Technical Report 67-54 October 1967

Spectral Properties of Matrices Which Have Invariant Cones

By

James S. Vandergraft

Assistant Professor

Computer Science Center

This work was supported in part by grant NsG 398 of the National Aeronautics and Space Administration to the University of Maryland.

Abstract

The Perron-Frobenius theory of non-negative matrices has been found to be of great value in the study of iterative processes. In this paper, we extend much of this theory to include those matrices which leave invariant a closed convex cone with non-empty interior. Such matrices are completely characterized in terms of certain spectral properties. The notion of irreducibility is generalized to these matrices, and several theorems are proved to show that this is a suitable extension of the classical concept. The basic results of the Perron-Frobenius theory are then extended, and several useful comparison theorems for iterative processes are derived.

Spectral Properties of Matrices

Which Have Invariant Cones[1)]

by

James S. Vandergraft[2)]

1. Introduction.

The basic results of the Perron-Frobenius theory of non-negative matrices [2] are:

I. If A is an nxn matrix with non-negative elements ($A \geq 0$), then a) $\rho(A)$ is an eigenvalue, b) there is a corresponding eigenvector which is non-negative, and c) if $B \geq A$, i.e. $B-A \geq 0$, then $\rho(B) \geq \rho(A)$ where $\rho(a)$ is the spectral radius of A.

II. If $A \geq 0$ and irreducible[3)] then a) $\rho(A)$ is a simple eigenvalue, b) there is a corresponding eigenvector which is positive, and c) if $B \geq A$ and $B \neq A$ then $\rho(B) > \rho(A)$.

III. If A has all positive elements ($A > 0$), then a) $\rho(A)$ is a simple eigenvalue, greater than the magnitude of any other eigenvalue, b) properties IIb and IIc hold.

The existence statments in these theorems (Ia, Ib, IIa, IIb, IIIa) have been generalized to operators on a Banach space which leave a cone invariant. (See [5], [6], [8]). The other statements, i.e., those which compare the spectral radii of two matrices,

---

1) This work was supported in part by grant NsG 398 of the National Aeronautics and Space Administration to the University of Maryland.

2) Computer Science Center, University of Maryland.

3) The matrix A is irreducible if no permutation matrix P exists such that $P^TAP = \begin{bmatrix} B & C \\ O & D \end{bmatrix}$ where B and D are square, and O is a block of zeros.

have not been adequately generalized in spite of their usefulness in the study of iterative processes. Irreducibility, which is important to this part of the theory, is usually replaced by a stronger condition. Recently, for example, Marek [7] proved some comparison-type theorems for Krasnoselskii's u_o-positive operators. It can be shown (see section 4) that, in the finite dimensional case, all non-negative u_o-positive operators are irreducible, but not conversely. Shaeffer [8] has defined a class of operators on a Banach space, which, in E^n becomes the class of irreducible matrices; however, his results are only of the existence type.

The generalizations referred to above, involve two extensions of the classical theory. The spaces are usually assumed to be infinite dimensional, and positivity is replaced by the assumption that the operator leaves a cone invariant. In this paper, we will retain the latter extension but, in order to obtain stronger results, will consider only finite dimensional spaces. Our first theorem gives necessary and sufficient conditions for a matrix to leave a cone invariant. We then extend the notion of irreducibility and prove the corresponding Perron-Frobenius type theorems. In the final section, these results are used to generalize the comparison theorems of Stein-Rosenberg and Fiedler-Ptak.

I would like to thank Professor Werner Rheinboldt, of the University of Maryland, who suggested several of the ideas developed here, and aided in their development by frequent discussions.

