If, in addition, $B(t)$ is periodic of period T, there exists a periodic solution of period T.

All the material in this thesis after section 2.1 is the original work of the author unless explicitly stated otherwise.

1.4. Notation. Throughout this thesis we will use capital letters, such as A, to denote matrices and the corresponding subscripted lower case letters a_{ij} to denote the (i,j)-entry of A. Alternately $A = [a_{ij}]$ will mean A is the matrix with a_{ij} as its (i,j) entry. All matrices are assumed to be real unless stated otherwise. We will use column vectors, which will be denoted by lower case letters, such as v, and the subscripted letter v_i will denote the ith component of the vector v. We will also use the notation $v = (v_i)$ to define v as the column vector with ith component v_i. Eigenvalues of matrices will generally be denoted by lower case Greek letters. Since real matrices can have complex eigenvalues and eigenvectors, we will not assume that any eigenvalue or eigenvector is real unless so stated. If unspecified, matrices will be assumed to be $n \times n$ and vectors to have dimension n.

The transpose of a vector or a matrix will be denoted by v^T or A^T, and the conjugate transpose by v^* (or A^* should A be specified as a complex matrix). The symbol v' will indicate the derivative of v with respect to time.

$D = \text{Diag}[d_1 \ldots d_n]$ will mean that D is a diagonal matrix with entries $d_{ii} = d_i$, $d_{ij} = 0$ if $i \neq j$. When

v is a vector, $D(v)$ will denote the diagonal matrix $\text{Diag}[v_1 \ldots v_n]$.

The identity matrix will be denoted by I. If unspecified it will be $n \times n$. To specify a different dimension, say $m \times m$, we will use a subscript I_m. For two vectors u and v (or two matrices A and B) $u \ll v$ (or $A \ll B$) will mean that each entry of $v - u$ (of $B - A$) is positive.

The spectral radius of a matrix A, the largest absolute value of an eigenvalue of A, will be denoted by $\sigma(A)$.

The symbol $\|v\|$ will denote any vector norm. We will often use the vector norms defined by

$$\|v_1\| = \sum_{i=1}^{n} |v_i| \tag{9.1}$$

$$\|v\|_2 = \sqrt{v^T v} \tag{9.2}$$

$$\|v\|_\infty = \sup_i |v_i| \tag{9.3}$$

Any matrix norm, $\|A\|$, satisfying

$$\|Ax\| \leq \|A\|\|x\|, \tag{10}$$

where $\|Ax\|$ and $\|x\|$ are a vector norm, is said to be consistent with that vector norm. We will make use of the matrix norms

$$\|A\|_1 = \sup_{\|v\|_1=1} \|Av\|_1 = \max_j \sum_{i=1}^{n} |a_{ij}|, \tag{11.1}$$

$$\|A\|_2 = \sup_{\|v\|_2=1} \|Av\|_2 = \sqrt{\sigma(A^T A)}, \tag{11.2}$$

$$\|A\|_\infty = \sup_{\|v\|_\infty=1} \|Av\|_\infty = \max_i \sum_{j=1}^{n} |a_{ij}| \tag{11.3}$$

which are consistent with the corresponding vector norms. It is well known that for any consistent matrix norm, $\sigma(A) \leq \|A\|$. (See, for example, Isaacson and Keller [18] for a proof of this and equations (11.1)-(11.3).) Whenever we use the symbol $\|A\|$, it will stand for a consistent matrix norm.

1.5. Background. The question of the location of the eigenvalues of a matrix has often been examined, although the results are not usually stated in terms of perturbations of a stability matrix which preserve its stability. We now review some of these results and show how they are applicable to our problem. We assume throughout that the matrix A is a stability matrix.

1.5.1. Continuity of Eigenvalues. It is well known that the eigenvalues of a matrix are continuous functions of its entries. Thus we know that there is an open region in n^2 space containing A which contains only stability matrices.

1.5.2. Sensitivity Analysis. This approach to studying the effect of perturbations, based on the derivative of the eigenvalue, is often used in engineering applications. (See Porter and Crossley [31] or Tomović [38]). Let $M(\epsilon)$ be a matrix function of ϵ having eigenvalue $\lambda(\epsilon)$ and corresponding right and left eigenvectors $x(\epsilon)$ and $y^T(\epsilon)$. Differentiating

$$M(\epsilon)x(\epsilon) = \lambda(\epsilon)x(\epsilon) \tag{12}$$

with respect to ϵ, multiplying on the left by $y^T(\epsilon)$, and solving for $\lambda'(\epsilon)$, yields

$$\lambda'(\epsilon) = \frac{y^T(\epsilon)M'(\epsilon)x(\epsilon)}{y^T(\epsilon)x(\epsilon)} . \tag{13}$$

In particular, if $M = A + \epsilon B$ and $x = x(0)$, $y^T = y^T(0)$ are the right and left eigenvectors of A corresponding to $\lambda = \lambda(0)$,

$$\lambda'(0) = \frac{y^T Bx}{y^T x} . \tag{14}$$

Setting $B = E_{ij}$ yields

$$\frac{\partial \lambda}{\partial a_{ij}} = \frac{y_i x_j}{y^T x} . \tag{15}$$

The eigenvalues are most sensitive to changes in the entries of A for which $\frac{\partial \lambda}{\partial a_{ij}}$ is largest. For small enough ϵ the eigenvalues of $A + \epsilon B$ can be approximated by $\lambda(0) + \epsilon\lambda'(0)$. But we do not really know $\lambda'(\epsilon)$ for

$\epsilon \neq 0$, since we do not know $x(\epsilon)$ and $y(\epsilon)$ for $\epsilon \neq 0$. Without this knowledge we do not have an error bound for the approximation to $\lambda(\epsilon)$, so that we cannot say for certain that $\lambda(\epsilon)$ has not moved far enough away from $\lambda(0)$ to cross the imaginary axis into the right half plane.

1.5.3. Characteristic Equation. Since the eigenvalues of A satisfy the characteristic equation

$$\det[A-\lambda I] = \lambda^n + c_1\lambda^{n-1} + \ldots + c_{n-1}\lambda + c_n = 0, \tag{16}$$

it is necessary for all the coefficients c_i to be positive in order for all the eigenvalues to have negative real part. If all the eigenvalues of A are known to be real (for example if A is symmetric) this is both a necessary and sufficient condition for A to be a stability matrix. The Routh-Hurwitz criterion discussed in Chapter IV is a strengthening of this condition to make it necessary and sufficient for general matrices.

In particular, we note that it is necessary that $\operatorname{tr} A = -c_1$ = sum of the eigenvalues be negative and that $\det A = (-1)^n c_n$ = product of the eigenvalues have the same sign as $(-1)^n$ for A to be stable.

1.5.4. Two By Two Matrices. If A is a 2×2 matrix it is necessary and sufficient for A to be stable that $\operatorname{tr} A$ be negative and $\det A$ be positive, which is useful for constructing counterexamples. Furthermore, the

eigenvalues of A are given by

$$\lambda = \frac{a_{11}+a_{22}}{2} \pm \sqrt{\left(\frac{a_{11}-a_{22}}{2}\right)^2 + a_{12}a_{21}} \quad (17)$$

Perturbations of A which decrease the diagonal entries of A usually are stabilizing, but not always, since $\begin{bmatrix} -2 & 1 \\ -3 & 1 \end{bmatrix}$ is stable, but $\begin{bmatrix} -4 & 1 \\ -3 & 1 \end{bmatrix}$ is not. In this case changing a_{11} from -2 to -4 decreased the average value of the eigenvalues, $\frac{a_{11}+a_{22}}{2}$, but increased the distance between them, $2\sqrt{\left(\frac{a_{11}-a_{22}}{2}\right)^2 + a_{12}a_{21}}$. Perturbations of the off diagonal entries which increase $a_{12}a_{21}$ (symmetric perturbations) move the eigenvalues further apart and hence are destabilizing. Perturbations which decrease $a_{12}a_{21}$ (skew-symmetric perturbations) move the eigenvalues together. When $a_{12}a_{21}$ becomes negative enough to make the quantity under the radical zero, the eigenvalues meet and further reduction of $a_{12}a_{21}$ yields complex conjugate eigenvalues. Thus skew symmetric perturbations of a 2×2 matrix with negative trace tend to be stabilizing or at least not destabilizing, though they may lead to imaginary components in the eigenvalues, which correspond to oscillations in the solutions of equation (3). We will see evidence that these observations are also true for larger matrices. A precise mathematical formulation of this principle for general matrices would be valuable.

1.5.5. Gershgorin's Theorem and Generalizations. Gershgorin's theorem, that the eigenvalues λ of a matrix R lie within the union of the circles defined by either

$$|\lambda - r_{ii}| \leq \sum_j |r_{ij}| \tag{18}$$

or

$$|\lambda - r_{ii}| \leq \sum_j |r_{ji}| \tag{19}$$

is well known. Thus if R is diagonally dominant and has negative diagonal entries, it is a stability matrix and any perturbation which preserves this characteristic preserves its stability.

Generalizations of Gershgorin's theorem may be formulated (see Householder [17] or Wilkinson [43]) by writing a matrix R as $E + C$, where, for some non-singular matrix P, $D = P^{-1}EP$ is diagonal. (Here we drop our assumption that D and P be real.) Then $D + P^{-1}CP$ is similar to R, so that for any eigenvalue λ of R, there is an eigenvector x such that

$$Dx + P^{-1}CPx = \lambda x. \tag{20}$$

It follows that

$$1 \leq \|(D-\lambda I)^{-1} P^{-1}CP\| \tag{21}$$

and

$$\min_i |d_{ii} - \lambda| \leq \|P^{-1}CP\| \leq \|P^{-1}\| \, \|P\| \, \|C\|. \tag{22}$$

If E is the diagonal part of R, and $P = I$, using the norms $\| \|_1$ and $\| \|_\infty$ in (21) yields equations (18) and

(19) respectively. Various other choices of E and the norm used give various generalizations of Gershgorin's theorem.

This can be applied to perturbations of a stability matrix A in two ways. First, if A is diagonalizable, letting $E = A$ and $C = \epsilon B$ shows that $A + \epsilon B$ is stable for any ϵ satisfying

$$\epsilon \|P^{-1}BP\| \leq \min_i \operatorname{Re} |d_{ii}|. \tag{23}$$

Letting $E = A$ and $C = \sum_{i,j} \epsilon_{ij} E_{ij}$ shows that $A + \sum_{i,j} \epsilon_{ij} E_{ij}$ is stable as long as

$$\| \sum_{i,j} \epsilon_{ij} E_{ij} \| \leq \frac{1}{\|P^{-1}\| \|P\|} \min_i \operatorname{Re} |d_{ii}|. \tag{24}$$

If A is normal then P is unitary and $\|P\|_2 = \|P^{-1}\|_2 = 1$, but if A has nearly parallel eigenvectors, which generally happens when A has two almost equal eigenvalues, then the condition number $\|P^{-1}\| \|P\|$ may be very large. In this case, computation of D and P will be difficult and may be unstable. (Unstable is used here in the sense of numerical analysis, that small round-off errors will cause large errors in the final answer.)

Second, suppose that $A = E + C$, that E is stable, and that

$$\|P^{-1}CP\| \leq \min_i \operatorname{Re} |d_{ii}|. \tag{25}$$

Then A is stable and any perturbation of A which preserves (25), preserves its stability. It should be noted, however, that perturbations which change E, not only change the values of d_{ii} but also change the matrix P.

1.5.6. Symmetric and Skew Symmetric Parts. Any real matrix A can be written as

(26) $$A = M + K$$

where $M = \frac{1}{2}(A^T+A)$ is symmetric and $K = \frac{1}{2}(A-A^T)$ is skew symmetric. It is well known that the real parts of the eigenvalues of A lie between the smallest and largest eigenvalues of M, and that their imaginary parts lie between the conjugate pair of pure imaginary eigenvalues of K of largest magnitude. Hence for M to be negative definite is a sufficient condition for A to be stable. If A is normal, the real parts of the eigenvalues of A are the eigenvalues of M and their imaginary parts are the eigenvalues of K, so that M negative definite is a necessary and sufficient condition for a normal matrix A to be stable. The stability criterion of Lyapunov discussed in Chapter II is a weakening of this condition to make it both sufficient and necessary for general matrices. The next lemma is a special case of a well known inequality for the eigenvalues of the sum of two symmetric matrices. (See Wilkinson [43].)

Lemma 1.10. Let M and V be symmetric matrices, u, v, and λ be the largest eigenvalues of M, V, and $M + V$ respectively. Then

$$\lambda \leq u + v. \tag{27}$$

Proof:

$$\lambda = \sup_x \frac{x^*(M+V)x}{x^*x} \leq \sup_x \frac{x^*Mx}{x^*x} + \sup_x \frac{x^*Vx}{x^*x} = u + v. \tag{28}$$

Corollary 1.11. Let A satisfy (26), $B = V + L$ with V symmetric and L skew symmetric, and u and v be the largest eigenvalues of M and V, respectively. Assume $u < 0$. Then $A + \epsilon B$ is stable for all ϵ such that $u + \epsilon v < 0$.

Proof: The symmetric part of $A + \epsilon B$ is $M + \epsilon V$, which by the Lemma has largest eigenvalue $\lambda(\epsilon) \leq u + \epsilon v$.

This simple Corollary is often a very practical way to establish that a perturbation preserves stability, and is the starting point for the methods of Chapter II.

1.5.7. Sign Stable Matrices. Quirk and Ruppert [32] (see also Maybee and Quirk [29]) introduced the concept of sign stability in connection with economic problems and proved the following theorem.

Definition. A matrix A is sign-stable if every matrix C such that c_{ij} is negative, zero, or positive whenever a_{ij} is negative, zero, or positive, respectively, is stable.

Theorem 1.12. An indecomposable matrix A is sign-stable if and only if A satisfies the conditions

(i) $a_{ii} \leq 0$ for every i, $a_{ii} < 0$ for some i.

(ii) $a_{ij}a_{ji} \leq 0$ for every $i \neq j$.

(iii) For any sequence of 3 or more indices $i,j,k,\ldots,r$ (with $i \neq j \neq k \neq \ldots \neq r$), the product $a_{ij}a_{jk}\cdots a_{ri} = 0$.

(iv) There exists a nonzero term in the expansion of $\det A$.

Of course a perturbation of a sign-stable matrix which preserves the sign pattern of the entries preserves the stability. Since the signs of the α_{ij} of equation (5) are easily determined but their magnitudes are not, we will make use of sign stability in our applications.

1.5.8. Norm of $(A+I)(A-I)^{-1}$. Let $C = (A+I)(A-I)^{-1}$. Then each eigenvalue γ of C is given by

$$\gamma = \frac{\lambda+1}{\lambda-1}, \tag{29}$$

where λ is an eigenvalue of A. The linear fractional transformation $\lambda \to \frac{\lambda+1}{\lambda-1}$ maps the left half plane onto the interior of the unit circle. Therefore A is stable if and only if $\sigma(C) < 1$.

Since $\sigma(C) \leq \|C\|$ for any consistent matrix norm, $\|C\| < 1$ is a sufficient condition for A to be stable. But it is difficult to apply this condition to the stability

of the perturbed matrix $A + \epsilon B$, because of the difficulty of expressing $\|(A+\epsilon B+I)(A+\epsilon B-I)^{-1}\|$ as a function of ϵ.

1.5.9. Canonical Forms. Stability of a matrix A can be determined by transforming it to one of the various canonical forms which display the eigenvalues. Or one can use the Schwarz canonical form [34] (see also Barnett and Storey [4]), which does not display the eigenvalues, but has all positive entries on the subdiagonal if and only if A is stable. Canonical forms are not well suited to perturbation problems, since the transformations involve long computational processes which obscure the relationship of the canonical form of A to that of $A + \epsilon B$.

CHAPTER II

APPLYING LYAPUNOV'S EQUATION

2.1. The Lyapunov Equation. To study the preservation of the stability of a matrix A under perturbations we need a necessary and sufficient condition for A to be stable. The most useful of these is the following theorem due to Lyapunov:

Theorem 2.1 (Lyapunov). If there exist symmetric positive definite matrices S and Q such that

$$A^T S + SA = -Q \tag{1}$$

then A is a stability matrix. Conversely if A is a stability matrix, then for any positive definite symmetric matrix Q, there is a unique positive definite symmetric matrix S which satisfies equation (1).

Lyapunov's equation (1) has many uses and is discussed in many places in the literature. Bellman [5] and Barnett and Storey [4] both have detailed expositions. In this section we review the pertinent results from the literature. Section 2.2 contains original contributions to the theory of Lyapunov's equation, and sections 2.3,

2.4, and 2.5 are new applications of Lyapunov's equation, using it to study perturbations and bordering of stability matrices. Throughout this chapter, S and Q will always stand for symmetric positive definite matrices unless noted otherwise.

That equation (1) is sufficient for A to be stable is closely related to the stability of the linear differential equation

$$x' = Ax. \tag{2}$$

If we form the positive definite quadratic form x^TSx, then

$$(x^TSx)' = x^TA^TSx + x^TASx = -x^TQx, \tag{3}$$

and x^TSx is a positive definite Lyapunov function for (2) with a negative time derivative. By Theorem 1.9, this implies that the zero solution of (2) is asymptotically stable, which can happen only if all the eigenvalues of A have negative real part, i.e., only if A is a stability matrix.

A proof that a positive definite solution S for equation (1) exists when A is a stability matrix can be found in Bellman [5], who gives the integral representation

$$S = \int_0^\infty e^{A^Tt}\, Q\, e^{At}\, dt. \tag{4}$$

Olga Taussky [37] has given an algebraic proof.

To better understand equation (1), we examine the operator $\bar{A}$ which maps the space of $n \times n$ matrices to the space of $n \times n$ matrices and is defined by

$$\bar{A}X = A^T X + XA. \tag{5}$$

It is easily seen that $\bar{A}$ is linear, and that it maps symmetric matrices to symmetric matrices and skew-symmetric matrices to skew symmetric matrices. It is also apparent that equation (1) is linear in A, so that

$$\overline{\alpha A + \beta B} = \alpha\bar{A} + \beta\bar{B} \tag{6}$$

for any matrices A and B and scalars α and β.

For an $n \times n$ matrix X, let X_V denote the $n^2 \times 1$ column vector formed by taking X row by row. Then for any two matrices B and C,

$$(BXC)_V = (B \times C^T) X_V,$$

where $\times$ denotes the Kronecker product. It follows that

$$(\bar{A}X)_V = (A^T X + XA)_V = ((A^T \times I) + (I \times A^T)) X_V. \tag{7}$$

If

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \tag{8}$$

then

$$(9) \qquad (A^T \times I) + (I \times A^T) = \begin{bmatrix} A^T + a_{11}I & a_{21}I & \cdots & a_{n1}I \\ a_{12}I & A^T + a_{22}I & \cdots & a_{n2}I \\ \vdots & \vdots & & \\ a_{1n}I & a_{2n}I & \cdots & A^T + a_{nn}I \end{bmatrix}$$

It is readily seen that if A is diagonal, triangular, symmetric, or skew symmetric, then $(A^T \times I) + (I \times A^T)$ has the same property.

Theorem 2.2. If a matrix A has eigenvalues $\lambda_1 \cdots \lambda_n$, then the eigenvalues of the linear operator $\bar{A}$ are precisely the sums of all pairs of eigenvalues $\lambda_i + \lambda_j$.

Proof: Let RAR^{-1} be upper triangular, so that the eigenvalues of A are the diagonal elements of RAR^{-1}. Then

$$(10) \qquad (R^{T^{-1}} \times R^{T^{-1}})[(A^T \times I) + (I \times A^T)](R^T \times R^T)$$
$$= R^{T^{-1}} A^T R^T \times I + I \times R^{T^{-1}} A^T R^T$$

which is lower triangular. The eigenvalues of $\bar{A}$ are the diagonal entries of $R^{T^{-1}} A^T R^T \times I + I \times R^{T^{-1}} A^T R^T$ which are the sums of pairs of diagonal entries of RAR^{-1}.

In fact if x and y are eigenvectors of A^T, so that $A^T x = \lambda_1 x$ and $y^T A = y^T \lambda_2$, then xy^T is the

"eigenvector" of $\bar{A}$ corresponding to $\lambda_1 + \lambda_2$, since

$$\bar{A}(xy^T) = A^T(xy^T) + (xy^T)A = \lambda_1 xy^T + xy^T\lambda_2 = (\lambda_1+\lambda_2)xy^T. \tag{11}$$

Corollary 2.3. If no two eigenvalues λ_1, λ_2 (not necessarily distinct) of A satisfy $\lambda_1 + \lambda_2 = 0$, then the matrix equation $A^TX + XA = Y$ has a unique solution X for every matrix Y.

Proof: The hypothesis guarantees that the linear operator $\bar{A}$ has no zero eigenvalues, hence it is non-singular.

Corollary 2.4. If no two eigenvalues λ_1, λ_2 (not necessarily distinct) of A satisfy $\lambda_1 + \lambda_2 = 0$, and Q is symmetric, then the solution S to $A^TS + SA = -Q$ is symmetric.

Proof:

$$A^TS^T + S^TA = -Q^T = -Q. \tag{12}$$

By the uniqueness of solutions, $S^T = S$.

It is clear that any stability matrix satisfies the hypothesis of the preceding two corollaries, but any singular matrix or matrix with a conjugate pair of pure imaginary eigenvalues fails to do so.

2.2. The Lyapunov Equations and the Symmetric Part of a Matrix. It does not seem to be widely recognized that there is a close connection between the Lyapunov equation (1) and negative definiteness of the symmetric part of a matrix. Equation (1) shows that a matrix A is stable if and only if for some positive definite symmetric matrix S, the matrix SA has negative definite symmetric part. The following corollary to Lyapunov's theorem expresses the relation between stability matrices and negative definite matrices in another way.

Corollary 2.5. A is a stability matrix if and only if A is similar to a matrix with negative definite symmetric part.