2. Cones in E^n.

We begin with a brief discussion of convex cones in finite dimensional spaces. In this paper, we will define a cone to be a closed subset K of E^n which satisfies $K \cap (-K) = \{0\}$, $K + K = K$, and $\alpha K \subset K$ for any $\alpha \geq 0$. A cone K is solid if its interior K^o is non-empty, and K is reproducing if $E^n = K - K$. (In E^n, every solid cone is reproducing, and conversely.) If $\chi \in K$ we sometimes write $\chi \overset{K}{\geq} 0$, and $\chi \geq 0$ will mean that χ is in the cone consisting of all vectors in E^n with non-negative coordinates. If A is an n x n matrix, then we write $A \overset{K}{\geq} 0$ if $A\chi \in K$ whenever $\chi \in K$. If K is solid then $\chi \overset{K}{>>} 0$ means $\chi \in K^o$, and similarly, $A \overset{K}{>>} 0$ implies $A\chi \overset{K}{>>} 0$ whenever $\chi \in K$ and $\chi \neq 0$. An important fact about elements in K^o is that, if $\chi \overset{K}{>>} 0$, and $y \in E^n$, then for some $\lambda > 0$, $y \overset{K}{\leq} \lambda\chi$. (see [9].) A vector $\chi \in K$ is called extremal if $\chi = y+z$ with $y \in K$ and $z \in K$ implies that both y and z are non-negative multiples of χ. A cone K is generated by a set of vectors if any element in K can be written as a finite linear combination of these vectors, using only non-negative coefficients. Our first result shows the connection between these last two concepts.

Theorem 2.1 Any cone in E^n is generated by its extremal vectors.

Proof. By induction on n: For n = 2, the result is obvious. Suppose the theorem is true for spaces of dimension less than n. If χ is an interior point of $K \subset E^n$, then let $u \in K$ be linearly independent of χ, and let H be the plane spanned by χ and u. Then $\tilde{K} = H \cap K$ is a cone, with $\chi \in \tilde{K}$, so $\chi = \chi_1 + \chi_2$ where χ_1, χ_2 are on the boundary of $\tilde{K}$, and hence on the boundary of K. Thus, to prove the theorem, we need only consider points on the boundary δK of K. Let $\chi \in \delta K$ be arbitrary. We can assume χ is not extremal, in which case $\chi = u+v$, where u and v are linearly independent of χ. If u or v are in K^o, then so is χ, hence u, v $\in \delta K$. The set $\{\alpha u + \beta v \mid \alpha, \beta \geq o\}$ is a cone which contains χ and is contained in δK. (we use here the fact that, if $y \in \delta K$ and $o \stackrel{K}{\leq} \chi \stackrel{K}{\leq} y$ then $\chi \in \delta K$.) Let S be the largest cone such that $\chi \in S \subset \delta K$. If H_s is the smallest linear subspace containing S, then clearly the dimension of H_s is less than n. We can now apply the induction hypothesis, and the proof is complete, provided that all extremal vectors of S are also extremal vectors of K. But, suppose $y \in S$ is not an extremal vector of K. Then $y = u + v$, where $u, v \in K$. Suppose $u \in S$ and consider the cone

$$S' = \{ w + \alpha u \mid w \in S, \alpha \geq 0\}.$$

Since

$$w + \alpha u = w + \alpha(y-v) \stackrel{K}{\leq} w + \alpha y \in S \subset \delta K$$

if follows that $S' \subset \delta K$. Since $\chi \in S'$ and S' is larger than S, we have a contradiction to the definition of S. Hence, $u \in S$, and similarly, $v \in S$, so y is not extremal in S.

A <u>subcone</u> of K is any cone contained in K, and an <u>extremal subcone</u> is a subcone which is generated by some subset of the extremal vectors of K. If an extremal subcone is contained in the boundary, δK, of K then it is called a <u>face</u> of K. If F is any face, then it will be contained in a linear subspace of dimension less than n. The smallest such subspace will be denoted by H_F. To every $\chi \in \delta K$ there corresponds a particular face which has several useful properties. These are described by the next lemma.

<u>Lemma 2.1</u> Given any $\chi \in \delta K$, there exists a face F_χ such that

i) $\chi \in F_\chi^o$, relative to the space H_{F_χ}

ii) $F_\chi = \delta K \cap H_{F_\chi}$

iii) $o \stackrel{K}{\leq} y \stackrel{K}{\leq} \chi$ implies $y \in F_\chi$.

<u>Proof</u>. By theorem 2.1, any $\chi \in \delta K$ can be written as

$$\chi = \sum_{i=1}^{n} \gamma_i \chi_i$$

where $\gamma_i > 0$, χ_i is extremal, $i = 1, 2, \ldots, n$. The cone generated by $\chi_1, \ldots, \chi_n$ is a face which satisfies part i). Let F_χ be the largest such face. Then ii) is also true since obviously $F_\chi \subset \delta K \cap H_{F_\chi}$, and the cone $\delta K \cap H_{F_\chi}$ is a face which satisfies i) so, in fact, we must have $F_\chi = \delta K \cap H_{F_\chi}$. Finally, if $0 \leq^K y \leq^K \chi$ then $y \in \delta K$. Suppose $y \notin H_{F_\chi}$. Then, let H' be the subspace spanned by H_{F_χ} and y. If $F' = \delta K \cap H'$ then F' is a face, and $\chi - y \in F'$ so χ is interior to F', relative to H'. This contradicts the definition of F_χ, so $y \in H_{F_\chi}$ and, by ii) it follows that $y \in F_\chi$.