Proof: If A is similar to a matrix with negative definite symmetric part it is clear that A is a stability matrix.

If A is a stability matrix, equation (1) holds. Let $S = R^TR$, R nonsingular. Multiplying (1) on the left by $R^{T^{-1}}$ and on the right by R^{-1} we have

$$R^{T^{-1}}A^TR^T + RAR^{-1} = -R^{T^{-1}}QR^{-1}. \tag{13}$$

Thus $-R^{T^{-1}}QR^{-1}$, which is negative definite, is twice the symmetric part of RAR^{-1}.

If S is a symmetric positive definite matrix, we can form an inner product $(x,y)_S = x^*Sy$. We call a matrix B S-symmetric if $(x,By)_S = (Bx,y)_S$ and S-skew-symmetric if $(x,By)_S = -(Bx,y)_S$. It is readily verified that V is S-symmetric if and only if SV is symmetric, and that an S-symmetric matrix has all real eigenvalues $\{\nu_j\}$ with

$$\min_j \nu_j = \min_x \frac{(x,Vx)_S}{(x,x)_S} = \min_x \frac{x^*SVx}{x^*Sx} \tag{14}$$

$$\max_j \nu_j = \max_x \frac{(x,Vx)_S}{(x,x)_S} = \max_x \frac{x^*SVx}{x^*Sx} . \tag{15}$$

Likewise W is S-skew-symmetric if and only if SW is skew symmetric and an S-skew-symmetric matrix has pure imaginary eigenvalues $\{\pm i\omega_j\}$ with

$$\min_j \omega_j = \min_x \left|\frac{(x,Wx)_S}{(x,x)_S}\right| = \min_x \left|\frac{x^*SWx}{x^*Sx}\right| \tag{16}$$

$$\max_j \omega_j = \max_x \left|\frac{(x,Wx)_S}{(x,x)_S}\right| = \max_x \left|\frac{x^*SWx}{x^*Sx}\right| . \tag{17}$$

Lemma 2.6. Let S be a symmetric positive definite matrix. Any matrix A can be written as a sum of an S-symmetric matrix V and an S-skew-symmetric matrix W. If ν_1 and ν_n are the smallest and largest eigenvalues of V, respectively, and $i\omega_1$ and $i\omega_n$, $\omega_i \geq 0$, are the smallest and largest (in absolute value) eigenvalues of W, respectively, then any eigenvalue λ of A satisfies

(18) $$\nu_1 \leq \mathrm{Re}(\lambda) \leq \nu_n$$

(19) $$\omega_1 \leq |\mathrm{Im}(\lambda)| \leq \omega_n .$$

Proof: Let $V = \frac{1}{2}(S^{-1}A^TS+A)$ and $W = \frac{1}{2}(-S^{-1}A^TS+A)$.

Furthermore, for some x

(20) $$\lambda x = Ax = Vx + Wx$$

from which it follows that

(21) $$\lambda = \frac{x^*SVx}{x^*Sx} + \frac{x^*SWx}{x^*Sx} .$$

From the symmetry, x^*Sx and x^*SVx are real, and SW skew-symmetric implies that x^*SWx is pure imaginary. Equations (18) and (19) now follow from (14)-(17).

This leads us naturally to the following theorem, first proved in separate parts by Vogt [39], and Barnett and Storey [4] using different methods.

Theorem 2.7. Let A be a matrix with eigenvalues $\{\lambda_k\}$ and S a positive,definite symmetric matrix. Let $A^TS + SA = M$ and $-A^TS + SA = K$. Then $V = \frac{1}{2}S^{-1}M$ has real eigenvalues $\{\nu_1 \leq \nu_2 \leq \ldots \leq \nu_n\}$ and

(22) $$\frac{1}{2}\min_x \frac{x^*Mx}{x^*Sx} = \nu_1 \leq \mathrm{Re}(\lambda_k) \leq \nu_n = \frac{1}{2}\max_x \frac{x^*Mx}{x^*Sx}$$

for every k.

Likewise $W = \frac{1}{2}S^{-1}K$ has pure imaginary eigenvalues $\{\pm i\omega_j\}$, $0 \le \omega_1 \le \omega_2 \le \ldots \le \omega_n$ and

$$(23) \qquad \frac{1}{2}\min_x \left|\frac{x^*Kx}{x^*Sx}\right| = \omega_1 \le \mathrm{Im}(\lambda_k) \le \omega_n = \frac{1}{2}\max_x \left|\frac{x^*Kx}{x^*Sx}\right|$$

for every k.

Proof: V is the S-symmetric part of A and W is the S-skew symmetric part of A. The inequalities in (22) and (23) now follow from (18) and (19). The equalities in (22) and (23) follow from (14)-(17) since $x^*SVx = \frac{1}{2}x^*Mx$ and $x^*SWx = \frac{1}{2}x^*Kx$.

When $M = -Q$ with Q positive definite, Theorem 2.7 gives us another proof of the sufficiency of the Lyapunov condition for A to be stable.

Corollary 2.8. Let $A^TS + SA = -I$, A have eigenvalues with real parts $\alpha_1 \le \ldots \le \alpha_n < 0$, and σ_1 and σ_n be the smallest and largest eigenvalues of S, respectively. Then

$$(24) \qquad \sigma_1 \le -\frac{1}{2\alpha_1} \le -\frac{1}{2\alpha_n} \le \sigma_n \le \|S\|.$$

Proof: The smallest and largest eigenvalues of $-\frac{1}{2}S^{-1}$ are $-\frac{1}{2\sigma_1}$ and $-\frac{1}{2\sigma_n}$, respectively. By Theorem 2.7 with $M = -I$

$$(25) \qquad -\frac{1}{2\sigma_1} \le \alpha_1 \le \alpha_n \le -\frac{1}{2\sigma_n}.$$

Equation (24) follows immediately, $\sigma_n \le \|S\|$ being true for any consistent norm.

The following stability concepts have been introduced in economic theory:

Definition 2.9. A matrix A is D-stable if for any matrix $D = \text{diag}[d_1, d_2, \ldots, d_n]$, DA is stable if and only if $d_i > 0$ for every i.

Definition. A matrix A is S-stable if for any symmetric matrix S, SA is stable if and only if S is positive definite.

Clearly S-stable $\Rightarrow$ D-stable $\Rightarrow$ stable. Arrow and McManus [2] have shown the following:

Theorem 2.10. If $A^T + A$ is negative definite then A is S-stable.

We now show that Theorem 2.10 is equivalent to Lyapunov's theorem.

Suppose Lyapunov's theorem is true and $A^T + A$ is negative definite. If SA is stable,

$$(A^T S)S^{-1} + S^{-1}(SA) = A^T + A, \tag{26}$$

and by Lyapunov's theorem S^{-1} is positive definite, which implies S is positive definite. Conversely if S is positive definite, so is S^{-1} and (26) is Lyapunov's equation for SA with $-Q = A^T + A$.

On the other hand suppose the theorem of Arrow and McManus is true. If equation (1) holds with S and Q positive definite, then VSA is stable for any positive definite V. Taking $V = S^{-1}$ shows that A is stable. Conversely, if A is stable we know that for any Q a solution S to equation (1) exists. We need only show that S is positive definite. But SA satisfies the hypothesis of Arrow and McManus. $S^{-1}(SA) = A$ is stable implies S^{-1}, and hence S, is positive definite.

Corollary 2.11. Assume S and Q are symmetric, positive definite, $A^TS + SA = -Q$, and V is symmetric and commutes with S. Then VA and AV are stable if and only if V is positive definite.

Proof: Since V and S commute, VS^{-1} is symmetric. $VA = VS^{-1}SA$, which by Arrow and McManus is stable if and only if VS^{-1} is positive definite, which happens if and only if V is positive definite.

Corollary 2.12. If there is a positive definite diagonal matrix N and positive definite Q such that $A^TN + NA = -Q$, then A is D-stable.

Proof: Since any diagonal matrix D commutes with N, by the preceding corollary, DA is stable if and only if D is positive definite.

Definition 2.13. A matrix A is D-negative-definite if there exists a diagonal matrix D such that DA has negative definite symmetric part.

We will have applications of D-stability and D-negative definiteness in Chapter V.

2.3. Stability Preserving Perturbations by Lyapunov's Equation. We now come to our main purpose for studying the Lyapunov equation, namely finding sufficient conditions for a perturbed matrix $A + \epsilon B$ to be stable when A is known to be stable. The basic idea is given by the following lemma, which we will call the Fundamental Lemma.

Lemma 2.14. Let $A^TS + SA = -Q$, with S and Q symmetric positive definite matrices. If $-Q + \epsilon(B^TS+SB)$ is negative definite, then $A + \epsilon B$ is stable.

Proof:

$$(A+\epsilon B)^TS + S(A+\epsilon B) = -Q + \epsilon(B^TS+SB). \tag{27}$$

The right hand side of (27) is negative-definite.

Thus by Lyapunov's theorem $A + \epsilon B$ is stable.

The next three theorems illustrate how the fundamental lemma can be applied.

Theorem 2.15. Let $A^TS + SA = -Q$, S and Q symmetric positive-definite. Let ν be the largest eigenvalue of $B^TS + SB$ and q be the smallest eigenvalue of Q. Then $A + \epsilon B$ is stable for all $\epsilon > 0$ such that

$$\epsilon\nu - q < 0. \tag{28}$$

Proof: By Lemma 1.10 the largest eigenvalue of $-Q + \epsilon(B^TS+SB)$ is less than $\epsilon\nu - q < 0$. Thus $-Q + \epsilon(B^TS+SB)$ is negative-definite. The theorem now follows from Lemma 2.14.

Theorem 2.16. Let $A^TS + SA = -Q$, S and Q symmetric negative-definite. If

$$\epsilon(\|B^T\| + \|B\|) \leq \frac{1}{\|S\|\|Q^{-1}\|}, \tag{29}$$

then $A + \epsilon B$ is stable.

Proof: Let q be the smallest eigenvalue of Q and ν the largest eigenvalue of $B^TS + SB$. Then

$$\epsilon\nu \leq \epsilon\|B^TS + SB\| \leq \epsilon(\|B^T\| + \|B\|)\|S\| \leq \frac{1}{\|Q^{-1}\|} \leq q. \tag{30}$$

Thus $\epsilon\nu - q$ is negative and $A + \epsilon B$ is stable by Theorem 2.15.

We note that if we use the 2-norm, then $\|B^T\| = \|B\|$, and the condition (29) for $A + \epsilon B$ to be stable becomes

(31) $$\epsilon\|B\|_2 \leq \frac{1}{2\|S\|_2\|Q^{-1}\|_2} .$$

Using the 1-norm or the ∞-norm, $\|B^T\|_1 = \|B\|_\infty$, and condition (29) becomes

(32) $$\|B\|_1 + \|B\|_\infty \leq \frac{1}{\|S\|_\infty\|Q^{-1}\|_\infty} = \frac{1}{\|S\|_1\|Q^{-1}\|_1} .$$

Theorem 2.17. Let S and Q be symmetric positive-definite matrices, $A^TS + SA = -Q$, and $\eta_1 \leq \eta_2 \leq \ldots \leq \eta_n$ be the eigenvalues of $Q^{-1}(B^TS+SB)$. Then $A + \epsilon B$ is stable:

(33a) for all $\frac{1}{\eta_1} < \epsilon < \frac{1}{\eta_n}$, if $\eta_1 < 0 < \eta_n$;

(33b) for all $\frac{1}{\eta_1} < \epsilon < \infty$, if $\eta_1 < \eta_n \leq 0$;

(33c) for all $-\infty < \epsilon < \frac{1}{\eta_n}$, if $0 \leq \eta_1 < \eta_n$.

Proof: $(A+\epsilon B)^TS + S(A+\epsilon B) = -Q + \epsilon(B^TS+SB)$, and $A + \epsilon B$ is stable for all ϵ such that $V(\epsilon) = -Q + \epsilon(B^TS+SB)$ is negative definite. Since $V(0)$ is negative definite, and the eigenvalues of $V(\epsilon)$ are all real continuous functions of ϵ, we may increase ϵ and still have $V(\epsilon)$ negative definite until $\det V(\epsilon) = 0$. But

(34) $$0 = \det V(\epsilon) = \det(-Q+\epsilon(B^TS+SB))$$

is equivalent to

(35) $$0 = \det(\tfrac{1}{\epsilon} I-Q^{-1}(B^TS+SB)),$$

which implies $\frac{1}{\epsilon}$ is an eigenvalue of $Q^{-1}(B^TS+SB)$. The smallest positive ϵ for which this occurs is $\frac{1}{\epsilon} = \eta_n$ unless $\eta_n \leq 0$, in which case $V(\epsilon)$ is negative definite for all positive ϵ. Likewise $V(\epsilon)$ remains negative definite as we decrease ϵ from zero until $\frac{1}{\epsilon} = \eta_1$, or if $\eta_1 \geq 0$, $V(\epsilon)$ is negative definite for all negative ϵ.

In the preceding three theorems, it is often computationally convenient to let Q be the identity matrix. In this case the conditions of Theorems 2.15 and 2.17 for $A + \epsilon B$ to be stable both reduce to

$$\frac{1}{\nu_1} < \epsilon < \frac{1}{\nu_n} \tag{36}$$

where ν_1 is the smallest and ν_n the largest eigenvalue of $B^TS + SB$. The condition of Theorem 2.16 reduces to

$$\epsilon(\|B^T\|+\|B\|) \leq \frac{1}{\|S\|} . \tag{37}$$

A stronger result, however, will often be obtained by using the criterion implied by the first inequality in (30):

$$\epsilon\|B^TS + SB\| \leq 1. \tag{38}$$

For some types of perturbation matrices B, the eigenvalues of $B^TS + SB$ are particularly easy to find. For example if B has rank one, B^TS and SB have rank one and their sum has rank at most 2. Thus there are at most 2 nonzero eigenvalues of $B^TS + SB$; they are easily found, since their sum is the trace of $B^TS + SB$ and their

product is the sum of all the 2×2 principal minors of $B^TS + SB$.

Assume that only one column of A is perturbed, which without loss of generality we may assume to be the last one, making B a rank one matrix of the form

$$B = [O,b] \tag{39}$$

where O denotes an $n \times (n-1)$ zero matrix and b denotes an n-dimensional column vector. We partition S into

$$S = \begin{bmatrix} M \\ s^T \end{bmatrix} \tag{40}$$

where M is an $n - 1$ by n matrix and s^T is an n-dimensional row vector. Then

$$B^TS + SB = \begin{bmatrix} O & Mb \\ b^TM & 2s^Tb \end{bmatrix} \tag{41}$$

The nonzero eigenvalues satisfy

$$\nu^2 - 2s^Tb\nu - b^TMMb = O \tag{42}$$

or

$$\nu = s^Tb \pm \sqrt{(s^Tb)^2 + b^TMMb} \tag{43}$$

Noting that the quantity under the radical in (43) is just b^TS^2b, and that a similar argument can be carried out for a perturbation of other columns, we have the following theorem by applying equation (36) with $\epsilon = 1$:

Theorem 2.18. Let $A^TS + SA = -I$, let the jth column of A be perturbed by adding a column vector b, and let s_j^T be the jth row of S. The perturbed matrix is stable if

$$s_j^Tb + \sqrt{b^TS^2b} < 1. \tag{44}$$

Equation (43) shows that the quantity $s_j^Tb + \sqrt{b^TS^2b}$ will always be positive.

Now let us assume that only the last row of A is perturbed so that B is the rank one matrix.

$$B = \begin{bmatrix} 0 \\ B^T \end{bmatrix} \qquad b^T = (b_1, b_2, \ldots, b_n) \tag{45}$$

and partition S into

$$S = [M \quad s]$$

where s is the column vector $(s_1, s_2, \ldots, s_n)^T$. Now

$$B^TS + SB = bs^T + sb^T$$

$$= \begin{bmatrix} 2s_1b_1 & s_1b_2+s_2b_1 & \cdots & s_1b_n+s_nb_1 \\ s_2b_1+s_1b_2 & 2s_2b_2 & \cdots & s_2b_n+s_nb_2 \\ & & & \\ s_nb_1+s_1b_n & \cdots & & 2s_nb_n \end{bmatrix} \tag{46}$$

which is again rank two since sb^T and bs^T are both rank one. The two nonzero eigenvalues satisfy

$$\nu^2 - (2\sum_k s_kb_k)\nu - \sum_{k<\ell}(s_kb_\ell - s_\ell b_k)^2 = 0 \tag{47}$$

$$(48) \qquad \nu = \sum_k s_k b_k \pm \sqrt{\left(\sum_k s_k b_k\right)^2 + \sum_{k<\ell} (s_k b_\ell - s_\ell b_k)^2}$$

Again a similar computation can be carried out for other rows of A, and equation (36) applied with $\epsilon = 1$ to yield the companion theorem to the Theorem 2.18:

Theorem 2.19. Let $A^T S + SA = -I$, let the ith row of A be perturbed by adding a row vector $b^T = (b_1 \ldots b_n)$, and let s_i be the ith column of S. The perturbed matrix is stable if

$$(49) \qquad s_i^T b + \sqrt{(s_i^T b)^2 + \sum_{k<\ell} (s_{ki} b_\ell - s_{\ell i} b_k)^2} \; < 1.$$

We note that although equation (44) involves the entire matrix S, equation (49) involves only the ith column of S.

Corollary 2.20. Let $A^T S + SA = -I$. Let the i,jth element of A be perturbed from a_{ij} to $a_{ij} + b_{ij}$. The perturbed matrix is stable if

$$(50) \qquad \frac{1}{s_{ij} - \sqrt{\sum_k s_{ik}^2}} < b_{ij} < \frac{1}{s_{ij} + \sqrt{\sum_k s_{ik}^2}}.$$

Proof: By taking b to be the column vector with b_{ij} in the ith position and zeroes elsewhere in equation (44) (or b^T to be a row vector with b_{ij} in the

jth position and zeroes elsewhere in equation (49)), we get

$$s_{ji}b_{ij} + \sqrt{\sum_k (s_{ki}b_{ij})^2} < 1. \tag{51}$$

Hence

$$b_{ij} < \frac{1}{s_{ji} + \sqrt{\sum_k s_{ki}^2}}, \tag{52}$$

since $s_{ji} + \sqrt{\sum_k s_{ki}^2}$ is positive.

Putting $b_{ij} = -c_{ij}$ in equation (51), we get

$$-s_{ji}c_{ij} + \sqrt{\sum_k s_{ki}^2 c_{ij}^2} < 1 \tag{53}$$

and hence, noting that $s_{ji} - \sqrt{\sum_k s_{ki}^2}$ is negative,

$$b_{ij} = -c_{ij} > \frac{1}{s_{ji} - \sqrt{\sum_k s_{ki}^2}}. \tag{54}$$

Alternately, we could return to the left inequality in equation (36). Equation (50) follows from (52) and (54) since $s_{ki} = s_{ik}$ for every k.

If $A^TS + SA = -Q$, but Q is not the identity, Theorem 2.15 shows that we may replace the 1 in equations (44), (49), and (50) by q, the smallest eigenvalue of Q.

2.4. Convexity of S-Permissible Perturbations.

Definition 2.21. Let $A^TS + SA$ be negative-definite. A matrix B is an S-permissible perturbation of A if $(A+B)^TS + S(A+B)$ is negative definite.

In other words, a matrix B is an S-permissible perturbation of A if $A + B$ is stable by the fundamental lemma, 2.14.

Lemma 2.22. For a given symmetric, positive-definite matrix S, the set of matrices A such that $A^TS + SA$ is negative definite is a convex cone.

Proof: Assume $A^TS + SA$ and $B^TS + SB$ are negative definite. Then for any $\alpha > 0$, $(\alpha A^T)S + S(\alpha A) = \alpha(A^TS + SA)$ is negative definite. If $\alpha,\beta > 0$, $\alpha + \beta = 1$, then $(\alpha A+\beta B)^TS + S(\alpha A+\beta B) = \alpha(A^TS+SA) + \beta(B^TS+SB)$ is negative definite.

Theorem 2.23. Let $A^TS + SA$ be negative definite. The set of S-permissible perturbations of A is an open convex set.

Proof: Let B and C be S-permissible perturbations of A. If $\beta,\gamma > 0$ and $\beta + \gamma = 1$, then

$$A + \beta B + \gamma C = \beta(A+B) + \gamma(A+C). \tag{55}$$

Convexity now follows from Lemma 2.22.

Since $(A+B)^T S + S(A+B)$ is negative definite, and eigenvalues are continuous functions of the entries of the matrix, there is an ϵ such that if M is symmetric and $\|M\| < \epsilon$, $(A+B)^T S + S(A+B) + M$ is negative definite. Let $\delta = \frac{\epsilon}{2\|S\|}$. For any matrix R with $\|R\| < \delta$ and $\|R^T\| < \delta$, $\|R^T S + SR\| < \epsilon$, so that $(A+B+R)^T S + S(A+B+R)$ is negative definite. Thus the set of S-permissible perturbations is open.

This theorem is important because it allows us to answer our second question and compute open (and convex) neighborhoods of the origin which contain only stability preserving perturbations. Thus we can show that stability is preserved when the entries of A are perturbed independently, and not just in a fixed ratio. Equivalently, a matrix can be shown to be stable when its entries are only known to lie in certain intervals:

Corollary 2.24. Let $A^T S + SA = -I$, and let $\beta_{ij} \geq 0$ for every i,j and $\sum_{i,j} \beta_{ij} = 1$. Then any matrix R whose entries satisfy

$$(56) \qquad a_{ij} + \frac{\beta_{ij}}{s_{ij} - \sqrt{\sum_k s_{ik}^2}} < r_{ij} < a_{ij} + \frac{\beta_{ij}}{s_{ij} + \sqrt{\sum_k s_{ik}^2}}$$

for every i,j is stable.

Proof: R can be expressed as A + B, and from (50) B is a convex combination of S-permissible perturbations.