If χ_o is an extremal vector of K, then clearly $\lambda \chi_o$ is also extremal, for any $\lambda \geq 0$. Hence, when referring to the number of extremal vectors, we will consider only distinct vectors which are normalized in some sense. (For example, we might assume their euclidean norms are equal to some constant.) A cone with a finite number of extremal vectors (in the above sense) is called <u>polyhedral</u>. If a cone contains n linearly independent vectors $\chi_1, \ldots, \chi_n$, then the vector $\chi_1 + \chi_2 + \ldots + \chi_n$ is interior to the cone. Hence, a cone is solid if and only if it has n linearly independent extremal vectors. A solid polyhedral cone which has exactly n extremal vectors is called <u>simplicial</u>.

3. Matrices and Invariant Cones.

It follows from the theory of invariant cones in a Banach space ([5], [6]) that if a matrix A leaves invariant a solid cone, then $\rho(A)$ is an eigenvalue, and a corresponding eigenvector lies in the cone. This result can also be proved directly, using the Brouwer fixed point theorem. However, Birkhoff [1] has given an elementary proof of this result which uses instead the Jordan Canonical form. The advantage of Birkhoff's proof is that it can be extended to prove a further property of $\rho(A)$, which turns out to be a sufficient condition for A to leave a cone invariant. In order to state this condition, we need the following definition.

Definition 3.1. If λ is an eigenvalue of a matrix A, then the degree of λ is the size of the largest diagonal block, in the Jordan canonical form of A, which contains λ.

Theorem 3.1. If K is a solid cone, and $A \overset{K}{\geq} 0$, then

i) $\rho(A)$ is an eigenvalue,

ii) the degree of $\rho(A)$ is no smaller than the degree of any other eigenvalue having the same modulus,

iii) K contains an eigenvector corresponding to $\rho(A)$.

Furthermore, conditions i) and ii) are sufficient to insure that A leaves invariant a solid cone.

Proof. Birkhoff's proof gives the necessity of conditions i) and iii). We will sketch his proof in order to show how it can be extended to prove ii). Let $\{x_{ij}\}$ be a linearly independent set of vectors which satisfy

$$(3.1) \qquad A x_{ij} = \lambda_i x_{ij} + x_{ij-1} \qquad i = 1,\dots,k, \quad j = 1,\dots,m_i$$

$$x_{io} = 0, \quad \sum_{i=1}^{k} m_i = n,$$

where λ_i are eigenvalues of A. Since $A x_{i1} = \lambda_i x_{i1}$, the vectors

$\{\chi_{i1}\}$ are eigenvectors, and hence may be complex, in which case they occur in conjugate pairs. The same is true of the principle vectors χ_{ij}, $j > 1$. These vectors form a basis for E^n in the sense that any element $Y \in E^n$ can be written as

$$Y = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \alpha_{ij}\, \chi_{ij}, \qquad \alpha_{ij} = \bar{\alpha}_{pq} \text{ if } \chi_{ij} = \bar{\chi}_{pq} \tag{3.2}$$

We assume the eigenvalues satisfy

$$|\lambda_1| = |\lambda_2| = \ldots = |\lambda_\nu| > |\lambda_{\nu+1}| \geq \ldots |\lambda_k|.$$

By induction, it can be proven that

$$A^r \chi_{ij} = \sum_{p=0}^{j-1} \lambda_i^{r-p} \binom{r}{p} \chi_{ij-p}$$

where $\binom{r}{p}$ is the binomial coefficient, and hence is a polynomial in r of degree p.