Example 2.25. We illustrate the techniques of sections 2.3 and 2.4 with the matrix A, given by

$$A = \begin{bmatrix} -10.0 & 0 & -7.0 & -2.0 & 0 & 0 & 0 \\ 0 & -9.0 & -1.0 & -5.0 & 0 & 0 & 0 \\ 5.0 & 1.0 & -2.0 & 0 & -10.0 & 0 & 0 \\ 1.0 & 5.0 & 0 & -2.0 & -8.0 & -3.0 & 0 \\ 0 & 0 & 5.0 & 2.0 & -1.0 & 0 & -4.0 \\ 0 & 0 & 0 & -4.0 & 0 & -1.0 & -4.0 \\ 0 & 0 & 0 & 0 & 5.0 & 3.0 & -0.5 \end{bmatrix}.$$

The solution to $A^TS + SA = -I$, rounded to 3 decimal places is

$$S = \begin{bmatrix} .077 & -.004 & .057 & -.015 & -.010 & .003 & .001 \\ -.004 & .088 & -.026 & .064 & -.045 & .031 & .012 \\ .057 & -.026 & .265 & -.053 & .080 & -.023 & -.005 \\ -.015 & .064 & -.053 & .232 & -.067 & .098 & -.012 \\ -.010 & -.045 & .080 & -.067 & .474 & -.032 & .048 \\ .003 & .031 & -.023 & .098 & -.032 & .295 & .030 \\ .001 & .012 & -.005 & -.012 & .048 & .030 & .377 \end{bmatrix}$$

It is not immediately obvious that S is positive definite or that A is stable, but if we take D = diag[0.2,0.2,0.2,0.2,0.5,0.2,0.4] then D + S is diagonally dominant and hence positive-definite and $A^T(D+S) + (D+S)A$ is

$$\begin{bmatrix} -5.0 & 0 & -0.4 & -0.2 & 0 & 0 & 0 \\ 0 & -4.6 & 0 & 0 & 0 & 0 & 0 \\ -0.4 & 0 & -1.8 & 0 & 0.5 & 0 & 0 \\ -0.2 & 0 & 0 & -1.8 & -0.6 & 0.2 & 0 \\ 0 & 0 & 0.5 & -0.6 & -2.0 & 0 & 0 \\ 0 & 0 & 0 & 0.2 & 0 & -1.4 & 0.4 \\ 0 & 0 & 0 & 0 & 0 & 0.4 & -1.4 \end{bmatrix}$$

which is diagonally dominant with negative diagonal entries and hence negative-definite. This implies that A is stable, which implies that S is positive definite.

The upper and lower bounds for S-permissible perturbations b_{ij} computed from equation (50) are given by

$$U = \begin{bmatrix} 5.726 & 11.223 & 6.468 & 12.032 & 11.433 & 9.939 & 10.092 \\ 7.957 & 4.671 & 10.027 & 5.262 & 12.426 & 6.385 & 7.276 \\ 2.887 & 3.796 & 1.805 & 4.232 & 2.704 & 3.748 & 3.516 \\ 3.845 & 2.951 & 4.515 & 1.973 & 4.821 & 2.685 & 3.808 \\ 2.079 & 2.242 & 1.749 & 2.358 & 1.036 & 2.177 & 1.854 \\ 3.163 & 2.881 & 3.406 & 2.415 & 3.519 & 1.636 & 2.892 \\ 2.608 & 2.539 & 2.652 & 2.538 & 2.324 & 2.429 & 1.317 \end{bmatrix}$$

$$L = -\begin{vmatrix} 48.64 & 9.812 & 24.623 & 8.920 & 9.281 & 10.571 & 10.404 \\ 7.697 & 26.853 & 6.596 & 16.313 & 5.852 & 10.556 & 8.780 \\ 4.304 & 3.172 & 40.321 & 2.921 & 4.787 & 3.206 & 3.398 \\ 3.460 & 4.758 & 3.052 & 23.672 & 2.927 & 5.662 & 3.490 \\ 1.994 & 1.864 & 2.434 & 1.791 & 58.621 & 1.911 & 2.257 \\ 3.194 & 3.506 & 2.953 & 4.583 & 2.873 & 47.147 & 3.490 \\ 2.628 & 2.701 & 2.584 & 2.702 & 2.986 & 2.837 & 216.374 \end{vmatrix};$$

Thus the (i,j) element of A may be changed from a_{ij} to $a_{ij} + b_{ij}$ for any $\ell_{ij} < b_{ij} < u_{ij}$ while leaving the other entries fixed, and stability is preserved. For example, A will be stable if $-216.87 \leq a_{77} \leq +.817$ and the other entries are as given.

Using the convexity of S-permissible perturbations, we know that $A + B$ is stable for any B whose entries b_{ij} satisfy

$$\beta_{ij}\ell_{ij} < b_{ij} < \beta_{ij}u_{ij},\ \beta_{ij} > 0,\ \sum_{i,j} \beta_{ij} = 1. \tag{57}$$

For example $A + B$ will be stable for any B, $\frac{1}{49} L \ll B \ll \frac{1}{49} U$. Or by taking $\beta_{67} = \frac{2}{3}$, $\beta_{77} = \frac{1}{3}$, we see that we can add $-2.327 < b_{67} < 1.928$ to a_{67} and $-72.124 < b_{77} < 0.439$ to a_{77} while leaving the other entries unchanged and preserve stability.

By comparison, $\|S\|_\infty = 0.757$, so that by equation (37) $A + B$ is stable if $\|B\|_1 + \|B\|_\infty \leq 1.32$. Equation (28) shows that if $b_{67} = 2$, $b_{77} = 1$, and all other entries of B are zero, $A + \varepsilon B$ is stable for $-2.86 < \varepsilon < .814$.

As another example of how the convexity of the S-permissible perturbation of a matrix can be applied, we prove the following theorem.

Theorem 2.26. If $M = [m_{ij}]$ has negative diagonal entries and the nonzero off diagonal entries satisfy

$$|m_{ij}| < 2\alpha_{ij}|m_{ii}| \tag{58}$$

for some α_{ij} with $\sum_{\substack{i,j \\ i\neq j}} \alpha_{ij} = 1$, then M is a stability matrix.

Proof: Let $A = \operatorname{diag}[m_{ii}]$ and $S = \operatorname{diag}[\frac{1}{-2m_{ii}}]$. Since the m_{ii} are all negative, S is positive definite and $A^TS + SA = -I$. Letting $B = M - A$, M will be stable if B is an S-permissible perturbation.

Let E_{ij} denote the matrix with 1 in position (i,j) and zeros elsewhere. By equation (50), $\frac{m_{ij}}{\alpha_{ij}} E_{ij}$ will be an S-permissible permutation if

$$2m_{ii} = -\frac{1}{s_{ii}} \leq \frac{m_{ij}}{\alpha_{ij}} \leq \frac{1}{s_{ii}} = -2m_{ii}\,, \tag{59}$$

since all off diagonal entries of S are zero. Equation (59) is true for all nonzero off diagonal entries m_{ij} of M by hypothesis. Thus $B = \sum \alpha_{ij}(\frac{m_{ij}}{\alpha_{ij}} E_{ij})$ is a convex combination of S-permissible perturbations and hence is S-permissible itself.

Example 2.27. The matrix

$$M = \begin{bmatrix} -1.0 & 0.2 & 0 \\ 0 & -1.0 & 0.2 \\ 4.4 & 0 & -3.0 \end{bmatrix}$$

is seen to be stable by taking $\alpha_{12} = \alpha_{23} = \frac{1}{8}$ and $\alpha_{31} = \frac{3}{4}$.

2.5. Bordering a Stability Matrix. Another important way to change a matrix is to border the matrix with a new row and column. Thus if we know that an $n \times n$ matrix A is stable, we seek conditions for the matrix $\begin{bmatrix} A & b \\ c^T & d \end{bmatrix}$ to be stable, where b and c are $n \times 1$ column vectors and d is a scalar. We may treat this as a type of perturbation of A and study the stability of the new matrix by the methods of this chapter.

Let $A^TS + SA = -Q$, q be the smallest eigenvalue of Q, and r and k be arbitrary positive scalars. We will choose optimal values for r and k later.

Then

$$\begin{bmatrix} A^T & c \\ b^T & d \end{bmatrix}\begin{bmatrix} rS & 0 \\ 0 & k \end{bmatrix} + \begin{bmatrix} rS & 0 \\ 0 & k \end{bmatrix}\begin{bmatrix} A & b \\ c^T & d \end{bmatrix}$$

$$(60) \quad = \begin{bmatrix} -rQ & (rSb+kc) \\ (rSb+kc)^T & 2kd \end{bmatrix} = - \begin{bmatrix} rQ & 0 \\ 0 & rq \end{bmatrix}$$

$$+ \begin{bmatrix} 0 & rSb+kc \\ (rSb+kc)^T & 2kd+rq \end{bmatrix}.$$

The smallest eigenvalue of $\begin{bmatrix} rQ & 0 \\ 0 & rq \end{bmatrix}$ is rq, so that the left hand side of (60) is negative definite if the largest eigenvalue of $\begin{bmatrix} 0 & rSb+kc \\ (rSb+kc)^T & 2kd+rq \end{bmatrix}$ is less than rq, that is if

$$(61) \quad \frac{2kd + rq + \sqrt{(2kd+rq)^2 + 4(rSb+kc)^T(rSb+kc)}}{2} < rq$$

Isolating the radical, squaring both sides, and collecting terms we have

$$(62) \quad 2kdrq + (rSb+kc)^T(rSb+kc) < 0.$$

Since $k, r,$ and q are all positive, this is equivalent to

$$(63) \quad d < -\frac{1}{2q}\left(\frac{r}{k} b^T S^2 b + 2c^T Sb + \frac{k}{r} c^T c\right).$$

The quantity in parenthesis is positive, and the least restrictive condition on d will be found by choosing k and r to minimize this quantity. We note that it is only the ratio of $\frac{k}{r}$ which matters. The

function $f(x) = a\frac{1}{x} + b + cx$ is readily seen to have a minimum at $x = \sqrt{\frac{a}{c}}$ by taking first and second derivatives. Putting $\frac{k}{r} = \sqrt{\frac{b^T S^2 b}{c^T c}}$ in (63), we have the following theorem:

Theorem 2.28. Let $A^T S + SA = -Q$, S and Q symmetric and positive definite, with q the smallest eigenvalue of Q. Then $\begin{bmatrix} A & b \\ c^T & d \end{bmatrix}$ is stable if

$$d < -\frac{1}{q}\left(\sqrt{c^T c b^T S^2 b} + c^T S b\right). \tag{64}$$

We note that the right hand side of (64) is always less than or equal to zero, being zero when $c = -\alpha Sb$, for some scalar α. Equation (64) is clearly only a sufficient condition for $\begin{bmatrix} A & b \\ c^T & d \end{bmatrix}$ to be stable. Examples which are stable but have $d > 0$ are easy to construct.

It is, however, the best result that can be obtained by using a matrix of the form $\begin{bmatrix} rS & 0 \\ 0 & k \end{bmatrix}$ in equation (60). A better result might be possible using $\begin{bmatrix} rS & x \\ x^T & k \end{bmatrix}$, but this makes the right hand side of (60) so complicated that bounds on the eigenvalues are difficult to find.

If we wish to find a better (possibly positive) upper bound for d that guarantees $\begin{bmatrix} A & b \\ c^T & d \end{bmatrix}$ to be stable,

could choose a d_o satisfying (64), solve the apunov equation for

$$\begin{bmatrix} A & b \\ c^T & d_o \end{bmatrix} S_1 + S_1 \begin{bmatrix} A & b \\ c^T & d_o \end{bmatrix} = \begin{bmatrix} -Q & 0 \\ 0 & -1 \end{bmatrix}$$

using $\begin{bmatrix} S & 0 \\ 0 & \frac{1}{2d_o} \end{bmatrix}$ as an initial guess in the iterative methods discussed in the next chapter, and compute how far d_o may be perturbed while preserving stability by equation (50).

Example 2.29. Let A be the same as in example 2.25, S the solution to $A^TS + SA = -I$, $b^T = [0,0,0,0,-1,-3,0]$, and $c^T = [0,0,0,0,2,4,0]$. Equation (64) becomes

$$d < -(\sqrt{c^Tc}\ \sqrt{b^TS^2b} + c^TSb) = -(4.30 \times 0.96 - 3.31) = -0.99.$$

If we let $b^T = [0,0,0,0,-3,-1,0]$ and $c^T = [0,0,0,0,+4,+2,0]$; then $d < -(4.47 \times 1.47 - 6.34) = -0.24$, is a sufficient condition for the bordered matrix $\begin{bmatrix} A & b \\ c^T & d \end{bmatrix}$ to be stable.

In this latter case, taking $d = -0.24$, computing the solution to the Lyapunov equation for the bordered matrix, and applying equation (50) shows that the bordered matrix is stable for $d < 0.119$.

CHAPTER III

ITERATIVE SOLUTION OF THE LYAPUNOV EQUATION

The methods of the previous chapter depend on solving the Lyapunov matrix equation

$$A^T S + SA = -Q. \tag{1}$$

As noted in Section 2.1, the operator $\bar{A}$ on the space of $n \times n$ matrices defined by

$$\bar{A}S = A^T S + SA \tag{2}$$

is a linear operator. The major difficulty in solving equation (1) is the size of the system. Since S, Q and $\bar{A}S$ are $n \times n$ matrices, there are n^2 equations in n^2 unknowns, and even using the symmetry of Q and S there are $\frac{1}{2}n(n+1)$ equations in $\frac{1}{2}n(n+1)$ unknowns. To solve such a system by Gaussian elimination would take $\frac{1}{3}(\frac{1}{2}n(n+1))^3 + (\frac{1}{2}n(n+1))^2 + \frac{2}{3}(\frac{1}{2}n(n+1)) = \frac{1}{24}n^6 + \frac{1}{8}n^5 + \frac{3}{8}n^4 + \frac{13}{24}n^3 + \frac{4}{3}n^2 + \frac{n}{3}$ multiplications and divisions. (See Westlake [42], p.100.)

Several authors have given methods to solve equation (1) based on the characteristic polynomial of A. (See Brickley and McNamee [6], Frame [8], Jameson [19].) But the best methods to find characteristic

equations require n matrix multiplications, and then these methods require another n matrix multiplications to get S after finding the characteristic polynomial, for a total of $2n^4$ multiplications. There are other methods (see Smith [35], and Kalman and Bertram [20]) which require first transforming A to a special form, usually a difficult and potentially an unstable procedure. A survey of several of these methods is found in Barnett and Storey [4], Chapter VI.

Iterative procedures are often used to solve large systems of linear equations. One interative method to solve (1) is given by Barnett and Storey [4], but they report that it converges too slowly to be useful in many cases. This chapter adapts some of the well known iterative procedures for solving ordinary systems of linear equations to the solution of the matrix equation (1) and evaluates their effectiveness for this equation. Because of the nature of the matrices in the Lotka-Volterra system, we will be especially interested in the effectiveness of these procedures for matrices A with large skew symmetric components.

3.1. General Iterative Procedures. Equation (1) may be written in the operator notation as

$$\bar{A}S + Q = 0. \tag{3}$$

Let $\bar{N}$ be an operator which is easy to invert, and w any real number. Then (3) is equivalent to

$$\frac{1}{w}\bar{N}S = \frac{1}{w}\bar{N}S + \bar{A}S + Q. \tag{4}$$

$$S = S + w\bar{N}^{-1}(\bar{A}S + Q). \tag{5}$$

$$S = (I_{n^2} + w\bar{N}^{-1}\bar{A})S + w\bar{N}^{-1}Q. \tag{6}$$

The choice of the number w, called the overrelaxation parameter, will be discussed later. An alternate formulation is to let $\bar{A} = \bar{P} - \bar{N}$, so that

$$\frac{1}{w}\bar{N}S = (\frac{1}{w} - 1)\bar{N}S + \bar{P}S + Q \tag{4a}$$

$$S = (1-w)S + w\bar{N}^{-1}(\bar{P}S+Q) \tag{5a}$$

$$S = ((1-w)I_{n^2} + w\bar{N}^{-1}\bar{P})S + w\bar{N}^{-1}Q \tag{6a}$$

We will find both forms useful. The iterative scheme derived from (5),

$$S^{(i+1)} = S^{(i)} + w\bar{N}^{-1}(\bar{A}S^{(i)}+Q) \tag{5b}$$

converges to the solution of (3), for any initial choice $S^{(0)}$, if and only if the eigenvalues $\{u\}$ of $(I+w\bar{N}^{-1}\bar{A})$ satisfy $|u| < 1$. But

$$\{u\} = \{1 + w\lambda : \lambda \text{ is an eigenvalue of } \bar{N}^{-1}\bar{A}\} \tag{6}$$

so that (4b) converges if and only if the eigevalues λ of $\bar{N}^{-1}\bar{A}$ satisfy

$$|1 + w\lambda| < 1. \tag{7}$$

Lemma 3.1. Let $\bar{N}$ and $\bar{A}$ be any linear operators. Then the eigenvalues of $\bar{N}^{-1}\bar{A}$ have negative real part if and only if there is a positive real number w such that equation (5b) converges to the solution of (1) for any initial matrix $S^{(0)}$.

Proof: Let $\{\lambda_k = x_k+iy_k\}$ be the set of eigenvalues of $\bar{N}^{-1}\bar{A}$, and choose $0 < w < \min_k \frac{-2x_k}{x_k^2+y_k^2}$. This is possible when all the x_k are negative. For every k,

$$|1 + w\lambda_k|^2 = 1 + 2x_k w + (x_k^2+y_k^2)w^2 \tag{8}$$

is a convex quadratic function of w, which has value 1 for $w = 0$ and $w = \frac{-2x_k}{x_k^2+y_k^2}$. Therefore $|1 + w\lambda_k| < 1$ for

$$0 < w < \min_k \frac{-2x_k}{x_k^2+y_k^2} < \frac{-2x_k}{x_k^2+y_k^2}. \tag{9}$$

On the other hand, suppose for some $\lambda_k = x_k + iy_k$, $x_k \geq 0$. Then $1 + w\lambda_k = 1 + wx_k + iy_k$ is an eigenvalue of $I_{n^2} + w\bar{N}^{-1}\bar{A}$ and has absolute value ≥ 1 for any choice of $w > 0$.

3.2. Rate of Convergence. For the convergence scheme (5b), $\rho(w) = \sigma(I_{n^2}+w\bar{N}^{-1}\bar{A})$ is approximately the factor by which the norm of the error is reduced with each iteration. If $M = I_{n^2} + w\bar{N}^{-1}\bar{A}$, the rate of convergence is usually defined as $|\log \sigma(M)| = -\log \sigma(M)$. (See Isaacson and Keller [18].)

Let $\{\lambda_k\}$ be the set of eigenvalues of the operator $\bar{N}^{-1}\bar{A}$. To get the best rate of convergence, we want to choose w to minimize $\sigma(I+w\bar{N}^{-1}\bar{A})$, which is equivalent to choosing w to minimize

$$\rho^2(w) = \max_k |1 + w\lambda_k|^2 = \max_k (1+2x_k w+(x_k^2+y_k^2)w^2). \tag{10}$$

For a fixed k,

$$\min_w |1 + w\lambda_k|^2 = 1 - \frac{x_k^2}{x_k^2+y_k^2} = \frac{y_k^2}{x_k^2+y_k^2} \tag{11}$$

and this minimum occurs for $w = \dfrac{-x_k}{x_k^2+y_k^2}$. Since $\min_w \rho^2(w) > \min_w |1 + w\lambda_k|^2$, convergence will be slow for even the best choice of w if there is an eigenvalue $\lambda_k = x_k + iy_k$ with y_k much larger than x_k.

If all the eigenvalues of $\bar{N}^{-1}\bar{A}$ are real with $\lambda_n < \ldots < \lambda_2 < \lambda_1 < 0$, then Figure 1 shows that the best choice for w is where $|1 + w\lambda_n|^2 = |1 + w\lambda_1|^2$, that is for $w = -\dfrac{2}{\lambda_1+\lambda_n}$. (An algebraic proof is in Isaacson and Keller [18].)

For this choice of w, we have

$$\rho(w) = \left(1+2\lambda_1\left(\frac{-2}{\lambda_1+\lambda_n}\right)+\lambda_1^2\left(\frac{4}{(\lambda_1+\lambda_n)^2}\right)\right)^{\frac{1}{2}} = \left|\frac{\lambda_1-\lambda_n}{\lambda_1+\lambda_n}\right|. \tag{12}$$

Convergence will be slow if $|\lambda_1|$ is very small compared to $|\lambda_n|$, that is if the condition number $\left|\dfrac{\lambda_n}{\lambda_1}\right|$ of $\bar{N}^{-1}\bar{A}$, is large.