Thus, if $Y = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \alpha_{ij}\, \chi_{ij}$ is any element in E^n,

$$A^r Y = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \alpha_{ij} \sum_{p=o}^{j-1} \lambda_i^{r-p} \binom{r}{p} \chi_{ij-p}$$

$$= \sum_{i=1}^{k} \sum_{j=1}^{m_i} \left(\sum_{p=o}^{m_i-j} \alpha_{ij+p} \binom{r}{p} \lambda_i^{r-p} \right) \chi_{ij} \tag{3.3}$$

The sequence $\left\{ \frac{A^r Y}{\|A^r Y\|} \right\}$ is bounded, so there is a convergent subsequence. From (3.3) it follows that the limit of this subsequence must have the form

$$Y^* = \sum_{i \in d} \beta_i\, \chi_{i1} \tag{3.4}$$

where

$$d = \{ i \leq \nu \mid m_i \geq m_j,\ j=1,\ldots,\nu \}.$$

Since K is solid, there is an element in K of the form

$$x_o = \sum_{i=1}^{K} \sum_{j=1}^{m_i} \alpha_{ij} x_{ij} \quad , \qquad \alpha_{ij} \neq 0.$$

and since A maps K into itself, $A^r x_o \in K$, and the above argument shows that K contains an element of the form

$$(3.5) \qquad x^* = \sum_{i \in d} \beta_i x_{i1} \ .$$

Now, suppose for some $i_o \in d$, λ_{i_o} is not positive. An elementary lemma says that there is a finite set of positive numbers $w_o, \ldots, w_q$, such that

$$\sum_{p=0}^{q} w_p \ \lambda_{i_o}^{p} = 0$$

Let

$$x' = \sum_{p=0}^{q} w_p A^p x^* = \sum_{p=0}^{q} w_p \left(\sum_{i \in d} \beta_i \lambda_i^p x_{i1} \right)$$

$$= \sum_{i \in d} \beta'_i x_{i1}$$

Then

$$\beta'_i = \beta_i \sum_{p=0}^{q} w_p \lambda_i^p$$

and hence,

$$\beta'_{i_o} = 0.$$

Thus, given any element of the form (3.5) in K, if $\beta_{i_o} \neq 0$ then either $\lambda_{i_o} > 0$, or we can find another <u>non-zero</u> element in K, with $\beta_{i_o} = 0$. Repeating this process, we finally get an element in K of the form (3.5), with $\beta_i \neq 0$ only if $\lambda_i = \rho(A)$. This element is an eigenvector, with eigenvalue $\rho(A)$, which completes Birkhoff's proof. To prove ii) we note that, if this statement were false, then λ_i would be non-positive, for every $i \in d$. Thus, by the above construction, we would be able to produce a <u>non-zero</u> element in K, of the form (3.5), in which all the coefficients β_i are zero. This contradiction proves ii). To prove the final statement of the theorem, let $\lambda_1 = \rho(A)$, and normalize the x_{ij} so that (3.1) becomes

$$Ax_{ij} = \lambda_i x_{ij} + \mathcal{E}\, x_{ij-1}, \quad i = 1, \ldots, k, \; j = 1, \ldots, m_i \tag{3.6}$$

where $\mathcal{E} = 1$ if $\nu = \kappa$, otherwise

$$\mathcal{E} = \lambda_1 - |\lambda_{\nu+1}|$$

We assume

$$m_1 \geq m_i, \; i = 1, 2, \ldots, \nu$$

and will show that the set

(3.7)

$$K = \left\{ x \in E^n \,\middle|\, x = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \alpha_{ij}\, x_{ij}, \; \begin{array}{l} |\alpha_{ij}| \leq \alpha_{1j}, \; j \leq m_1 \\ |\alpha_{ij}| \leq \alpha_{1m_1}, \; j \geq m_1 \\ \alpha_{ij} = \bar{\alpha}_{pq} \text{ if } x_{ij} = \bar{x}_{pq} \end{array} \right\}$$

is a solid invariant cone. Clearly K is a cone. To show solidness, let $Y = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \gamma_{ij}\, x_{ij}$ be an arbitrary element in E^n. Then

(3.8)

$$Y = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \beta_{ij}\, x_{ij} - \sum_{j=1}^{m_1} \delta_j\, x_{1j}$$

where

$$\beta_{ij} = \gamma_{ij}, \quad i \neq 1$$

$$\beta_{1j} = \begin{cases} \max\left(|\gamma_{1j}|, \; \{|\gamma_{ij}|, \; m_i \leq j\}\right), & j \leq m_1 \\ \max\left(|\gamma_{1,m_1}|, \; \{|\gamma_{ij}|, \; m_i \leq j\}\right), & j \geq m_1 \end{cases}$$

$$\delta_j = \beta_{1j} - \gamma_{1j}$$

This choice of β_{ij} and δ_j insures that both terms in (3.8) are in K, hence K is reproducing. Since any reproducing cone in E^n is also solid, our assertion is true. To show that K is invariant, let

$$X = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \alpha_{ij}\, x_{ij} \in K$$

and,

$$AX = \sum_{i=1}^{k} \sum_{j=1}^{m_i} \beta_{ij}\, x_{ij}$$

By (3.6) we have

$$\beta_{ij} = \begin{cases} \alpha_{ij}\ \lambda_i + \varepsilon\ \alpha_{i,j+1} & j < m_i \\ \alpha_{ij}\ \lambda_i & j = m_i \end{cases}$$