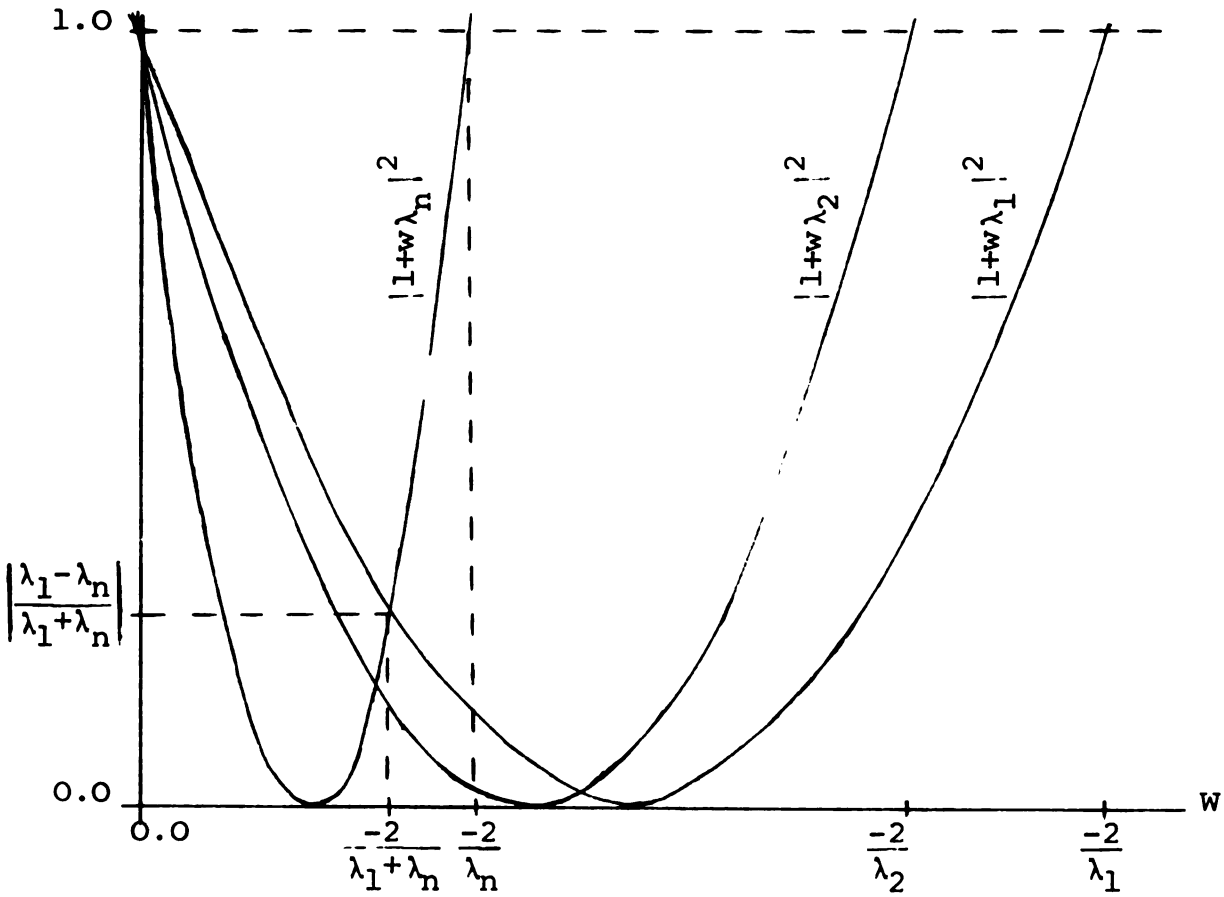

Figure 1. Graph of $|1 + w\lambda_i|^2$ versus w, where $\lambda_n < \ldots < \lambda_2 < \lambda_1 < 0$ are real eigenvalues of $\bar{N}^{-1}\bar{A}$. Convergence occurs for w which makes $|1 + w\lambda_i|^2 < 1$ for all i, that is for $0 < w < \frac{-2}{\lambda_n}$. Convergence is most rapid for that w which minimizes $\max_i |1 + w\lambda_i|^2$, that is for $w = \frac{-2}{\lambda_1+\lambda_2}$.

In practice the determination of the best choice for w is complicated not only by complex eigenvalues, but also by a lack of knowledge of the exact location of the eigenvalues.

3.3. Practical Iterative Procedures. In practice, equation (5b) is used in the form

$$P^{(i)} = A^T S^{(i)} + S^{(i)} A + Q \tag{13a}$$

$$N^T V^{(i)} + V^{(i)} N = P^{(i)} \tag{13b}$$

$$S^{(i+1)} = S^{(i)} + wV^{(i)}, \tag{13c}$$

where N is a matrix for which it is easy to solve equation (13b) for $V^{(i)}$. Since we know that $\bar{A} = A^T \times I + I \times A^T$ and $\bar{N} = N^T \times I + I \times N^T$ have eigenvalues that are sums of pairs of eigenvalues of A and N respectively, we can find criteria for convergence in terms of the eigenvalues or other properties of the matrices A and N. The following theorem is the most general in this direction.

Theorem 3.2. A necessary condition for the matrix iterative scheme,

$$N^T S^{(i+1)} + S^{(i+1)} N = N^T S^{(i)} + S^{(i)} N + w(A^T S^{(i)} + S^{(i)} A + Q) \tag{14}$$

to converge for all initial values $S^{(0)}$, is that the vector iterative scheme

$$Nx^{(i+1)} = Nx^{(i)} + wAx^{(i)} + wq \tag{15}$$

converges for all initial values $x^{(0)}$ and the rate of convergence of the matrix scheme can be no faster than that of the vector scheme.

Proof: In operator form, (14) is $S^{(i+1)} = (I_{n^2}+w\bar{N}^{-1}\bar{A})S^{(i)} + w\bar{N}^{-1}Q$ and (15) is equivalent to $x^{(i+1)} = (I+wN^{-1}A)x^{(i)} + wN^{-1}q$. Let μ be the eigenvalue of $(I+wN^{-1}A)$ with largest absolute value. Then (15) converges if and only if $|\mu| < 1$ and the rate of convergence is $-\log|\mu|$.

Since μ is an eigenvalue of $I + wN^{-1}A$, $\det[\mu I - (I+wN^{-1}A)] = 0$, which implies $\det[(\mu-1)N - wA] = 0$, so that 0 is an eigenvalue of $(\mu-1)N - wA$.

The eigenvalues of the operator $\overline{(\mu-1)N - wA} = (\mu-1)\bar{N} - w\bar{A}$ are the sums of all pairs of eigenvalues of $(\mu-1)N - wA$. In particular $0 = 0 + 0$ will be an eigenvalue. This implies $\det[(\mu-1)\bar{N} - w\bar{A}] = 0$, which implies $\det[\mu I_{n^2} - (I_{n^2}+w\bar{N}^{-1}\bar{A})] = 0$. Thus μ is an eigenvalue of $I_{n^2} + w\bar{N}^{-1}\bar{A}$, and $\rho = \sigma[I_{n^2} + w\bar{N}^{-1}\bar{A}] \geq |\mu|$. But (14) converges only if $1 \geq \rho \geq |\mu|$, and the rate of convergence is $-\log\rho \leq -\log|\mu|$. This completes the proof.

Convergence of (15), however, is not sufficient for the convergence of (14), as seen in the following example:

Example 3.3. Let $A = \begin{bmatrix} -1 & 4 \\ -1 & 3 \end{bmatrix}$ and $N = \begin{bmatrix} 1 & 0 \\ 0 & 4 \end{bmatrix}$ then $N^{-1}A = \begin{bmatrix} -1 & 4 \\ -\frac{1}{4} & \frac{3}{4} \end{bmatrix}$, which has eigenvalues $-\frac{1}{8} \pm i\sqrt{15}/8$. Since they have negative real part, by Lemma 3.1 there is a positive real number w such that (15) converges. But

$$\bar{A} = A^T \times I + I \times A^T = \begin{bmatrix} -2 & -1 & -1 & 0 \\ 4 & 2 & 0 & -1 \\ 4 & 0 & 2 & -1 \\ 0 & 4 & 4 & 6 \end{bmatrix}$$

$$\text{and } \bar{N} = N^T \times I + I \times N^T = \begin{bmatrix} 2 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 8 \end{bmatrix}$$

$$\text{so that } \bar{N}^{-1}\bar{A} = \begin{bmatrix} -1 & -\frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{4}{5} & \frac{2}{5} & 0 & -\frac{1}{5} \\ \frac{4}{5} & 0 & \frac{2}{5} & -\frac{1}{5} \\ 0 & \frac{1}{2} & \frac{1}{2} & \frac{3}{4} \end{bmatrix}$$

which has eigenvalues $-\frac{1}{8} \pm i\sqrt{15}/8$, as predicted by the proof of Theorem 3.2, and a double eigenvalue $2/5$. Since there are eigenvalues with both positive and negative real part, again by Lemma 3.1 there is no positive real w which will make (14) converge.

Even if we restrict ourselves to operating only on symmetric matrices S, (14) will not converge. If $S = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix}$, $(\bar{N}^{-1}\bar{A}(S)_V$ is given by

$$\begin{bmatrix} -1 & -\frac{1}{2} & -\frac{1}{2} & 0 \\ \frac{4}{5} & \frac{2}{5} & 0 & -\frac{1}{5} \\ \frac{4}{5} & 0 & \frac{2}{5} & -\frac{1}{5} \\ 0 & \frac{1}{2} & \frac{1}{2} & \frac{3}{4} \end{bmatrix} \begin{bmatrix} s_{11} \\ s_{12} \\ s_{21} \\ s_{22} \end{bmatrix} .$$

Since S is symmetric, we can reduce the dimensions to a 3×3 system by identifying s_{12} and s_{21}. This requires us to add columns 2 and 3 and delete row 2 or row 3 in the matrix $\bar{N}^{-1}\bar{A}$, which then collapses to

$$\begin{bmatrix} -1 & -1 & 0 \\ \frac{4}{5} & \frac{2}{5} & -\frac{1}{5} \\ 0 & 1 & \frac{3}{4} \end{bmatrix} .$$

This matrix also has eigenvalues $-\frac{1}{8} \pm i\sqrt{15}/8$ and $\frac{2}{5}$. In other words, there are eigenvalues of the operator $\bar{N}^{-1}\bar{A}$ with both positive and negative real part that correspond to symmetric "eigenvectors".

An important consideration for using any iterative procedure is a good initial guess for the unknown. One possibility is to take $S^{(0)} = \bar{N}^{-1}Q$, another is to let $S^{(0)}$ be the identity matrix.

A convenient choice for Q is the identity matrix. If A is normal, $A = M + K$ where $M = \frac{1}{2}(A^T+A)$ and $K = \frac{1}{2}(-A^T+A)$ and M^{-1} and K commute. Then

$$(16) \qquad A^T(-\tfrac{1}{2} M^{-1}) + (-\tfrac{1}{2} M^{-1})A = -\tfrac{1}{2}[MM^{-1} - KM^{-1} + M^{-1}M + M^{-1}K] = -I.$$

Thus $-(A^T+A)^{-1} = -\frac{1}{2} M^{-1}$ can be used as an initial guess for S when A is close to normal and $Q = I$. Our experience shows, however, that when A is close to the form $-D + K$ with D diagonal and K skew symmetric, taking $S^{(0)} = \frac{1}{2} D^{-1}$ is easier and works as well.

As mentioned, to actually use the iterative scheme (5b) (or equivalently (13)) we must find matrices N for which $\bar{N}$ is easy to invert, that is for which it is easy to solve (13b) for $V^{(i)}$. Since $\bar{N} = N^T \times I + I \times N^T$ is diagonal or triangular when N is diagonal or triangular, respectively, these are the most natural choices for N.

3.4. Simple Iteration. $\overline{\frac{1}{2} I}$ is the identity operator on $n \times n$ matrices, since $\frac{1}{2} IS + S(\frac{1}{2} I) = S$. Choosing $N = \frac{1}{2} I$ gives the simple iterative scheme

$$(17) \qquad S^{(i+1)} = S^{(i)} + w(A^T S^{(i)} + S^{(i)} A + Q).$$

Criteria for convergence is given by the following corollary to Lemma 3.1.

Corollary 3.4. Let A be a stability matrix. Then there is a positive real number w such that the

iterative scheme (17) converges to the solution of $A^TS + SA = -Q$, for any initial guess $S^{(0)}$.

Proof: $N = \frac{1}{2} I$, $\bar{N}$ is the identity operator, $\bar{N}^{-1}\bar{A}$ is just $\bar{A}$, and (5b) reduces to the iterative scheme (17). But the eigenvalues of $\bar{A}$ are sums of pairs of eigenvalues of A, and hence all have negative real part. The conclusion follows from Lemma 3.1.

If the ratio of the largest real part to the smallest real part of the eigenvalues of A is large, or if A has an eigenvalue with large imaginary part, the same will be true of $\bar{A}$. In Section 3.2 we saw that the rate of convergence of (17) for such a matrix will be slow. Since each iteration of (17) takes n^3 multiplications, if it takes over $2n$ iterations to achieve the desired accuracy, the characteristic polynomial methods will be faster.

3.5. Jacobi Iteration. If $N = \mathrm{diag}[d_1,\ldots,d_n]$, then equation (13b) componentwise is

$$d_i v_{ij} + v_{ij} d_j = p_{ij} \tag{18}$$

$$v_{ij} = \frac{p_{ij}}{d_i + d_j}$$

When we use the negatives of the diagonal entries of A for N in equations (13), the result is Jacobi iteration with overrelaxation applied to the operator $\bar{A}$. Componentwise this can be written as

$$(19)\qquad s_{ij}^{(\ell+1)} = (1-w)s_{ij}^{(\ell)} - w\Big(\sum_{\substack{k=1\\k\neq i}}^{n} a_{ki}s_{kj}^{(\ell)} + \sum_{\substack{k=1\\k\neq j}}^{n} s_{ik}^{(\ell)}a_{kj} + q_{ij}\Big)/(a_{ii}+a_{jj})$$

where $s_{ij}^{(\ell)}$ is the value of the (i,j)-entry of S after the ℓth iteration. This can be done with $n^3 + n^2$ multiplications per iteration.

A sufficient condition for the Jacobi method with $w = 1$ to converge for ordinary vector equations $Ax = y$ is that A be diagonally dominant. This is also a sufficient condition for the convergence of (19), since $A^T \times I + I \times A^T$ is diagonally dominant whenever A is.

For matrices which are exactly or nearly of the form $A = -D + K$ with D diagonal and K skew symmetric, which we expect to encounter often in a Lotka-Volterra system, the following theorem is useful:

Theorem 3.5. Let $A = -D + K$, K skew symmetric, $D = \mathrm{diag}[d_1,\ldots,d_n]$, with $d_i > 0$ $\forall i$. Then there is a real $w > 0$, such that

$$(20)\qquad S^{(i+1)} = S^{(i)} + w\bar{D}^{-1}\bar{A}S^{(i)} + w\bar{D}^{-1}Q$$

converges to the solution S of $A^TS + SA = -Q$. The eigenvalues of $\bar{D}^{-1}\bar{K}$ are pure imaginary. Let them be denoted by $\pm iy_1,\ldots,\pm iy_n$ with $0 \leq y_1 \leq y_2 \leq \ldots \leq y_n$. The

fastest rate of convergence is $-\frac{1}{2}\log\frac{y_n^2}{1+y_n^2}$, which is obtained by taking $w = \frac{1}{1+y_n^2}$.

Proof: $\bar{A} = -\bar{D} + \bar{K}$ and $\bar{D} = D \times I + I \times D$ is diagonal and $\bar{K} = K^T \times I + I \times K^T$ is skew symmetric. $\bar{D}$ has positive diagonal entries, so that $(\bar{D})^{1/2}$ exists. $\bar{D}^{1/2}(\bar{D}^{-1}\bar{K})\bar{D}^{-1/2} = \bar{D}^{-1/2}\bar{K}\bar{D}^{-1/2}$, which is skew symmetric. It follows that $\bar{D}^{-1}\bar{K}$, being similar to a skew symmetric matrix, has pure imaginary eigenvalues, which we denote by $\pm iy_1,\ldots,\pm iy_n$. $\bar{D}^{-1}\bar{A} = -I + \bar{D}^{-1}\bar{K}$, has eigenvalues $-1 \pm iy_1,\ldots,-1 \pm iy_n$. It now follows from Lemma 3.1 that there is a real $w > 0$ which makes (20) convergent.

The eigenvalues of the operator $I + w\bar{D}^{-1}\bar{A}$ have the form $1 - w \pm iwy_k$. Since for any fixed w, $|1 - w \pm iwy|^2 = (1+y^2)w^2 - 2w + 1$ is strictly increasing as y increases, $\max_k |1 - w \pm iwy_k|^2 = |1 - w \pm iwy_n|^2$, and the best rate of convergence is obtained for that w which minimizes $(1+y_n^2)w^2 - 2w + 1$, that is for $w = \frac{1}{1+y_n^2}$. (See Figure 2) For this w, $\sigma|I + w\bar{D}^{-1}\bar{A}| = |1 - w \pm iwy_n| = \sqrt{\frac{y_n^2}{1+y_n^2}}$, which makes the rate of convergence $\left|\log\sqrt{\frac{y_n^2}{1+y_n^2}}\right| = -\frac{1}{2}\log\frac{y_n^2}{1+y_n^2}$.

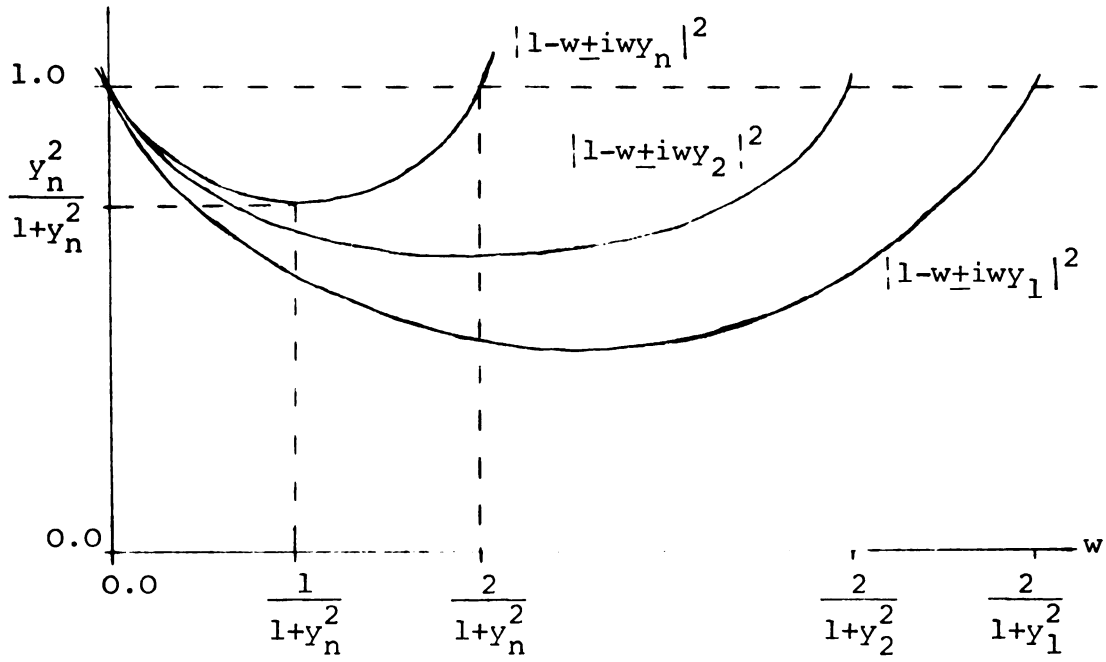

Figure 2. Graph of $|1 - w \pm iwy_n|^2$ versus w, where $\pm iy_n$ are the eigenvalues of $\bar{D}^{-1}\bar{K}$ with $0 \leq y_1 \leq y_2 \leq \ldots \leq y_n$. Convergence occurs if $|1 - w \pm iwy_n|^2 < 1$ for all n, that is if $w < \frac{2}{1+y_n^2}$. Best rate of convergence is for w which minimizes $\max_n |1 - w \pm iwy_n|^2$, that is for $w = \frac{1}{1+y_n^2}$.

Although this theorem guarantees convergence, it also shows that if the eigenvalues of $\bar{D}^{-1}\bar{K}$ are much larger than 1 in absolute values, convergence will be too slow to be practical.

3.6. Seidel Iteration. A third possibility is to take N to be the upper (or lower) triangular part of A. Equation (13b) is then equivalent to the system of equations

$$(21)\qquad \sum_{k=1}^{i} a_{ki}v_{kj} + \sum_{k=1}^{j} v_{ik}a_{kj} = p_{ij} \qquad i = 1...n,\ j = i...n,$$

which can be readily solved by taking the equations in the order $i = 1,\ j = 1...n;\ i = 2,\ j = 2...n;$ up to $i = n,\ j = n,$ and utilizing the fact that $v_{kj} = v_{jk}$.

Using this choice for N in equation (5) and (13) is equivalent to Seidel Iteration applied to the operator $\bar{A}$. Componentwise this can be written analogously to equation (19):

$$(22)\qquad s_{ij}^{(\ell+1)} = (1-w)s_{ij}^{(\ell)} - w\Big(\sum_{k=1}^{i-1} a_{ki}s_{kj}^{(\ell+1)} + \sum_{k=i+1}^{n} a_{ki}s_{kj}^{(\ell)} + \sum_{k=1}^{j-1} s_{ik}^{(\ell+1)}a_{kj} + \sum_{k=j+1}^{n} s_{ik}^{(\ell)}a_{kj} + q_{ij}\Big)/(a_{ii}+a_{jj})$$

$$i = 1...n,\quad j = i...n.$$

Again this can be done with $n^3 + n^2$ multiplications per iteration. It is usually more efficient than the Jacobi

method and is actually easier to program, since the old entries are replaced by the new entries as soon as they are computed, instead of requiring separate storage. Like the Jacobi method, the Seidel method with $w = 1$ will converge if A is diagonally dominant. The vector Seidel method is also known to converge when A is symmetric positive (negative) definite (see Fox [7], p.193), and A symmetric positive (negative) definite implies $\bar{A}$ is symmetric positive (negative) definite, so that this result also carries over to the matrix iterative procedure of (22). If $A = (L + D - L^T)$ where L is strictly lower triangular, D is diagonal, and $L + D + L^T$ is positive (negative) definite, the Seidel method can be shown to converge by mimicking Fox's proof.

3.7. Block Seidel Method. Iterative schemes from equation (5b) work best when $\bar{N}^{-1}$ is approximately $\bar{A}^{-1}$, but this is not the case for the Jacobi or Seidel schemes when A has large off diagonal entries and especially when the off diagonal part is nearly skew symmetric. In the latter case these schemes try to approximate the inverse of a matrix having eigenvalues with large imaginary components by inverting a matrix with real eigenvalues. The block Seidel method described below allows us to pull some of these large off diagonal entries into the part which is inverted, giving much more rapid convergence.