Obviously $\beta_{ij} = \bar{\beta}_{pq}$ if $\chi_{ij} = \bar{\chi}_{pq}$, so we need only prove

$$|\beta_{ij}| \leqq \beta_{1j} \qquad j \leqq m_1$$
$$|\beta_{ij}| \leqq \beta_{1,m_1} \qquad j \geqq m_1$$

Consider the various cases:

$\underline{j < m_1}$

$$j < m_i \ :\ |\beta_{ij}| = |\alpha_{ij}\ \lambda_i + \varepsilon\ \alpha_{i,j+1}| \leqq \alpha_{1j}\ \lambda_1 + \varepsilon\ \alpha_{1,j+1} = \beta_{1j}$$
$$j = m_i \ :\ |\beta_{ij}| = |\alpha_{i,m_i}\ \lambda_i| \leqq \alpha_{1,m_i}\ \lambda_1 + \varepsilon\ \alpha_{1,m_{i+1}} = \beta_{1,m_i}$$

$\underline{j \geqq m_1}$

$j < m_i$: by (3.6) $|\lambda_i| \quad \lambda_1 - \varepsilon$, so

$$|\beta_{ij}| = |\alpha_{ij}\ \lambda_i + \varepsilon\ \alpha_{i,j+}\ |$$
$$\leqq \alpha_{1m_1}\ (\lambda_1 - \varepsilon) + \varepsilon\ \alpha_{1m_1} = \alpha_{1m_1}\ \lambda_1 = \beta_{1m_1}$$

$$j = m_i \ :\ |\beta_{ij}| = |\alpha_{ij}\ \lambda_i| \leqq \alpha_{1j}\ \lambda_1 = \beta_{1j}$$

Thus, $AX \in K$ and the result is proved.

An interesting corollary of this theorem is that, if A is a symmetric matrix, then either A or -A leaves some cone invariant.

4. Irreducibility.

Before generalizing the notion of irreducibility, we will give an alternative definition, which emphasizes the geometric nature of this concept. If $e_1, \ldots, e_n$ are the unit coordinate vectors in E^n, then a coordinate subspace is a subspace spanned by any subset of $\{e_1, \ldots, e_n\}$. An irreducible matrix is a matrix which has no invariant coordinate subspace of dimension less than n. Since the positive hyperoctant is generated by the vectors $e_1, \ldots, e_n$, a non-negative irreducible matrix maps the positive hyperoctant into itself and leaves no face invariant. It is clear that this definition is equivalent to that given in section 1. (Gantmacher [4] and others use this as a basic definition.)

If we replace the positive hyper-octant by an arbitrary solid cone, the above definition leads to the following generalization.

Definition 4.1: The matrix $A \overset{K}{\geqq} 0$ is K-irreducible if A leaves no face of K invariant. A matrix which is not K-irreducible is called K-reducible.

To further justify this definition, we will prove several properties of K-irreducible matrices which are known to be true for non-negative irreducible matrices. We assume always that K is a solid cone.

Theorem 4.1: $A \overset{K}{\geqq} 0$ is K-irreducible if and only if no eigenvector of A lies on the boundary of K.

Proof: Suppose $A \geqq 0$ is K-reducible, and let F be an invariant face of K. A, restricted to the subspace H_F, leaves the solid cone F invariant; hence this restricted operator has an eigenvector $\chi_1 \in F$. But χ_1 is also an eigenvector for A, operating on the entire space and χ_1 is on the boundary of K. Conversely, suppose χ is an eigenvector on the boundary of K, and let F_χ be the face defined in lemma 2.1. Then, for any $Y \in F_\chi$, there exists an $\alpha > 0$, such that $Y \overset{F_\chi}{\leqq} \alpha\chi$ and hence $Y \overset{K}{\leqq} \alpha\chi$. Thus, $AY \overset{K}{\leqq} A(\alpha\chi) = \alpha\lambda\chi \in F_\chi$ so by lemma 2.1, $Ay \in F_\chi$. Thus F_χ is an invariant face and A is K-reducible.