Let the $n \times n$ matrices A, S, and Q be partitioned into $m \times m$ blocks,

$$A = \begin{bmatrix} A_{11} & A_{12} & \cdots & A_{1r} \\ A_{21} & A_{22} & \cdots & A_{2r} \\ & & & \vdots \\ A_{r1} & A_{r2} & \cdots & A_{rr} \end{bmatrix} \qquad S = \begin{bmatrix} S_{11} & S_{12} & \cdots & S_{1r} \\ S_{21} & S_{22} & \cdots & S_{2r} \\ & & & \vdots \\ S_{r1} & S_{r2} & \cdots & S_{rr} \end{bmatrix}$$

(23)

$$Q = \begin{bmatrix} Q_{11} & Q_{12} & \cdots & Q_{1r} \\ Q_{21} & Q_{22} & \cdots & Q_{2r} \\ & & & \vdots \\ Q_{r1} & Q_{r2} & \cdots & Q_{rr} \end{bmatrix}$$

where $r = \frac{n}{m}$. For simplicity we assume that m divides n. It should be noted that since S and Q are symmetric, S_{ii}, Q_{ii} are symmetric and $S_{ji} = S_{ij}^T$, $Q_{ji} = Q_{ij}^T$. Equation (1) written out block by block becomes

$$(24) \qquad \left(\sum_{k=1}^{r} S_{jk}A_{ki}\right)^T + \sum_{k=1}^{r} S_{ik}A_{kj} = -Q_{ij}$$

or

$$(25) \qquad A_{ii}^T S_{ij} + S_{ij}A_{jj} = -\sum_{\substack{k=1 \\ k\neq i}}^{r} A_{ki}^T S_{kj} - \sum_{\substack{k=1 \\ k\neq j}}^{r} S_{ik}A_{kj} - Q_{ij}.$$

If m is small, equation (25) can be solved for the m^2 entries in block S_{ij} directly, by simply solving the m^2 vector equation

$$(26) \qquad (A_{ii}^T \times I + I \times A_{jj}^T)(S_{ij})_V = Y_V$$

where Y is the right hand side of (25). This allows us to do a block Seidel iteration with overrelaxation completely analogous to equation (22) with single entries replaced by matrix blocks:

$$(27a) \quad A_{ii}^T V_{ij} + V_{ij} A_{jj}^T = \sum_{k=1}^{i-1} A_{ki}^T S_{kj}^{(\ell+1)} + \sum_{k=i+1}^{r} A_{ki}^T S_{kj}^{(\ell)} + \sum_{k=1}^{j-1} S_{ik}^{(\ell+1)} A_{kj} + \sum_{k=j+1}^{r} S_{ik}^{(\ell)} A_{kj} + Q_{ij}$$

$$(27b) \qquad S_{ij}^{(\ell+1)} = (1-w) S_{ij}^{(\ell)} - w V_{ij}.$$

The equations are solved in the order $i = 1$, $j = 1...r$; $i = 2$, $j = 2...r$;...; $i = r$, $j = r$; and each time a block S_{ij} is found, S_{ji} is set equal to S_{ij}^T if $i \neq j$.

To obtain the maximum benefit from this procedure it is best to first permute the rows and columns of A to bring the largest possible off diagonal elements into the diagonal blocks, so that instead of solving equation (1) directly one solves

$$(28) \qquad P^T A^T P (P^T S P) + (P^T S P) P^T A P = -P^T Q P$$

for $P^T S P$, where P is a permutation matrix. On the computer this can be done by leaving A, S, and Q fixed but permuting the indices of the rows and columns.

This makes possible the option of using a sequence of permutations to bring different off diagonal elements into the diagonal blocks on successive iterations.

The number of multiplications for each iteration of the block Seidel method depends on the block size m. The right hand side of equation (27a) takes $2(\frac{n}{m} - 1)$ multiplications, and solving for the m^2 unknowns in V_{ij} by Gaussian elimination takes $\frac{1}{3} m^6 + m^4 + \frac{2}{3} m^2$ multiplications and divisions. Equation (27b) takes an additional m^2 multiplications. This must be done for each of the $\frac{1}{2} \frac{n}{m}(\frac{n}{m} + 1)$ blocks, for a total of $n^3 + \frac{1}{2}(\frac{1}{3} m^4 + m^2 + \frac{5}{3})n^2 + (\frac{1}{6} m^5 + \frac{1}{2} m^3 - m^2 + \frac{5}{6})n$ multiplications. For $m = 2$, this is $n^3 + \frac{11}{2} n^2 + \frac{19}{2} n$. This number could be reduced somewhat by taking advantage of the existing zeroes in the matrices $A_{ii}^T \times I + I \times A_{jj}^T$. At the cost of storing an additional $\frac{1}{2}(m^2n^2 + m^3n)$ real numbers, the matrices $A_{ii}^T \times I + I \times A_{jj}$ could be stored in factored form after the first iteration, so that later iterations would require only $m^4 + m^2$ multiplications and divisions to solve for V_{ij}.

In spite of the extra work involved, the examples which follow show that convergence is enough faster for the block Seidel method than for the Seidel or Jacobi method to make it the preferred method. As n gets

larger the number of additional operations becomes less significant, and for smaller matrices it would be better to solve them directly or use a characteristic polynomial method than an iterative method.

3.8. Examples and Comparisons. The methods described were tested on three different matrices and the results are given in the tables below. In each case the best overrelaxation parameter w was found somewhat experimentally and the result for the best w is given. It was observed that as w is increased the rate of convergence increases to a certain point, after which it rapidly decreases until the procedure becomes divergent.

In the examples below D stands for the negative of the diagonal of A, w is the overrelaxation parameter, and $\|R\|$ is the square root of the sum of squares of the entries of the residual $A^TS + SA + Q$. In each case the procedure was run for 20 iterations unless the norm squared of the correction matrix $\bar{N}^{-1}(A^TS+SA+Q)$ became less than 10^{-10} before that.

Since a matrix A of odd dimension cannot be divided evenly into 2×2 blocks, the block Seidel method was actually performed on the bordered matrix $\begin{bmatrix} A & 0 \\ 0 & -1 \end{bmatrix}$.

Example 3.6.

$$A = \frac{1}{10}\begin{bmatrix} -5 & 1 & 2 & 0 & 1 \\ -1 & -4 & 7 & -1 & 0 \\ -2 & -7 & -6 & 0 & 0 \\ 0 & 1 & 0 & -6 & -4 \\ -1 & 0 & 0 & 4 & -5 \end{bmatrix}$$

	Q	$S^{(0)}$	w	no. of iterations	$\|R\|$
Simple iteration	$\frac{1}{10}$ diag[5,5,12,9,10]	I	.584	20	8.28×10^{-2}
Jacobi	I	$\frac{1}{2}D^{-1}$	.333	20	1.12×10^{-2}
Seidel	I	$\frac{1}{2}D^{-1}$	.300	20	2.87×10^{-4}
Block Seidel*	I	$\frac{1}{2}D^{-1}$	1.0	6	7.67×10^{-8}

*Two by two blocks were used.

Example 3.7.

$$A = \begin{bmatrix} -5 & 2 & 0 & 1 & 5 \\ -1 & -4 & 0 & 2 & 4 \\ 0 & 0 & -3 & -3 & 0 \\ 1 & -2 & 2 & -2 & 0 \\ -5 & -6 & 0 & 0 & -1 \end{bmatrix}$$

	Q	$S^{(0)}$	w	no. of iterations	$\|R\|$
Simple iteration	D	I	0.1	diverges	
Jacobi	I	I	0.125	20	1.32
	D	I	0.125	40	6.32×10^{-1}
Seidel	I	$\frac{1}{2} D^{-1}$	0.15	20	1.76×10^{-1}
Block Seidel*	I	$\frac{1}{2} D^{-1}$	0.8	10	1.41×10^{-5}

*Two by two blocks were used and the rows and columns were first permuted to come in the order (1,5,3,4,2).

Example 3.8.

$$A = \begin{bmatrix} -10 & 0 & -7 & -2 & 0 & 0 & 0 \\ 0 & -9 & -1 & -5 & 0 & 0 & 0 \\ 5 & 1 & -2 & 0 & -10 & 0 & 0 \\ 1 & 5 & 0 & -2 & -8 & -3 & 0 \\ 0 & 0 & 5 & 2 & -1 & 0 & -4 \\ 0 & 0 & 0 & 4 & 0 & -1 & -4 \\ 0 & 0 & 0 & 0 & 5 & 3 & -0.5 \end{bmatrix}$$

	Q	$S^{(0)}$	w	no. of iterations	$\|R\|$
Simple iteration	I	$-(A^T+A)^{-1}$	0.02	60	11.9
Jacobi	I	$-(A^T+A)^{-1}$	0.02	60	diverges slowly
Seidel	I	$\frac{1}{2} D^{-1}$	diverges	for all	$w > 1 \times 10^{-4}$
Block Seidel*	I	$\frac{1}{2} D^{-1}$	0.4	20	6.16×10^{-1}
Block Seidel with sequence of 3 permutations**	I	$\frac{1}{2} D^{-1}$	0.3	20	2.53×10^{-1}

*Two by two blocks were used and the rows and columns were first permuted to come in the order (1 3 2 4 5 7 6 8).

**The 3 permutations used put the rows and columns in the orders (1 3 2 4 5 7 6 8), (1 8 2 3 4 5 6 7) and (1 2 3 5 4 6 7 8).

CHAPTER IV

APPLYING THE ROUTH-HURWITZ STABILITY CRITERION

In this chapter we use the criterion of Routh and Hurwitz to establish bounds on the size of ϵ for which the perturbed matrix $A + \epsilon B$ will be stable when A is stable. This method usually gives better bounds on the size of ϵ than the methods of Chapter II, when only a single entry of A is perturbed, but is not applicable to all perturbations B. It also requires calculation of the characteristic polynomial with its associated difficulties, and does not yield any result like the convexity of the set of S-permissible perturbations that would allow us to combine the bounds for perturbations of individual entries to establish stability when more than one entry of A is perturbed.

4.1. The Stability Criterion of Routh and Hurwitz.

Let A be a real n by n matrix and let

$$p(\lambda) = a_0\lambda^n + b_0\lambda^{n-1} + a_1\lambda^{n-2} + b_1\lambda^{n-3} + \dots \tag{1}$$

be the characteristic polynomial of A. The Hurwitz matrix H is defined to be the $n \times n$ square matrix

$$H = \begin{bmatrix} b_0 & b_1 & b_2 & b_3 & \cdots \\ a_0 & a_1 & a_2 & a_3 & \cdots \\ 0 & b_0 & b_1 & b_2 & \cdots \\ 0 & a_0 & a_1 & a_2 & \cdots \\ 0 & 0 & b_0 & b_1 & \cdots \\ 0 & 0 & a_0 & a_1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \end{bmatrix}$$

where each row is terminated with as many zeroes as necessary to make H $n \times n$.

The following theorem, first proved by Hurwitz in 1895, can be found in Gantmacher [9].

Theorem 4.1 (Hurwitz, 1895). Assume $a_0 > 0$. All the roots of the polynomial $p(\lambda)$ (or equivalently all the eigenvalues of the matrix A) have negative real part if and only if all the leading principal minors of the matrix H are positive.

Let us reduce the matrix H to upper triangular form by elementary row operations as follows: Multiply rows $1,3,5,\ldots$ by $-\frac{a_0}{b_0}$ and add them to rows $2,4,6,\cdots$. Call the entry in position $(2,2)$ c_0, multiply rows $2,4,6,\ldots$ by $-\frac{b_0}{c_0}$ and add them to rows $3,5,7,\cdots$. Call the entry in position $(3,3)$ d_0, multiply rows

3,5,7 by $-\frac{c_0}{d_0}$ and add them to rows 4,6,8,$\cdots$. Continue in this manner until the matrix is upper triangular.

We note that none of these operations change the determinant of H or any of its leading principal minors. The first, second, third,..., up to the nth principal minor of H are called the Hurwitz numbers $D_1, D_2, D_3 \ldots D_n$. The numbers b_0, c_0, d_0 on the diagonal of the triangularized matrix are called the Routh numbers. Clearly $D_1 = b_0$, $D_2 = b_0c_0$, $D_3 = b_0c_0d_0$, etc. This shows that the following criterion for stability, first given by Routh in 1877, is equivalent to Hurwitz' criterion.

Theorem 4.2 (Routh, 1877). Assume $a_0 > 0$. All the roots of the polynomial $p(\lambda)$ (or equivalently all the eigenvalues of the matrix A) have negative real part if and only if the Routh numbers $a_0, b_0, c_0, d_0, \ldots$ are all positive.

An independent proof of this theorem can also be found in Gantmacher, as well as a proof of the following theorem due to Orlando:

Theorem 4.3 (Orlando, 1911). Let $\lambda_1, \lambda_2, \ldots, \lambda_n$ be the roots of $p(\lambda)$. Then the (n-1)th Hurwitz number is given by

$$D_{n-1} = (-1)^{\frac{n(n-1)}{2}} a_0^{n-1} \prod_{1 \le i < k \le n} (\lambda_i + \lambda_k), \tag{2}$$

and the nth Hurwitz number $D_n = \det H$ is given by

$$D_n = (-1)^{\frac{n(n+1)}{2}} a_0^n \lambda_1 \lambda_2 \cdots \lambda_n \prod_{1 \le i < k \le n} (\lambda_i + \lambda_k). \tag{3}$$

Corollary 4.4. Det H is zero if and only if for some root λ_0 of $p(\lambda)$, $-\lambda_0$ is also a root of $p(\lambda)$. In particular Det H = 0 if zero or a conjugate pair of pure imaginary numbers are roots of $p(\lambda)$.

4.2. Application to Perturbations. If A and B are real, $A + \epsilon B$ must have a zero eigenvalue or a conjugate pair of pure imaginary eigenvalues when it passes from stability to instability as ϵ changes. Employing Orlando's theorem to recognize when this happens is the key to the proof of our main theorem of this chapter.

Theorem 4.5. Let A be a real stability matrix with characteristic polynomial $p(\lambda)$ and let B be a real matrix such that $A + \epsilon B$ has characteristic polynomial $p(\lambda) + \epsilon q(\lambda)$, with Hurwitz matrix $H_p + \epsilon H_q$. If $\mu_1 < 0 < \mu_2$ are the smallest and largest real eigenvalues of $H_p^{-1} H_q$, then $A + \epsilon B$ is stable for all ϵ such that

$$-\frac{1}{\mu_2} < \epsilon < -\frac{1}{\mu_1}. \tag{4}$$

If there is no real eigenvalue of $H_p^{-1}H_q$ less than 0, (greater than 0), then we can replace $-\frac{1}{\mu_1}$ by $+\infty$ ($-\frac{1}{\mu_2}$ by $-\infty$).

Proof: Since A is a stability matrix, H_p can be transformed by elementary row operations to an upper triangular matrix with the Routh numbers on the diagonal and hence H_p^{-1} exists.

Let $H(\epsilon) = H_p + \epsilon H_q$. $A + \epsilon B$ is stable for $\epsilon = 0$, and since its eigenvalues are continuous functions of ϵ, $A + \epsilon B$ remains stable as ϵ is increased up to some ϵ_0, which is the smallest positive number such that $A + \epsilon_0 B$ has a zero eigenvalue or a pair of conjugate pure imaginary eigenvalues. By Orlando's theorem, this implies that $\det H(\epsilon_0) = 0$.

But

$$\det H(\epsilon_0) = \det(H_p + \epsilon_0 H_q) = 0 \tag{5}$$

is equivalent to

$$\det\left(-\frac{1}{\epsilon_0} I - H_p^{-1}H_q\right) = 0, \tag{6}$$

which implies $-\frac{1}{\epsilon_0}$ is an eigenvalue of $H_p^{-1}H_q$. Since ϵ_0 is the smallest positive number for which this happens, $-\frac{1}{\epsilon_0} = \mu_1$, and $A + \epsilon B$ is stable for all $0 \leq \epsilon < \epsilon_0 = -\frac{1}{\mu_1}$. If $H_p^{-1}H_q$ has no negative eigenvalues

there cannot be an $\epsilon_0 > 0$ with $\det H(\epsilon_0) = 0$, and $A + \epsilon B$ is stable for all positive ϵ.

The left hand inequality in (4) is established similarly by decreasing ϵ from 0 until $\det H(\epsilon)$ is zero. This completes the proof.

The characteristic polynomial of $A + \epsilon B$, $\det(\lambda I-(A+\epsilon B))$, will in general be a polynomial in both λ and ϵ. Unfortunately it will not be linear in ϵ, as the hypothesis of Theorem 4.5 requires, except in special cases. J.S. Frame, in private communication, showed that if B has rank 1, the characteristic polynomial will be linear in ϵ. The next lemma generalizes this result to perturbations B of arbitrary rank.

Lemma 4.6. The degree in ϵ of the characteristic polynomial of $A + \epsilon B$ is less than or equal to the rank of B.

Proof: Let m be the rank of B. Then there exist $n \times m$ matrices X and Y of rank m, such that $B = XY^T$. Since

$$(7) \qquad \begin{bmatrix} \lambda I_n - A & X \\ \epsilon Y^T & I_m \end{bmatrix} = \begin{bmatrix} \lambda I_n - A - \epsilon XY^T & X \\ 0 & I_m \end{bmatrix} \begin{bmatrix} I_n & 0 \\ \epsilon Y^T & I_m \end{bmatrix}$$

$$(8) \qquad \det \begin{bmatrix} \lambda I_n - A & X \\ \epsilon Y^T & I_m \end{bmatrix} = \det(\lambda I_n - A - \epsilon B).$$

Since ϵ occurs in only m rows of the matrix on the left in (7) and (8), the highest power of ϵ that can occur when the determinant is expanded is ϵ^m.

The degree in ϵ of the characteristic polynomial may be strictly less than the rank of B, as can be seen by taking A upper triangular and B strictly upper triangular.

Suppose that B has all zero entries for one row or column, which for ease of notation we take to be the last column. Then B has rank 1,

$$(9) \qquad A + \epsilon B = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n}+\epsilon b_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n}+\epsilon b_{2n} \\ \vdots & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn}+\epsilon b_{nn} \end{bmatrix}$$

and

$$\det(\lambda I-(A+\epsilon B)) = \det \begin{bmatrix} \lambda-a_{11} & -a_{12} & \cdots & -a_{1n} \\ a_{21} & \lambda-a_{22} & \cdots & -a_{2n} \\ & & & \\ a_{n1} & a_{n2} & \cdots & \lambda-a_{nn} \end{bmatrix}$$

$$+ \epsilon \det \begin{bmatrix} \lambda-a_{11} & -a_{12} & \cdots & -b_{1n} \\ a_{21} & \lambda-a_{22} & \cdots & -b_{2n} \\ & & & \\ a_{n1} & a_{n2} & \cdots & -b_{nn} \end{bmatrix} \tag{10}$$

Thus the characteristic polynomial can easily be written as $p(\lambda) + \epsilon q(\lambda)$ and the Hurwitz matrix as $H_p + \epsilon H_q$.

4.3. Examples.

Example 4.7. We illustrate this technique by computing bounds on ϵ such that the matrix

$$A(\epsilon) = \begin{bmatrix} -5 & -10 & 0 & \epsilon \\ 8 & -1 & -4 & 0 \\ 0 & 1 & -1 & -2 \\ 0 & 0 & 2 & -2 \end{bmatrix} \tag{11}$$

is stable. $A = A(0)$ is stable, since it is sign stable. The characteristic polynomial of A is

$$\rho(\lambda) = \lambda^4 + 9\lambda^3 + 113\lambda^2 + 319\lambda + 390 \tag{12}$$

and the characteristic polynomial of $A(\epsilon)$ is

$$p(\lambda) + \epsilon q(\lambda) = \lambda^4 + 9\lambda^3 + 113\lambda^2 + 319\lambda + 390 + \epsilon(-16). \tag{13}$$

The Hurwitz matrix for $A(\epsilon)$ is

$$H_p + \epsilon H_q = \begin{bmatrix} 9 & 319 & & \\ 1 & 113 & 390 & \\ & 9 & 319 & \\ & 1 & 113 & 390 \end{bmatrix}$$

(14)

$$+ \epsilon \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 0 & -16 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & -16 \end{bmatrix}$$

We form $H_p^{-1} H_q$ by elementary row operations on the augmented matrix.

$$\left[\begin{array}{cccc|cccc} 9 & 319 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 113 & 390 & 0 & 0 & 0 & -16 & 0 \\ 0 & 9 & 319 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 113 & 390 & 0 & 0 & 0 & -16 \end{array}\right]$$

First we reduce the left half to upper triangular form with the Routh numbers on the diagonal:

$$\left[\begin{array}{cccc|cccc} 9 & 319 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 77.56 & 390 & 0 & 0 & 0 & -16 & 0 \\ 0 & 0 & 273.74 & 0 & 0 & 0 & +1.86 & 0 \\ 0 & 0 & 0 & 390 & 0 & 0 & -.527 & -16 \end{array}\right]$$

From this we see that the (n-1)th Hurwitz number for $A(\epsilon)$ is $D_{n-1} = 9 \times 77.56 \times (273.74+\epsilon\ 1.86)$. Orlando's theorem shows that $A(\epsilon)$ has a pair of conjugate pure imaginary eigenvalues when $\epsilon = \frac{-273.74}{1.86} = -147.17$. The determinant of $A(\epsilon)$ is $(390-16\epsilon)$, showing that $A(\epsilon)$ has a zero eigenvalue when $\epsilon = \frac{390}{16} = 24.375$. Completing the computation we find

$$H_p^{-1}H_q = \begin{bmatrix} 0 & 0 & 8.54 & 0 \\ 0 & 0 & -.241 & 0 \\ 0 & 0 & \frac{1}{147.17} & 0 \\ 0 & 0 & -.0013 & \frac{-1}{24.375} \end{bmatrix} \tag{15}$$

which has nonzero eigenvalues $\frac{1}{147.17}$ and $-\frac{1}{24.375}$. Thus $A(\epsilon)$ is indeed stable for

$$-147.17 < \epsilon < 24.375. \tag{16}$$

Example 4.8. It will not always be possible to compute D_{n-1} as a linear function of ϵ as in the previous example. For example, if we perturb the (2,4) element of the same matrix A,

$$A(\epsilon) = \begin{bmatrix} -5 & -10 & 0 & 0 \\ 8 & -1 & -4 & \epsilon \\ 0 & 1 & -1 & -2 \\ 0 & 0 & 2 & -2 \end{bmatrix} \tag{17}$$

the characteristic polynomial is

$$(18)\quad p(\lambda) + \epsilon q(\lambda) = \lambda^4 + 9\lambda^3 + 113\lambda^2 + 319\lambda + 390 + \epsilon(-2\lambda - 10),$$

and

$$(19)\quad H_p + \epsilon H_q = \begin{bmatrix} 9 & 319 & 0 & 0 \\ 1 & 113 & 390 & 0 \\ 0 & 9 & 319 & 0 \\ 0 & 1 & 113 & 390 \end{bmatrix} + \epsilon \begin{bmatrix} 0 & -2 & 0 & 0 \\ 0 & 0 & -10 & 0 \\ 0 & 0 & -2 & 0 \\ 0 & 0 & 0 & -10 \end{bmatrix}.$$

D_{n-1} is now a more complicated function of ϵ, but

$$(20)\quad H_p^{-1} H_q = \begin{bmatrix} 0 & -4.92\times10^{-1} & 4.57 & 0 \\ 0 & 7.61\times10^{-3} & -1.44\times10^{-1} & 0 \\ 0 & 9.43\times10^{-5} & 3.07\times10^{-3} & 0 \\ 0 & 1.88\times10^{-5} & 1.18\times10^{-3} & -1.99 \end{bmatrix}$$

which has eigenvalues $-1.99, 0, 1.01\times10^{-3}$, and 9.67×10^{-3}. $A(\epsilon)$ is stable for

$$(21)\quad -103 = \frac{-1}{9.67\times10^{-3}} < \epsilon < \frac{-1}{-1.99} = 0.502.$$

When we can apply Theorem 4.5 we get the true bounds on ϵ for $A + \epsilon B$ to be stable, since $A + \epsilon B$ has zero or pure imaginary eigenvalues when ϵ is equal to the upper or lower bound. Our method based on the Lyapunov equation gives us sufficient bounds on ϵ for $A + \epsilon B$ to be stable, but may be much more restrictive than necessary. This point is readily seen in the following simple example.

Example 4.9. Let $A = \begin{bmatrix} -1 & 0 \\ 10 & -1 \end{bmatrix}$. We know that $\begin{bmatrix} -1 & \epsilon \\ 10 & -1 \end{bmatrix}$ will be stable if and only if $\epsilon < \frac{1}{10}$. The solution to $A^TS + SA = -2I$, is

$$S = \begin{bmatrix} 51 & 5 \\ 5 & 1 \end{bmatrix} \tag{22}$$

Equation (50) of Chapter II says that we have stability for

$$-.0432 = \frac{2}{5 - \sqrt{51^2+5^2}} < \epsilon < \frac{2}{5 + \sqrt{51^2+5^2}} = .0356, \tag{23}$$

much more restrictive than actually necessary. But

$$H_p + \epsilon H_q = \begin{bmatrix} 2 & 0 \\ 1 & 1 \end{bmatrix} + \epsilon \begin{bmatrix} 0 & 0 \\ 0 & -10 \end{bmatrix} \tag{24}$$

and $H_p^{-1}H_q = \begin{bmatrix} 0 & 0 \\ 0 & -10 \end{bmatrix}$ has eigenvalues 0, -10, and equation (4) tells us that we have stability for

$$-\infty < \epsilon < \frac{1}{10}, \tag{25}$$

which are the true bounds.

The eigenvalues u_1 and u_2 of $H_p^{-1}H_q$ in (4) are usually the easiest eigenvalues to compute numerically. If the characteristic polynomial is easy to calculate and the perturbation B has rank one (in particular if only one entry of the matrix is being perturbed), the method of this chapter is preferable to that of Chapter II. But if we wish to perturb entries in several different

rows and columns, especially if we wish to perturb them independently, the Lyapunov method is more suitable.

CHAPTER V

APPLICATIONS TO LOTKA-VOLTERRA AND OTHER MODELS OF ECOSYSTEMS

5.1. The Community Matrix. We now return to our original problem, to investigate the stability of an ecosystem whose population levels are at equilibrium. We use the Lotka-Volterra model introduced in Chapter I, and investigate the stability of an equilibrium point in the first quadrant for the system of differential equations

$$p_i' = p_i\left(r_i + \sum_{j=1}^{n} \alpha_{ij}p_j\right) \quad i = 1\ldots n. \tag{1}$$

Changing to matrix and vector notation (1) becomes

$$p' = D(p)[r + \mathcal{A}p], \tag{2}$$

where, as before, $D(p)$ is the matrix $\operatorname{diag}[p_1, p_2, \ldots, p_n]$. We will call the matrix $\mathcal{A} = [\alpha_{ij}]$ the Volterra Matrix. The process of linearizing (2) about an equilibrium point as done in the introduction, may be written in this notation as follows: Find p^E such that

$$r + \mathcal{A}p^E = 0. \tag{3}$$

We assume that $p^E >> 0$. Letting

$$x = p - p^E, \tag{4}$$

We obtain

(5) $$x' = D(p)\mathcal{A}x = D(p^E)\mathcal{A}x + D(x)\mathcal{A}x.$$

Levins [23] has appropriately named the matrix

(6) $$A = D(p^E)\mathcal{A} = [p_i^E \alpha_{ij}],$$

the community matrix. Mathematicians will recognize it as the Jacobian of (1) evaluated at p^E. Since Ax is the linear part of (5) and

(7) $$D(x)\mathcal{A}x = o(x),$$

the zero solution of (5) and hence the equilibrium point p^E of (2) is asymptotically stable by Theorem 1.6 if A is a stability matrix. In this chapter we investigate the implications of our studies about stability preserving perturbations of A (and the insights we have gained about the nature of stability matrices) for stability of ecosystems.

First we note some of the strengths and weaknesses of our approach to the problem of stability. The major weakness is that we can only prove local stability, not global stability. That is, we know that if the perturbation $x(0)$ of the initial populations away from the equilibrium point is small enough, the populations will return to their equilibrium values, but we do not know that any initial population will move toward the equilibrium values. We will later find some sufficient conditions for $x(0)$ to be small enough. On the other hand we must remember that local stability is a necessary condition for global stability.

The major strength of our work is its applicability not only to Lotka-Volterra models, but also to other models of ecosystems, and in fact to general autonomous dynamical systems. For if

$$p' = f(p) \tag{8}$$

and

$$f(p^E) = 0, \tag{9}$$

then letting $x = p - p^E$ we have

$$x' = Ax + g(x) \tag{10}$$

where A is the Jacobian matrix given by

$$a_{ij} = \frac{\partial f_i}{\partial p_j}(p^E) \tag{11}$$

and $g(x) = o(x)$. May [28] has reviewed some of the forms commonly used for the function f. We could also study age dependent and sex dependent effects by treating age and sex classes as different species.

Gilpin and Justice [11] have shown that an arbitrary ecosystem model which they write in the form

$$p_i' = p_i R_i(p) \quad i = 1...n, \tag{12}$$

can be approximated near an equilibrium point p^E such that

$$R(p^E) = 0 \tag{13}$$

by a Lotka-Volterra model, which they (and many other authors) write in the form introduced by Gause [10],

$$p_i' = p_i \frac{r_i}{k_i} \left(k_i - \sum_{j=1}^{n} \gamma_{ij} p_j\right) \tag{14}$$

$$\gamma_{ii} = 1. \tag{15}$$

Although they give a detailed geometrical description of their procedure, what they have really done is to approximate $R(p)$ by a first order Taylor series about p^E.

The relationships between (1), (12), and (14) are given by

$$\alpha_{ij} = \frac{\partial R_i}{\partial p_j}(p^E) = \alpha_{ii}\gamma_{ij} \tag{16}$$

$$r_i = -\sum_{j=1}^{n} \frac{\partial R_i}{\partial p_j}(p^E) p_j^E = \sum_{j=1}^{n} \alpha_{ij} p_j^E \tag{17}$$

$$\frac{r_i}{k_i} = -\alpha_{ii} \tag{18}$$

But if one wishes to study the stability of the equilibrium point p^E by approximating (12) by a Lotka-Volterra model (1) and then forming the community matrix (6), he might as well take the Jacobian (11) directly, since if

$$f_i(p) = p_i R_i(p), \tag{19}$$

using (16) and (13) gives

$$\frac{\partial f_i}{\partial p_j}(p^E) = p_i^E \frac{\partial R_i}{\partial p_j}(p^E) = p_i^E \alpha_{ij} \quad \text{for} \quad i \neq j \tag{20}$$

$$(21) \qquad \frac{\partial f_i}{\partial p_i}(p^E) = R_i(p^E) + p_i^E \frac{\partial R_i}{\partial p_i}(p^E) = p_i^E \alpha_{ii},$$

and the Jacobian matrix calculated from (11) is the same as the community matrix calculated from (6).

The constants r_i, k_i, and α_{ij} in (1) and (14) are usually given the following biological interpretations: r_i is the intrinsic rate of growth of species i, i.e. births minus deaths per individual per unit of time when the population is near zero; k_i is the carrying-capacity of the environment for species i, i.e. the number of individuals of species i which causes the rate of growth to decline to zero due to overcrowding even in the absence of other species; and α_{ij} is the change in the growth rate of species i per each individual of species j, so that $\sum_{j=1}^{n} \alpha_{ij} p_j$ is the total change (either damping or acceleration) of the growth rate due to species interaction.

It is often better to measure population levels in terms of units of biomass (that is, the total mass of all members of a species) instead of numbers of individuals. When this is done, the word individual is replaced by the phrase unit of biomass throughout the above paragraph.

In equation (12) the intrinsic growth rate for species i would be $R_i(0)$, and the carrying capacity

would be the value of p_i which makes $R_i(0 \ldots 0\ p_i\ 0 \ldots 0) = 0$. As Gilpin and Justice point out, these are not necessarily the same values as r_i and k_i obtained when approximating (12) by (14). This is because, although

$$R_i(p) \approx r_i + \sum_j \alpha_{ij} p_j \tag{22}$$

near the equilibrium point p^E, it may not be a good approximation away from p^E, so that it may not be true that

$$R_i(0) \approx r_i \tag{23}$$

$$R_i(0 \ldots 0\ k_i\ 0 \ldots 0) \approx r_i + \alpha_{ii} k_i = 0. \tag{24}$$

Gilpin and Justice therefore state that when a general model of the form (12) is approximated by a Lotka-Volterra model the coefficients no longer have any biological significance, but what we must do is interpret them in terms of what happens near the equilibrium point. Thus we could reinterpret r_i as the intrinsic growth rate at equilibrium, which is just balanced by the damping (or acceleration) $\sum_{j=1}^{n} \alpha_{ij} p_j$ when $p = p^E$, and $\frac{k_i}{r_i} = -\frac{1}{\alpha_{ii}}$ as the number of additional individuals the environment would support if the levels of the other species are changed so as to decrease the damping by one unit.

We return to the community matrix A given by (6) or (11) and investigate the biological significance

of the entries of A. Whether the matrix A comes from a Lotka-Volterra model or a more complicated function f, the signs of the entries a_{ij} and a_{ji} correspond to the types of species interaction: (-+) species j preys on species i, (or is a parasite of species i), (--) both species compete for the same resource, (++) symbiosis, (OO) no interaction, (+O) commensalism, and (-O) amensalism. Because of inefficiencies in energy use and transfer, the gains to a predator in a predator prey interaction are never as great as the loss to the prey. Thus if j preys on i, we can assume $|a_{ij}| > |a_{ji}|$.

The diagonal entries a_{ii} indicate the amount of self interaction: a_{ii} negative means that growth of species i is self damped near the equilibruim by some type of inner competition for resources, a_{ii} zero means that growth of species i is only limited through interactions with other species. Positive a_{ii} in the Lotka-Volterra model is biologically unrealistic, since it would imply positive α_{ii} which would mean that in the absence of other species, species i grows without bounds and even faster than exponentially.

The trophic level of a species is the number of steps in the food chain between it and the abiotic nutrients. Thus green plants are in the first trophic level, herbivores in the second, carnivores which eat

herbivores in the third, etc. (Some species, however, may be in more than one trophic level.)

It is often assumed that at the first trophic level the species exhibit self damping since plant growth is limited by abiotic factors, but that there is no self damping at higher trophic levels, since herbivores and carnivores are limited only by the existence of prey and by being preyed upon. Holling [16], however, has found that many predators do exhibit self damping behavior if their numbers become too large.

The existence of some self damping is crucial for stability, because A cannot have all negative eigenvalues unless it has a negative trace. Of course if the a_{ii} are sufficiently negative, i.e. if there is enough self damping, then A is almost certainly stable. We have also seen that large symmetric off diagonal entries (competition and symbiosis), tend to be destabilizing, and that large skew symmetric entries (predator-prey) tend to have a neutral effect or else to be stabilizing in the sense that they pull the real parts of the eigenvalues closer together. This leads us to the conclusion that even though the system may be unstable at one trophic level, coupling it through predator-prey interactions with a stable trophic level may bring

about stability. (May [28] has calculated the conditions for this to happen in a 3 species two trophic level system.)

The matrix A in Example 2.25 (pp.42 and 73) example of a possible community matrix for an ecosystem with four trophic levels, 2 plant species, 2 herbivores, 2 carnivores, and one top carnivore. This example shows no direct competition within any trophic level. It would be reasonable to add competition terms between the plants by making both a_{12} and a_{21} negative. The matrix A(0) in Example 4.7 (p.81) is another example of a community matrix, representing a simple food chain with only one species at each trophic level. Alternatively, it might be considered as the result of lumping all the species together in each trophic level.

Our original goal was to take such a community matrix, with the entries known only within certain error bounds, and devise methods to decide whether any matrix within these error bounds gives a stable system. We can now do this using the methods of Chapters II and IV, if we are willing to do enough numerical computation. Unfortunately, the author found no readily available examples where the community matrix for a natural ecosystem has been estimated. Most of the empirical work in ecology has been at the micro-ecology level, studies of

interactions between a few competing species and a few laboratory studies of predator-prey relations. In spite of declarations by many of the leaders of the discipline about the importance of studying ecosystems as complete systems and the dangers of looking at only isolated parts (see for example Holling [15] and Hardin [14]), empirical work on a macroecology scale is still in the pioneering stage. We will let examples 2.25, 2.29, 4.7, and 4.8 serve as illustrations of what the mathematics developed in this thesis makes possible once the coefficients of a community matrix have been estimated.

Some recent theoretical work has been done which attempts to compute the interaction coefficients in a strictly competitive community (no predator-prey interactions) from the a measurement of "niche overlap." (See McArthur [26] and [27].) More work is clearly needed on methods to measure the strength of predator-prey interactions and their effect on birth and death rates.

It should be remembered that the community matrix A is obtained by multiplying each row of the original matrix $\mathcal{A} = [\alpha_{ij}]$ by the corresponding equilibrium values p_i^E, which in turn depend on the growth rates r_i. We observe in nature that due to energy losses as one moves up the food chain, the equilibrium values for lower trophic levels, measured in units of biomass, are

generally greater than those of higher trophic levels. A careful study of conditions for a Lotka-Volterra system to have a set of positive equilibrium values $\{p_i^E\}$, and the effect of these values on the community matrix needs to be undertaken.

The equilibrium value factors in the community matrix complicate the effect that perturbations of the coefficients in (1) have on the community matrix. If growth rate vector r is perturbed to $r + \Delta r$ and the Volterra matrix $\mathcal{A}$ is perturbed to $\mathcal{A} + \Delta\mathcal{A}$, then the new equilibrium values $p^E + \Delta p^E$ satisfy

$$r + \Delta r + (\mathcal{A}+\Delta\mathcal{A})(p^E+\Delta p^E) = 0, \tag{25}$$

and the new community matrix $A + B$ satisfies

$$A + B = D(p^E+\Delta p^E)(\mathcal{A}+\Delta\mathcal{A}) = D(p^E)(\mathcal{A}+\Delta\mathcal{A}) + D(\Delta p^E)(\mathcal{A}+\Delta\mathcal{A}). \tag{26}$$

Thus

$$\Delta p^E = -(\mathcal{A}+\Delta\mathcal{A})^{-1}(\Delta r+\Delta\mathcal{A}p^E) \tag{27}$$

and the perturbation to the community matrix is

$$B = D(p^E)\Delta\mathcal{A} + D(\Delta p^E)(\mathcal{A}+\Delta\mathcal{A}). \tag{28}$$

5.2. D-Stable Volterra Matrices. If the Volterra matrix $\mathcal{A}$ is D-stable (Def. 2.9), then the community matrix A will be stable for any vector of growth rates r which produces positive equilibrium

values. Of course negative equilibrium populations are unrealistic. In some systems the Volterra matrix $\mathcal{A}$ is clearly D-stable.

If we consider only a trophic level (or simple food chain) model, the matrix $\mathcal{A}$ will have the sign pattern

$$\begin{bmatrix} \leq 0 & - & & \\ + & \leq 0 & - & \\ & + & \leq 0 & - \\ & & + & \leq 0 \end{bmatrix}$$

which will be sign-stable and thus D-stable if any of the diagonal terms are strictly negative. Since the first level can be assumed to exhibit self damping, $a_{11} < 0$, and the community matrix A is stable.

Another special form of the matrix $\mathcal{A}$, used by Kerner [21], leads to D-stability. The basic assumption is that when species j preys on species i, the gain to species j is $\gamma_j \alpha_{ij}$, where $-\alpha_{ij}$ is the loss to species i, and that the proportionality constant γ_j depends only on the efficiency of species j in utilizing its food, and not on species i. This leads to the conclusion that equation (1) can be written in the form

$$p_i' = p_i (r_i + \lambda_i \sum_j \alpha_{ij} p_j), \tag{29}$$

$$\alpha_{ji} = -\alpha_{ij},$$

for which the community matrix A has (i,j) entry

(30) $$a_{ij} = p_i^E \lambda_i \alpha_{ij}.$$

Under the assumption that all the diagonal terms α_{ii} are zero, Kerner shows that

(31) $$\varphi = \sum_{i=1}^{n} (p_i(t) - p_i^E \ln p_i(t)) \frac{1}{\lambda_i}$$

is a constant of motion for this system and uses it to prove that there are neutrally stable oscillations about the equilibrium point and to build a "statistical mechanics" for the system. Of course if the diagonal entries are all negative,

(32) $$D = \operatorname{diag}\left[\frac{1}{p_i^E \lambda_i}\right]$$

gives

(33) $$A^T D + DA = \operatorname{diag}[\alpha_{ii}],$$

the community matrix is stable, and the system has oscillations spiraling in toward the equilibrium point. These results are often criticized as being "fragile" since they are based on the probably invalid assumption that $\alpha_{ji} = -\alpha_{ij}$. Neither Kerner's school nor his critics have realized that they can get the same conclusions with the following weaker hypothesis.

If $\mathcal{A}$ is D-negative-definite (Def. 2.13), which will be the case if $\mathcal{A}$ has the form

(34) $$\mathcal{A} = E(M+K)$$

with E positive diagonal, M symmetric, negative definite, and K skew symmetric, then $\mathcal{A}$ will be D-stable and any positive equilibrium point will be asymptotically stable. (Kerner's formulation (29) assumes that M is either 0 or a negative diagonal matrix.) Thus we can tolerate competition terms and departures from skew-symmetric in the predator-prey terms which are not too large compared to the self damping terms, and still have asymptotic stability for any realistic set of growth rates.

As the symmetric terms become larger, we may still have asymptotic stability for particular equilibrium values, which means for particular growth rates $\{r_i\}$, but not for all of them. Finally as the competition terms become too large, the eigenvalues are pushed far enough apart that some of them become positive, the system becomes unstable and some species become extinct. This is expressed in ecology as the competitive exclusion principle: that two species competing for exactly the same resources cannot co-exist indefinitely.

5.3. Addition of a New Species to an Ecosystem. Suppose we have an ecosystem modeled by equation (1) with stable community matrix given by equation (6), and add another species. We seek conditions for the new enlarged system to be stable. The new community matrix

will be the old community matrix bordered by a new row and column plus a perturbation of the old community matrix due to the change in the equilibrium values.

The enlarged system can be written as

$$(35)\qquad p_i' = p_i\Big(r_i + \sum_{j=1}^{n} \alpha_{ij}p_j + b_i p_{n+1}\Big), \quad i = 1\ldots n$$

$$p'_{n+1} = p_{n+1}\Big(r_{n+1} + \sum_{j=1}^{n} c_j p_j + d p_{n+1}\Big).$$

Thus the Volterra matrix $\mathcal{A}$ is replaced by the bordered matrix $\begin{bmatrix} \mathcal{A} & b \\ c^T & d \end{bmatrix}$ where b is the vector $(b_1\ldots b_n)^T$ and c^T is the vector $(c_1\ldots c_n)$. Let $p^E + \Delta p^E$ be the vector of new equilibrium values for species $1,\ldots,n$ and p^E_{n+1} the equilibrium value for the added species, and let r be the vector $(r_1,\ldots,r_n)^T$. Then

$$\begin{bmatrix} p^E+\Delta p^E \\ p^E_{n+1} \end{bmatrix} = -\begin{bmatrix} \mathcal{A} & b \\ c^T & d \end{bmatrix}^{-1}\begin{bmatrix} r \\ r_{n+1} \end{bmatrix}$$

$$(36a)\qquad = -\begin{bmatrix} \mathcal{A}^{-1} + \frac{1}{e}\,\mathcal{A}^{-1}bc^T\mathcal{A}^{-1} & -\frac{1}{e}\,\mathcal{A}^{-1}b \\ -\frac{1}{e}\,c^T\mathcal{A}^{-1} & \frac{1}{e} \end{bmatrix}\begin{bmatrix} r \\ r_{n+1} \end{bmatrix},$$

where

$$(36b)\qquad e = d - c^T\mathcal{A}^{-1}b.$$

Since $p^E = -\mathcal{A}^{-1}r$, (36) implies

$$(37a)\qquad p^E_{n+1} = -\frac{1}{e}(r_{n+1} - c^T\mathcal{A}^{-1}r) = -\frac{r_{n+1}+c^Tp^E}{d-c^T\mathcal{A}^{-1}b}$$

(37b) $$\Delta p^E = \frac{1}{e}(r_{n+1}-c^T\mathcal{A}^{-1}r)\mathcal{A}^{-1}b = -p^E_{n+1}\mathcal{A}^{-1}b.$$

Our first requirement for a stable ecosystem is that the new equilibrium values be positive, i.e.