The next lemma gives an interesting property of matrices with invariant cones, and allows us to prove another spectral characterization of K-irreducible matrices.

Lemma 4.1: If $A \overset{K}{\geqq} 0$ has two eigenvectors in K°, then A also has an eigenvector on the boundary of K. Furthermore, the corresponding eigenvalues are all equal.

Proof: Let $\chi_1, \chi_2 \in K^\circ$ be linearly independent eigenvectors, with eigenvalues λ_1, λ_2 and let

$$t_o = \min \{t > 0 \mid t\chi_2 - \chi_1 \in K\}$$

where we assume $0 \leqq \lambda_2 \leqq \lambda_1$. If $\chi_3 = t_o\chi_2 - \chi_1$ then χ_3 is on the boundary of K, and if $\lambda_1 \neq 0$, then

$$A\chi_3 = t_o\lambda_2\chi_2 - \lambda_1\chi_1 = \lambda_1\left\{t_o \frac{\lambda_2}{\lambda_1}\chi_2 - \chi_1\right\} \in K.$$

The definition of t_o implies $\lambda_2 \geqq \lambda_1$, hence, in fact, $\lambda_1 = \lambda_2$. If $\lambda_1 = 0$, then $\lambda_2 = 0$, and $A\chi_3 = 0$. In either case, χ_3 is an eigenvector on the boundary of K with eigenvalue $\lambda_1 = \lambda_2$.

The proof of the next theorem follows easily from the two previous results.

Theorem 4.2: $A \overset{K}{\geqq} 0$ is K-irreducible if and only if A has exactly one eigenvector in K, and this eigenvector is in K°.

Note that our concept of K-irreducibility depends on both the matrix and the cone. It is possible for a matrix to leave two cones invariant, but be K-irreducible with respect to only one. An example of this is the matrix

$$A = \begin{bmatrix} 2 & 0 \\ 0 & 1 \end{bmatrix}$$

If $\quad K_1 = \left\{(x,y) \mid x \geqq 0,\ y \geqq 0\right\}$

and $\quad K_2 = \left\{(x,y) \mid x \geqq 0,\ |y| \geqq x\right\}$

then $A \overset{K_i}{\geqq} 0$ and A is K_2-irreducible, but not K_1-irreducible.

Frobenius introduced the class of non-negative irreducible matrices because it is larger than the class of positive matrices, but retains many of the important spectral properties. Clearly, if A maps K into its interior, it can leave no face invariant, hence it is K-irreducible. That is, the class of K-irreducible matrices is larger than the class of matrices which satisfy $A \overset{K}{>>} 0$. Similarities in certain spectral properties of these two types of matrices are pointed out by the next two theorems, which generalize IIa,b and IIIa,b of section 1.

<u>Theorem 4.3</u>: If $A \overset{K}{\geqq} 0$ is K-irreducible then

i) $\rho(A)$ is a simple eigenvalue, and any other eigenvalue with the same modulus is also simple.

ii) There is an eigenvector corresponding to $\rho(A)$ in K^{o}, and no other eigenvector lies in K.

Furthermore, i) is sufficient for A to be K-irreducible with respect to some invariant solid cone.

<u>Proof</u>: Part i) follows from theorem 3.1, provided $\rho(A)$ is simple, and ii) is a restatement of theorem 4.2. Suppose $\rho(A)$ is not simple. Then, there exist vectors x_1, x_2, linearly independent, with $x_1 \in K^{o}$, $Ax_1 = \rho(A)x_1$ and either