(38a) $$p^E_{n+1} = -\frac{r_{n+1}-c^Tp^E}{d-c^T\mathcal{A}^{-1}b} > 0$$

(38b) $$p^E + \Delta p^E = p^E - p^E_{n+1}\mathcal{A}^{-1}b >> 0$$

The new community matrix is now seen to be

(39) $$D\begin{pmatrix} p^E+\Delta p^E \\ p^E_{n+1}\end{pmatrix}\begin{bmatrix}\mathcal{A} & b \\ c^T & d\end{bmatrix} = \begin{bmatrix} A-p^E_{n+1}D(\mathcal{A}^{-1}b)\mathcal{A} & D(p^E-p^E_{n+1}\mathcal{A}^{-1}b)b \\ p^E_{n+1}c^T & p^E_{n+1}d\end{bmatrix}.$$

If S is the solution to the Lyapunov equation

(40) $$A^TS + SA = -I,$$

we see from equations (28) and (64) of Chapter II that sufficient conditions for stability are

(41) $\nu p^E_{n+1} - 1 < 0$, ν the largest eigenvalue of
$$-[\mathcal{A}^TD(\mathcal{A}^{-1}b)S + SD(\mathcal{A}^{-1}b)\mathcal{A}]$$

(42a) $$d < \frac{-p^E_{n+1}}{1-\nu p^E_{n+1}}\left(\sqrt{c^Tcb^TES^2Eb} + c^TSEb\right),$$

where

(42b) $$E = D(p^E-p^E_{n+1}\mathcal{A}^{-1}b).$$

Equations (38b) and (41) lead to the surprising conclusion that we are most likely to have stability in the expanded system, if the vector b, which gives the effects of the added species on the old species, is directly proportional to the vector r, which gives the intrinsic growth rates of the old species. For if $b = \beta r$ with $\beta > 0$, then

$$\Delta p^E = \beta p^E_{n+1} p^E, \tag{43}$$

which shows that the inequality (38b) is satisfied, and the perturbation to A is

$$-p^E_{n+1} D(\alpha^{-1} b)\alpha = \beta p^E_{n+1} A, \tag{44}$$

which shows that $\nu = -\beta$ in (41) and (42). This may be hard to accomplish in nature, since the usual assumption that the plants have a positive growth rate would imply that the added species should be preyed upon by the plants!

The difficulty of adding a new species to an ecosystem and still maintaining stability, supports the theme stressed by Robert May, that stability is not enhanced by increased complexity. In fact, our results are another addition to the list of mathematical studies indicating that increasing the number of species makes it more difficult to have stability. (See May [28].)

Perhaps the reason that many field ecologists hold the opposite view, that complexity increases stability, is that they consider stability to be the lack of large fluctuations or oscillations over a short time period, which is something very different from Lyapunov asymptotic stability, which allows large oscillations as long as they are damped over a long time period.

5.4. Domain of Attraction by Lyapunov Functions

When p^E is asymptotically stable, solutions which start close enough to p^E approach p^E in the limit. We now want to investigate how close is close enough.

Definition 5.1. The domain of attraction of a point p^E for the differential equation (8) is the set of all points p_0 such that the solution $p(t)$ with initial value p_0 satisfies $\lim_{t\to\infty} p(t) = p^E$.

Expositions of the following theorem, due to LaSalle [22], can be found in Hahn [12], and Yoshizawa [44].

Theorem 5.2 (LaSalle, 1960). Given the differential equation

$$y' = F(y), \tag{45}$$

let $V(y)$ be a function with first order partial derivatives, $V(y) \geq 0$, and $V'(y(t)) \leq 0$ in a region $R = \{y: V(y) < a\}$, and let M be the largest invariant

subset of the set $\{y:V'(y) = 0\}$. Then every solution of (45) which starts in R approaches M as $t \to \infty$.

Corollary 5.3. Assume that in equation (45), $F(0) = 0$. Assume that $V(y)$ has first order derivatives, $V(y) > 0$ and $V'(y) < 0$ for $y \neq 0$ on the set $R = \{y:V(y) < a\}$, and $V(0) = V'(0) = 0$. Then 0 is an asymptotically stable equilibrium point for (45) and R is contained in the domain of attraction of 0.

We apply this corollary to get an estimate of the domain of attraction for the Lotka-Volterra system. The matrix norm used in the following theorem can be $\| \|_1$, $\| \|_2$, or $\| \|_\infty$.

Theorem 5.4. Let p^E be the equilibrium point of equation (1) given by equation (3). Let A defined by equation (6) be a stability matrix,

$$B = GD(p^E) = D^{-1}(p^E)AD(p^E), \tag{46}$$

Q a positive definite matrix with smallest eigenvalue q, and S the solution to

$$B^T S + SB = -Q \tag{47}$$

with smallest eigenvalue σ. Then p^E is asymptotically stable and any population vector p such that the vector y defined by

$$y_i = \frac{p_i - p_i^E}{p_i^E} \tag{48}$$

satisfies

$$\sqrt{y^T S y} < \frac{\sqrt{\sigma}\, q}{(\|B^T\|+\|B\|)\,\|S\|} \tag{49}$$

is in the domain of attraction of p^E.

Proof: For any solution $p(t)$ of equation (1), let $x(t)$ be given by equation (4), and $y(t)$ by (48). Then

$$y(t) = D^{-1}(p^E)x(t) \tag{50}$$

and from (5)

$$y'(t) = D^{-1}(p^E)x'(t) = \mathcal{A}D(p^E)y(t) + D(y(t))\mathcal{A}D(p^E)y(t). \tag{51}$$

Clearly p^E is asymptotically stable and a point p is in the domain of attraction of p^E if and only if 0 is asymptotically stable and y is in the domain of attraction of 0 for equation 51.

Since B and A are similar, B is a stability matrix and S is positive definite. It follows from (46), (47) and (51) that

$$(y^T S y)' = y^T[-Q + B^T D(y) S + S D(y) B]y. \tag{52}$$

Since

$$\|B^T D(y) S + S D(y) B\| \leq \|D(y)\| \cdot \|S\| (\|B^T\|+\|B\|), \tag{53}$$

and

$$\|D(y)\| = \max_i y_i \leq \sqrt{y^T y} \leq \sqrt{\frac{y^T S y}{\sigma}}, \tag{54}$$

the matrix in the brackets in (52) is negative definite, for any y satisfying (49, and $(y^TSy)^{\cdot} < 0$ if $y \neq 0$. The conclusion now follows from Corollary 5.3 to LaSalle's theorem.

If $Q = I$ and B (and hence A) have eigenvalues $\{\alpha_k+i\beta_k\}$ with $\alpha_1 \leq \ldots \leq \alpha_n < 0$, then $\|B^T\| + \|B\| \geq 2|\alpha_1| \geq \max_k 2|\alpha_k+i\beta_k|$, and from Corollary 2.8 $\|S\| \geq -\frac{1}{2\alpha_n}$ and $\sigma \leq -\frac{1}{2\alpha_1}$. Thus

$$(55) \qquad \frac{\sqrt{\sigma}}{(\|B^T\|+\|B\|)\,\|S\|} \leq \frac{|\alpha_n|}{\sqrt{2|\alpha_1|^3}} \leq \frac{|\alpha_n|}{\max_k |\alpha_k+i\beta_k|\sqrt{2|\alpha_1|}}.$$

Hence if B is ill-conditioned ($\frac{|\alpha_n|}{|\alpha_1|}$ large) or has large imaginary parts to its eigenvalues the inequality in (59) will be very restrictive.

If, however, $\mathcal{A}$ is D-negative-definite, the next theorem shows that the entire positive quadrant is in the domain of attraction of p^E. The Lyapunov function used was first applied to Lotka-Volterra systems by Aiken and Lapidus [1], who were following the school of thought of Kerner and of Montroll and Goel [30]. They applied it, however, only when $\mathcal{A}$ was a diagonal plus skew symmetric matrix, failing to realize once again that it gives the same results with much weaker hypothesis.

Theorem 5.5. Let the matrix $\mathcal{A} = [\alpha_{ij}]$ in equation (1) and (2) be such that $D\mathcal{A}$ has negative definite (negative semidefinite) symmetric part for some matrix $D = \text{diag}[d_1 \ldots d_n]$, $d_i > 0$ for all i. Then for any growth rates r such that the equilibrium vector p^E defined by equation (3) satisfies

$$p^E >> 0, \tag{56}$$

p^E is asymptotically stable (stable), and its domain of attraction includes all populations p with

$$p >> 0. \tag{57}$$

Proof: Let $p(t)$ be any solution of (1) with initial value p_0 satisfying (57). Define

$$V(p) = \sum_i d_i \left(p_i - p_i^E - p_i^E \ln\left(\frac{p_i}{p_i^E}\right)\right). \tag{58}$$

Since

$$z - 1 - \ln(z) \geq 0 \tag{59}$$

with equality only if $z = 1$, $V(p) \geq 0$ for any $p >> 0$ with equality only if $p = p^E$. Now

$$V'(p(t)) = \sum_i d_i \left(p_i'(t) - p_i^E \frac{p_i'(t)}{p_i(t)}\right) \tag{60}$$

and $p_i' = x_i'$, so that by (5) (or equation (6) of Chapter I)

$$V'(t) = \sum_i d_i \left((p_i^E + x_i) \sum_j \alpha_{ij} x_j - p_i^E \sum_j \alpha_{ij} x_j\right) \tag{61}$$

$$V'(t) = \sum_i \sum_j x_i d_i \alpha_{ij} x_j = x^T D\mathcal{A} x. \tag{62}$$

Stability when $D\mathcal{A}$ has negative semidefinite symmetric part now follows from Lyapunov's Theorem 1.9. Asymptotic stability and the domain of attraction when $D\mathcal{A}$ has negative definite symmetric part follow from Corollary 5.3 to LaSalle's theorem since any point $p \gg 0$ is in the set $\{p:V(p) < a\}$ for some a.

We note that the paper by Aiken and Lapidus contains an error. They claim asymptotic stability when $\mathcal{A}$ is a diagonal plus a skew symmetric matrix with non-positive diagonal entries but only one nonzero diagonal entry. But that only makes the symmetric part of $\mathcal{A}$ negative-semidefinite, proving stability but not asymptotic stability.

We also note that this generalizes a result of McArthur [27], who arrived at the same conclusion when $D\mathcal{A}$ was symmetric and negative-definite for some positive diagonal matrix D, by using $-x^T D\mathcal{A}x$ as a Lyapunov function. But this limits him to a restricted set of competition equations, since predator-prey or unbalanced competition interactions introduce a skew symmetric component to $\mathcal{A}$.

5.5. Time Varying Coefficients in the Lotka-Volterra Model. Certainly it is more realistic to expect the growth rates r_i and even the interaction coefficients α_{ij} to vary with time instead of being

constant. To investigate how this affects stability, we add time varying perturbations to the coefficients in the Lotka-Volterra model. Equation (1) becomes

$$p_i' = p_i(r_i+\rho_i(t) + \sum_{i=1}^{n} (\alpha_{ij}+\beta_{ij}(t))p_j). \tag{63}$$

Let $\rho(t) = (\rho_i(t))$ and $\mathcal{B}(t) = [\beta_{ij}(t)]$. There is no longer a fixed equilibrium point as in equation (1), but we define a time varying critical vector $c(t)$ by

$$(r+\rho(t)) + (\mathcal{A}+\mathcal{B}(t))c(t) = 0. \tag{64}$$

(We may think of $c(t)$ as a moving equilibrium point.) Let

$$z(t) = p(t) - c(t), \tag{65}$$

so that

$$\begin{aligned} z_i'(t) &= c_i(t)[\sum_{i=1}^{n} (\alpha_{ij}+\beta_{ij}(t))z_i(t)] \\ &\quad + z_i(t)[\sum_{i=1}^{n} (\alpha_{ij}+\beta_{ij}(t))z_i(t)] + c_i'(t) \end{aligned} \tag{66}$$

or in matrix notation

$$\begin{aligned} z'(t) &= D(c(t))[\mathcal{A}+\mathcal{B}(t)]z(t) + D(z(t))[\mathcal{A}+\mathcal{B}(t)]z(t) \\ &\quad + c'(t), \end{aligned} \tag{67}$$

$$c'(t) = -[\mathcal{A}+\mathcal{B}(t)]^{-1}[\rho'(t)-\mathcal{B}'(t)c(t)]. \tag{68}$$

We will show that if $c'(t)$ is small enough, $p(t)$ will stay close to the critical vector $c(t)$, or in other

words that solutions $z(t)$ of (67) will be bounded, the size of the bound depending on the size of $c'(t)$ and the eigenvalues of $\alpha + \beta(t)$.

Equation (67) contains a linear term, a quadratic term, and a forcing term $c'(t)$. We express the linear part as

$$z'(t) = [A+B(t)]z(t) \tag{69}$$

where

$$A + B(t) = D(c(t))(\alpha+\beta(t)). \tag{70}$$

Equation (70) does not define A and $B(t)$ uniquely. We could define them uniquely by defining A, say by equation (6), and letting (70) define $B(t)$, but it is not clear what is the best choice for A. We require A to be a stability matrix, so that we can solve the Lyapunov equation (40) for a positive-definite symmetric matrix S. Then if $B(t)$ satisfies the conditions of Theorems 2.15, 2.16, or 2.17 for all t, z^TSz will be a Lyapunov function for (69), the linear part of (67) will be asymptotically stable by Theorem 1.9, and this choice for S in equation (71) below will make q positive, as will be required in Theorems 5.7 and 5.8. Ideally we would like to find the positive-definite symmetric matrix S which makes q in (71) as large as possible, but it is not clear how to do so.

Lemma 5.6. Let S be a positive definite symmetric matrix,

$$-q = \sup_t \{\text{largest eigenvalue of } [A^T S + SA + B^T(t)S + SB(t)]\}, \tag{71}$$

$$d = \sup_t 2\|S\|_2 \|\alpha+\beta(t)\|_2, \tag{72}$$

$$L = \sup_t 2\|Sc'(t)\|_2, \tag{73}$$

and σ the smallest eigenvalue of S. Then

$$(z^T(t)Sz(t))' \leq -\frac{q}{\sigma} z^T S z + d\left(\frac{z^T S z}{\sigma}\right)^{\frac{3}{2}} + L\left(\frac{z^T S z}{\sigma}\right)^{\frac{1}{2}}. \tag{74}$$

Proof: From equations 67 to 71,

$$(z^T(t)Sz(t))' = z^T(A^T S + SA + B^T(t)S + SB(t))z + 2z^T(\alpha+\beta(t))^T D(z) Sz + 2z^T Sc'. \tag{75}$$

$$z^T(A^T S + SA + B^T(t)S + SB(t))z \leq -qz^T z. \tag{76}$$

By the Cauchy Schwarz inequality

$$\begin{aligned} 2z^T(\alpha+\beta(t))^T D(z) Sz &\leq 2\|z\|_2 \|(\alpha+\beta(t))^T D(z) Sz\|_2 \\ &\leq 2\|\alpha+\beta(t)\|_2 \|S\|_2 \|D(z)\|_2 \|z\|_2^2 \\ &\leq 2\|\alpha+\beta(t)\|_2 \|S\|_2 (z^T z)^{\frac{3}{2}}, \end{aligned} \tag{77}$$

since $\|D(z)\|_2 \leq \|z\|_2$. Also

$$2z^T Sc' \leq 2\|z\|_2 \|Sc'\|_2 \leq 2\|Sc'\|_2 (z^T z)^{\frac{1}{2}}. \tag{78}$$

Since $z^T z \leq \frac{1}{\sigma} z^T S z$, (74) follows.

Theorem 5.7. Let $q,d,L,$ and σ be as in the previous lemma. Assume that q is positive,

$$L < \frac{q^2}{4d} \tag{79}$$

and let

$$w = \sqrt{q^2 - 4dL}\,. \tag{80}$$

then for any solution $z(t)$ of (67) with $z(0) \in R = \{z:(z^T Sz)^{\frac{1}{2}} < \frac{\sqrt{\sigma}}{2d}(q+w)\}$, $z(t)$ remains in R for all t, and for any $\varepsilon > 0$ the values of $(z^T(t)Sz(t))^{\frac{1}{2}}$ are in the set $[0, \frac{\sqrt{\sigma}}{2d}(q-w) + \varepsilon]$ for sufficiently large t.

Proof: Let $W(t) = (z^T(t)Sz(t))^{\frac{1}{2}}$. From (74)

$$W' \leq \frac{1}{2}\left(\frac{d}{\sigma^{3/2}} W^2 - \frac{q}{\sigma} W + \frac{L}{\sigma^{1/2}}\right). \tag{81}$$

From the comparison principle (see Yoshizawa [44]), $W(t) \leq u(t)$ for all t, where $u(t)$ is the solution to

$$u'(t) = \frac{1}{2}\left(\frac{d}{\sigma^{3/2}} u^2 - \frac{q}{\sigma} u + \frac{L}{\sigma^{1/2}}\right),\ u(0) = W(0). \tag{82}$$

But from (79) and the quadratic formula, u' is negative for all u between $\frac{\sqrt{\sigma}}{2d}(q \pm w)$ and positive elsewhere. Thus $u(t) \to \frac{\sqrt{\sigma}}{2d}(q-w)$ for all $0 \leq u(0) \leq \frac{\sqrt{\sigma}}{2d}(q+w)$, which completes the proof.

Theorem 5.7 shows that if $z(0) = p(0) - c(0)$ is in R, $z(t)$ is bounded, and hence if $c(t)$ is bounded, $p(t) = c(t) + z(t)$ is bounded. It also shows that if $c'(t)$ is small enough in comparison to q, the solution vector $p(t)$ follows the critical vector $c(t)$ quite closely, because $\frac{\sigma^{1/2}}{2d}(q-w)$ goes to zero as $L = \sup_t 2\|Sc'(t)\|$ goes to zero. Thus as $c'(t)$ decreases the difference between $p(t)$ and $c(t)$ also decreases.

Figures 3, 4, and 5 illustrate these points. The two dimensional system

$$(83) \qquad p_1' = p_1[10 + (-2.5+\cos \omega t)p_1 + (-5 - 2 \sin \omega t)p_2]$$

$$p_2' = p_2[-5 + 4p_1 + (-1+\cos \omega t)p_2]$$

was solved numerically by fifth order Runge-Kutta and plotted along with the critical values $c_1(t)$ and $c_2(t)$ defined by (64), for three different values of ω. As ω is decreased q and d remain the same, but L decreases, resulting in smaller differences between $p_i(t)$ and $c_i(t)$. One can also observe the initial transient solution corresponding to the time period when $p(t) - c(t)$, as measured by $(z^T(t)Sz(t))^{1/2}$, is decreasing, followed by the steady state oscillations during the time period when $z^T(t)Sz(t)$ is in the set $[0, \frac{\sqrt{\sigma}}{2d}(q-w) + \epsilon]$.

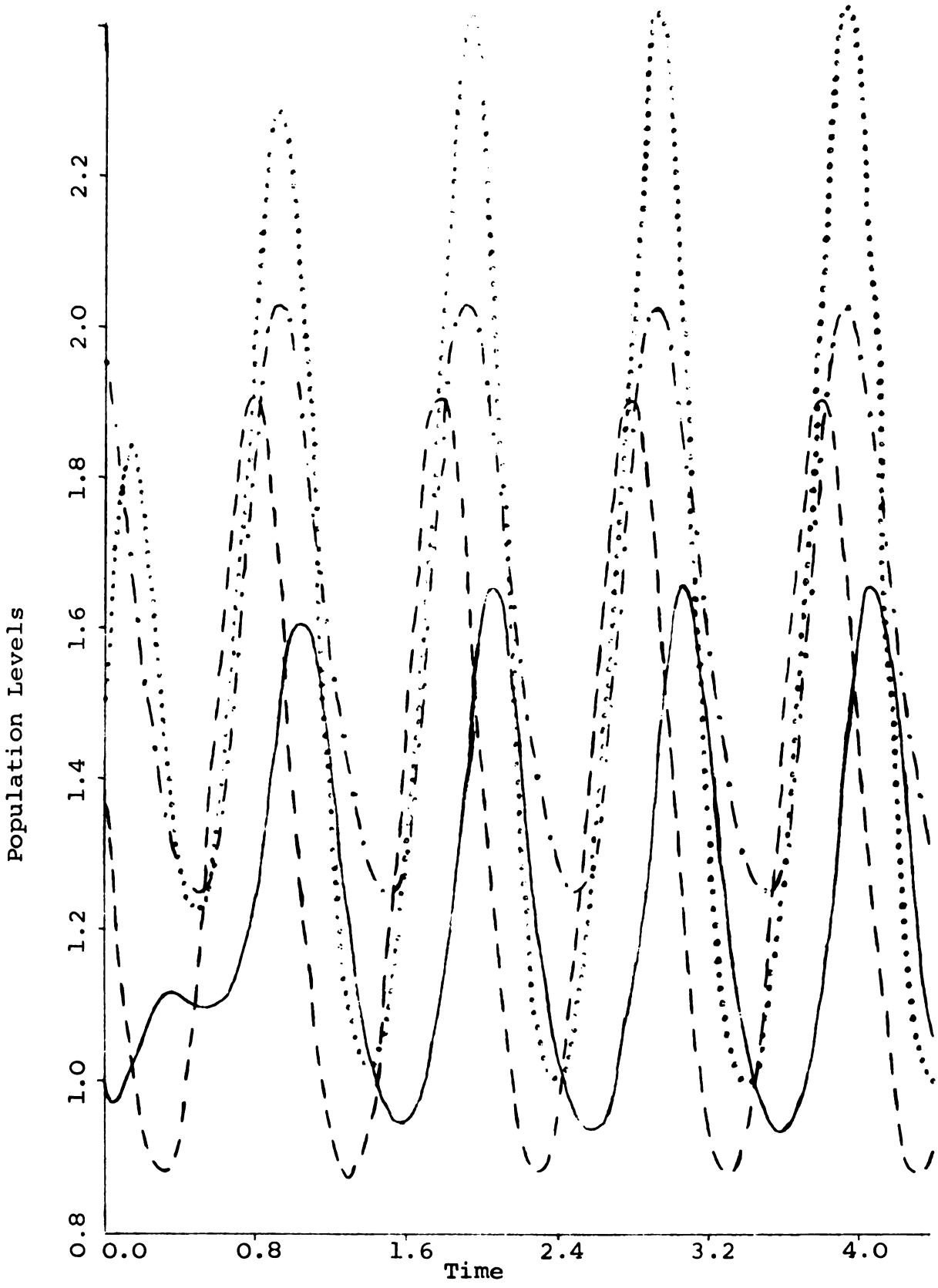

Figure 3. Solution and critical values of equation (83) with $\omega = 2\pi$.

Prey = $p_1(t)$ $\cdots$ Predator = $p_2(t)$ ——

$c_1(t)$ $\cdot - \cdot$ $c_2(t)$ $- -$

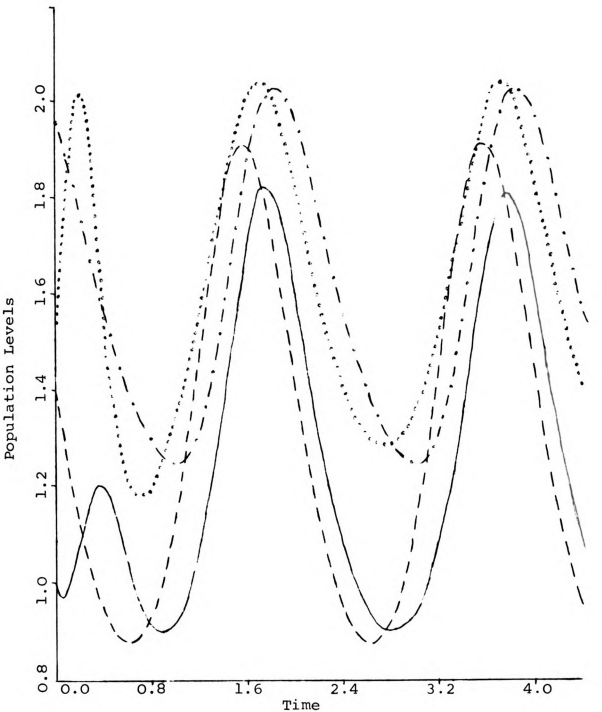

Figure 4. Solution and critical values of equation (83) with $\omega = \pi$.

Prey = $p_1(t)$ ···· Predator = $p_2(t)$ ——

$c_1(t)$ · — · $c_2(t)$ — —

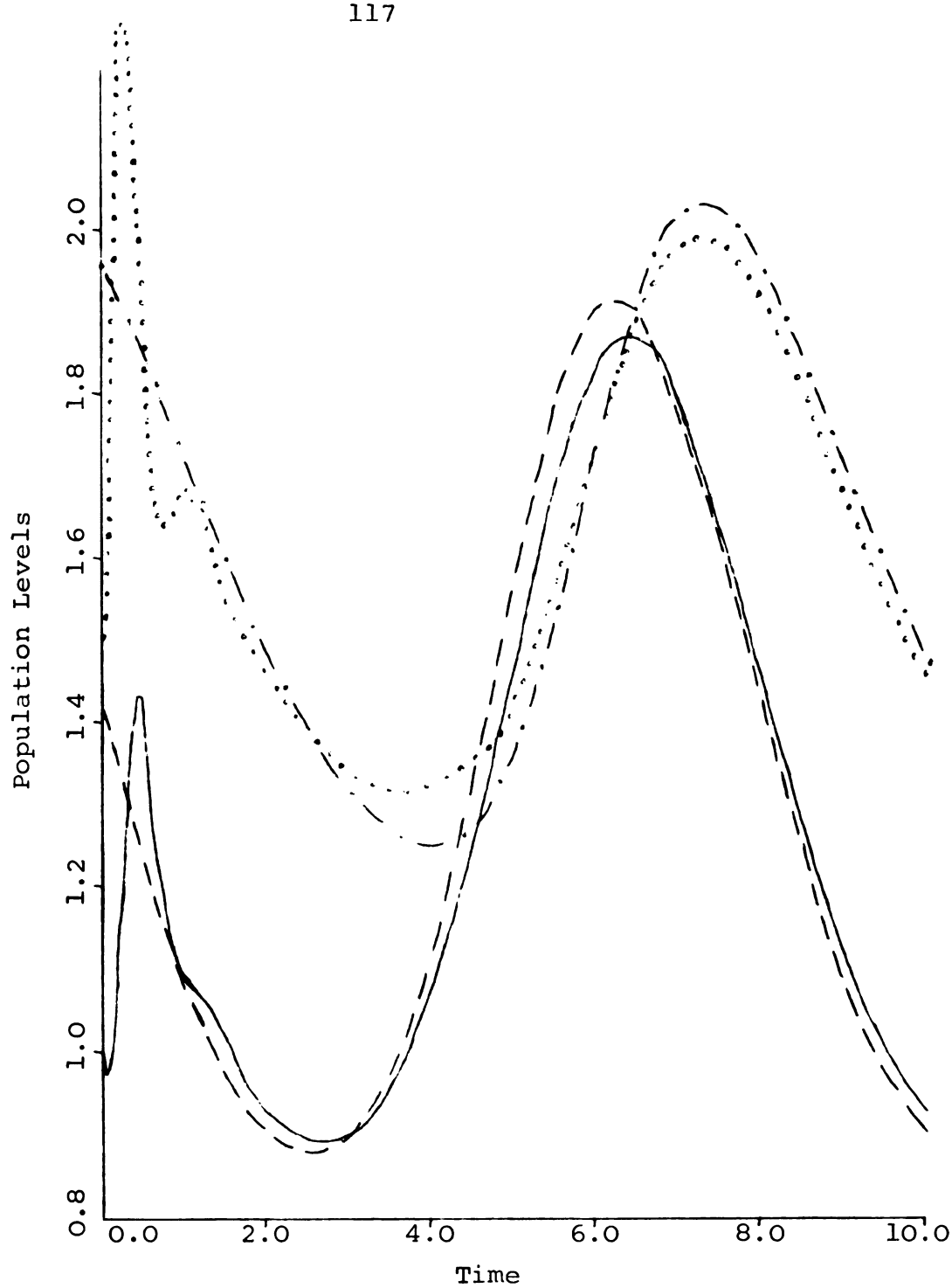

Figure 5. Solution and critical values of equation (83) with $\omega = \frac{\pi}{4}$.

Prey = $p_1(t)$ · · · · Predator = $p_2(t)$ ——

$c_1(t)$ · — · $c_2(t)$ — —

It would certainly be reasonable to model many natural ecosystems with periodic coefficients in equation (63). If the fluctuations are small enough, we can prove the existence of a periodic solution:

Theorem 5.8. Let q,d,L,σ and w be as in Theorem 5.7. Assume that q is positive and that equation (79) holds. If $\rho_i(t)$ and $\beta_{ij}(t)$ in equation (63) are periodic functions of period T for all i and j, then there exists a periodic solution to (63) with $p(0) - c(0) \leq \frac{\sqrt{\sigma}}{2d}(q-w)$.

Proof: Let $F:R^n \to R^n$ be defined by $F(z_0) = z(T)$, where $z(t)$ is the solution to (67) with initial value $z(0) = z_0$. From standard theorems of differential equations, F is well defined and continuous. Let $H = \{z_0 : \sqrt{z_0^T S z_0} \leq \frac{\sqrt{\sigma}}{2d}(q-w)\}$. H is closed and convex. Further F maps $H \to H$, since $(z^T(T)Sz(T))^{\frac{1}{2}} \leq u(T)$, where u is the solution of (82) with $u(0) = z_0 \leq \frac{\sqrt{\sigma}}{2d}(q-w)$, which implies $u(t) \leq \frac{\sqrt{\sigma}}{2d}(q-w)$ for all t.

Schauder's fixed point theorem (see Hale [13]) implies that F has a fixed point, i.e. $\bar{z}(T) = \bar{z}(0)$ for some solution $\bar{z}$ of (67) starting in H. From (64), $c(t)$ is periodic with period T and hence (67) is periodic with period T. From the uniqueness of

solutions with given initial values, it follows that $\bar{z}(t+T) = \bar{z}(t)$ for all t, and therefore $\bar{p}(t) = \bar{z}(t) + c(t)$ is periodic with period T, which completes the proof.

Putting Theorems 5.7 and 5.8 together, we see that there is not only a periodic solution $\bar{p}(t)$ to (63), but that we can expect it to be close to the critical vector $c(t)$ if $c'(t)$ is small. Finally, for any other solution $p(t)$, let $y(t) = p(t) - \bar{p}(t)$, so that

$$y_i' = \frac{\bar{p}_i'(t)}{\bar{p}_i(t)} y_i + \sum_j \bar{p}_i(t)(\alpha_{ij}+\beta_{ij}(t))y_j + \sum_j y_i(\alpha_{ij}+\beta_{ij}(t))y_j. \tag{84}$$

Let

$$A(t) = D(\bar{p}(t))[\mathcal{A} + \mathcal{B}(t)] + \mathrm{Diag}\left[\frac{\bar{p}_i'(t)}{\bar{p}_i(t)}\right], \tag{85}$$

then if there exists a positive definite symmetric matrix S, such that $A^T(t)S + SA(t)$ is negative definite for all t, $\bar{p}(t)$ is asymptotically stable.

CONCLUSIONS

This thesis develops methods to show that a stability matrix remains stable under perturbations. The problem was motivated by the Lotka-Volterra models of ecosystems, and the necessity for such a study to make these models useful in practice was pointed out, but it is clear that the topic is important for any large dynamical system whose parameters can only be estimated. This will usually be the case for models of biological systems. The results will also be important for the field of structural stability, since most of this thesis could be described as a study of the structural stability of linear systems under linear perturbations. The summary of pertinent results from matrix theory in the introduction should provide starting points for further research on the subject. Sensitivity methods based on derivatives of eigenvalues (Section 1.5.2), generalizations of Gershgorin's theorem (Section 1.5.5), and the transformation $A \rightarrow (A+I)(A-I)^{-1}$ (Section 1.5.8), appear the most promising to yield more information about the effect of perturbations on stability.

This thesis has exploited the two classical necessary and sufficient criteria for stability, those of Lyapunov and of Routh and Hurwitz. Of these, the Lyapunov criterion is the best suited to perturbation analysis. Theorems 2.15-2.19 and especially Corollary 2.20 give us sufficient bounds on the perturbations that are computationally feasible, as are the conditions for preserving stability when bordering a matrix in Theorem 2.28.

The convexity of the S-permissible perturbations, Theorem 2.23, is especially useful, since it provides a method to establish open regions in n^2 space that contain only stability matrices. This is what we really need, not just stability under a single perturbation, if we are to establish the stability of a matrix whose entries are only known to lie within certain intervals. Open regions about a matrix A containing only stability matrices can also be established using Gershgorin type theorems (Section 1.5.5). In practice these two methods should complement each other, since the Gershgorin methods are computationally easiest and give the best results when the eigenvalues of A are spread apart, whereas the Lyapunov equation methods are computationally easiest and give the best results when the eigenvalues of A are close together.

Improvements in the Lyapunov equation methods might be obtained by investigating the best choice for Q in the equation (1) of Chapter II. Further investigation of criteria for $B^TS + SB - Q$ in equation (27) to be negative-definite would also be useful, especially when B has rank two or higher.

The Lyapunov equation has many other applications besides perturbation analysis. In fact, we have seen some examples in Chapter V. Other examples can be found in Barnett and Storey [3] and Chapter 5 of May [28]. Thus the improved understanding of the relationship of this equation to the symmetric part of A (Section 2.2), and the iterative procedures to solve it (Chapter III) are valuable in themselves. The block Seidel method appears to be a strong candidate for the best procedure to solve the Lyapunov equation for large matrices (say dimension 10 or higher). While block Seidel and other iterative procedures for linear equations are well known, to the author's knowledge they have not previously been adapted to the solution of the Lyapunov equation. Theorem 3.2, which tells us how well we can expect these procedures to perform on the Lyapunov equation, is important not only for the iteration schemes studied in this thesis but also for any which might be studied in the future.

While the Routh-Hurwitz criterion is often used to prove stability of a particular matrix, Theorem 4.5, employing this criterion and Orlando's theorem to give necessary and sufficient upper and lower bounds for a rank one perturbation to preserve stability of a matrix, is new. Further research in this direction should include extensions of this result to perturbation matrices of higher rank. Perturbations of at least rank 2 will be important for ecosystem analysis, since changes in the strength of an interaction between two species will generally change both a_{ij} and a_{ji}.

After developing all this mathematical machinery, it is disappointing to not find any examples in the ecological literature to apply it to. Certainly an attempt to estimate the coefficients in the Lotka-Volterra model for a natural ecosystem would be valuable. Perhaps the results of this thesis, making it possible to evaluate the stability of the model with only estimates of the coefficients, will make such a study more attractive to the ecologists.

This thesis has contributed to a general understanding of the types of species interactions that contribute to stability. More specific contributions to mathematical ecology include the conditions for

maintaining the stability of the system when adding a new species, and estimating the domain of attraction for an asymptotically stable equilibrium point of a Lotka-Volterra model. We also weakened the hypothesis for the Lyapunov function of Aiken and Lapidus and thereby generalized results both of Aiken and Lapidus and of McArthur, showing that the domain of attraction is the entire first quadrant when the community matrix is D-negative definite. This leads us to hope that the general theorem on domains of attraction (Theorem 5.4) could be improved, and points out the need for further investigation of D-negative definiteness. We have also seen that D-stability is important for ecosystem models, and conjecture that the two may be equivalent. When are all the equilibrium values positive, and whether ecosystems are "trophic-level-stable", are other questions that merit additional research.

The final section, where the coefficients of the general Lotka-Volterra model are made time varying, is the first investigation of this problem. We saw that the solutions follow a moving critical vector and the closeness of the solutions to the critical vector depends on the derivative of the critical vector, and that if the coefficients are periodic there is a periodic solution.

The existence of a periodic solution when the time variations of the coefficients are small enough follows from a general theorem of differential equations (see Hale [13]), but we have shown an explicit bound on the size of the variations that is small enough. Weakening the hypothesis of Theorems 5.7 and 5.8 and investigating the uniqueness and stability of the periodic solutions should be goals of additional research in this area. But this will probably require non-linear techniques, whereas the methods developed in this thesis are essentially based on linearization.

There are still many unanswered questions about the differential equations proposed fifty years ago by Lotka and Volterra as a model for ecosystems. There has recently been a great increase in interest and research about them, but most of it of a theoretical rather than an applied nature. This thesis has helped to answer some of the theoretical questions about stability, and added others to the list of unanswered ones. But its major contribution is to improve the usefulness of the Lotka-Volterra equations for analyzing the stability of actual ecosystems.

BIBLIOGRAPHY

BIBLIOGRAPHY

1. Aiken, R., and Lapidus, L. 1973, The stability of interacting populations, Int. J. Systems Sci. 4, 691-695.

2. Arrow, K.J., and McManus, M. 1958, A note on dynamic stability, Econometrica 26, 448-454.

3. Barnett, S., and Storey, C. 1968, Some applications of the Liapunov matrix equation, J. Inst. Math. App. 4, 33-42.

4. Barnett, S., and Storey, C. 1970, Matrix methods in stability theory, New York: Barnes and Noble.

5. Bellman, R. 1970, Introduction to Matrix Analysis, New York, McGraw-Hill.

6. Brickley, W.G., and MacNamee, J. 1960, Matrix and other direct methods for the solution of systems of linear difference equations, Phil. Trans. Roy. Soc. London, Series A, 252, 69-131.

7. Fox, L. 1964, An introduction to numerical linear algebra, Oxford: Clarendon Press.

8. Frame, J.S. 1963, The rectangular matrix equation AY + B = YC, Notices Am. Math. Soc. 10, 566.

9. Gantmacher, F.R. 1959, Applications of the theory of Matrices, New York: Interscience.

10. Gause, G.F. 1934, The struggle for existence, Baltimore: Williams and Wilkins.

11. Gilpin, M.E., and Justice, K. 1973, A note on non-linear competition models, Math. Biosciences, 17, 57-63.

12. Hahn, W. 1967, Stability of motion, New York: Springer Verlag.

13. Hale, J.K. 1969, Ordinary differential equations, New York: Interscience.

14. Hardin, G. 1969, Not peace, but ecology. In Diversity and stability in ecological systems, Brookhaven Symposium in Biology, No. 22, Springfield VA: Nat. Bureau of Standard, U.S. Dept. of Commerce, p.151.

15. Holling, C.S. 1969, Stability in ecological systems. In Diversity and stability in ecological systems, op. cit., p.128.

16. Holling, C.S. 1966, The functional response of invertebrate predators to prey density, Mem. Entomol. Soc. Canada, 48.

17. Householder, A. 1953, Principles of numerical analysis, New York: McGraw-Hill.

18. Isaacson, E., and Keller, H. 1966, Analysis of numerical methods, New York: John Wiley and Sons.

19. Jameson, A. 1968, Solution of the equation $AX + XB = C$ by inversion of an $M \times M$ or $N \times N$ matrix, SIAM J. Appl. Math. 16, 1020-1023.

20. Kalman, R.E., and Bertram, J.E. 1960, Control system analysis and design via the "second method" of Liapunov, trans. A.S.M.E.J. Basic Engng., 82D, 371-400.

21. Kerner, E.H. 1957, A statistical mechanics of interacting biological species, Bull. Math. Biophys., 19, 121-146.

22. LaSalle, J.P. 1960, The extent of asymptotic stability, Proc. Nat. Acad. Sci. U.S.A., 46, 363-365.

23. Levins, R. 1968, Evolution in a changing environment, Princeton: Princeton University Press.

24. Lotka, A. 1925, Elements of physical biology, Baltimore: Williams and Wilkins. (Reissued as Elements of mathematical biology, Dover, 1956.)

25. Lyapunov, A. 1907, Problème general de la stabilité du mouvement, Annales de la Faculté des science de Toulouse, second series, 9. (Reissued in Annals Math. Studies, 19, Princeton: Princeton University Press, 1947.)

26. MacArthur, R.H. 1969, Species packing, or what competition minimizes, Proc. Nat. Acad. Sci., 64, 1369-1375.

27. MacArthur, R.H. 1970, Species packing and competitive equilibrium for many species, Theor. Pop. Biol., 1, 1-11.

28. May, R.M. 1973, Stability and complexity in model ecosystems, Princeton: Princeton University Press.

29. Maybee, J., and Quirk, J. 1969, Qualitative problems in matrix theory, SIAM Review, 11, 30-51.

30. Montroll, E., and Goel, N. 1971, On the Volterra and other nonlinear models of interacting populations Rev. of Modern Phys., 43, No. 2.

31. Porter, B., and Crossley, R. 1972, Model control, theory and applications, New York: Barnes and Noble.

32. Quirk, J., and Ruppert, R. 1965, Qualitative economics and stability of equilibrium, Rev. Economic Studies, 32, 311-326.

33. Rosen, R. 1970, Dynamical system theory in biology, New York: Interscience.

34. Schwarz, H.R. 1956, Ein Verfahren zur Stabilitätsfrage bei Matrizen-Eigenverte-probleme, Z. Angew. Math. Phys., 7, 473-500.

35. Smith, R.A. 1966, Matrix calculations for Liapunov quadratic forms, J. Diff. Eqns., 2, 208-217.

36. Strobeck, C. 1973, N species competition, Ecology, 54, 650-654.

37. Taussky, O. 1961, A remark on a theorem of Lyapunov, J. Math. Anal. Appl., 2, 105-107.

38. Tomovic, R. 1972, General sensitivity theory, New York: American Elsevier.

39. Vogt, W.G. 1965, Transient response from the Liapunov stability equation, Preprints Joint Automatic Control Conf., Paper V4, 23-30.

40. Volterra, V. 1926, Variazioni e fluttuazioni del numero d'individua en specie animali conviventi. (Translation by M.E. Wells in R.N. Chapman's Animal Ecology, New York: McGraw-Hill, 1931, pp.409-448.)

41. Volterra, V. 1931, Leçons sur la théorie mathematique de la lutte pour la vie, Paris: Gauthier-Villars.

42. Westlake, J. 1968, A handbook of numerical matrix inversion and solution of linear equations, New York: John Wiley and Sons.

43. Wilkinson, J.H. 1965, The algebraic eigenvalue problem, Oxford: Clarendon Press.

44. Yoshizawa, T. 1966, Stability theory by Liapunov's second method, Tokyo: The Mathematical Society of Japan.