(4.1) $$Ax_2 = \rho(A)x_2$$

or

(4.2) $$Ax_2 = \rho(A)x_2 + x_1$$

If (4.1) were true then, for large enough $t > 0$, $\chi_3 = t\chi_1 + t\chi_2 \in K$ and χ_3 is another eigenvector, contradicting theorem 4.2. If equation (4.2) holds, then $-\chi_2 \notin K$ and we can define

$$t_o = \min \{t > 0 | \ t\chi_1 - \chi_2 \in K\}.$$

Then, $\rho(A) = 0$ implies $A(t_o\chi_1 - \chi_2) = -\chi_1 \notin K$, and $\rho(A) \neq 0$ implies

$$\begin{aligned} A(t_o\chi_1 - \chi_2) &= t_o \ \rho(A)\chi_1 - \rho(A)\chi_2 - \chi_1 \\ &= \rho(A) \left\{\left(t_o - \frac{1}{\rho(A)}\right)\chi_1 - \chi_2\right\} \in K \end{aligned}$$

which contradicts the definition of t_o. Hence, $\rho(A)$ must be simple. To prove the last statement of the theorem, we use the proof of theorem 3.1. The cone K defined in that proof contains only elements of the form $\alpha\chi_1 + y$, where χ_1 is the eigenvector corresponding to $\rho(A)$ and $\alpha = 0$ only if $y = 0$. Hence, no other eigenvector can lie in K and by theorem 4.2, A is K-irreducible.

By replacing the assumption that A is K-irreducible by the condition that $A >\stackrel{K}{>} 0$, it is possible to make a stronger statement about $\rho(A)$.

<u>Theorem 4.4</u>: If $A >\stackrel{K}{>} 0$ then

i) $\rho(A)$ is a simple eigenvalue, greater than the magnitude at any other eigenvalue

ii) An eigenvector corresponding to $\rho(A)$ lies in K.

Furthermore, condition i) is sufficient for A to map some cone into its interior.

<u>Proof</u>: Most of the theorem follows as a corollary to the previous result. In fact, we need only prove the last part of i) and the final statement. Let λ_2 be any eigenvalue different from $\rho(A)$, with eigenvector χ_2, and suppose $|\lambda_2| = \rho(A)$. For simplicity, we assume $\rho(A) = 1$, in which case $\lambda_2 = e^{i\theta}$ for some θ. We will show that for some φ, $\mathrm{Re}(e^{i\varphi}\chi_2) \in K$ and from this, we will obtain a contradiction. For any φ, either $\mathrm{Re}(e^{i\varphi}\chi_2) \in K$, or else we can define a positive number t_φ by

$$t_\varphi = \min \left\{t > 0 | \ t\chi_1 + \mathrm{Re}(e^{i\varphi}\chi_2) \in K\right\}$$

where $\chi_1 \in K^\circ$ is the eigenvector corresponding to $\rho(A)$.
If $y_\varphi = t_\varphi \chi_1 + \mathrm{Re}(e^{i\varphi}\chi_2)$, then y_φ is on the boundary of K, and $Ay_\varphi = t_\varphi \chi_1 + \mathrm{Re}(e^{i(\varphi+\theta)}\chi_2) \in K^\circ$. Hence $t_\varphi > t_{\varphi+\theta}$ and thus $\inf_\varphi \{t_\varphi\} = 0$. From this, it follows that for some φ_o,

$$y_o = \mathrm{Re}(e^{i\varphi_o}\chi_2) \in K.$$

Now, if $\{\xi_k\}$ is any finite set of positive numbers, then $\Sigma \xi_k A^k y_o = 0$ implies $y_o = 0$. But, by a basic lemma, which was also used in the proof of theorem 3.1, if $\theta \not\equiv 0 \pmod{2\pi}$ then there exists a finite set of positive numbers $\{\xi_k\}$ such that

$$\Sigma \xi_k \, e^{ik\theta} = 0.$$

Hence,

$$\sum_k \xi_k A^k y_o = \Sigma \xi_k \mathrm{Re}(e^{ik\theta} e^{i\varphi_o}\chi_2)$$
$$= \mathrm{Re}(\Sigma \xi_k e^{ik\theta} e^{i\varphi_o}\chi_2) = 0$$

so, $y_o = 0$, i.e., $e^{i(\varphi_o+\pi)}\chi_2 = y_2$ where y_2 is real. Since $Ay_2 = \lambda_2 y_2$, where $|\lambda_2| = 1$, clearly $\lambda_2 = \pm 1$. We can assume $y_2 \notin K$, and if we let

$$t_o = \min \{t > 0 | \; t\chi_1 + y_2 \in K)$$

then

$$A^2(t_o\chi_1 + y_2) = t_o\chi_1 + y_2 \in K^\circ$$

which contradicts the definition of t_o. Hence $|\lambda_2| < \rho(A)$. To prove the last statement of the theorem, we again use the notation in the proof of theorem 3.1. The cone (3.7) becomes

$$K = \left\{ \chi \,\middle|\, \chi = \alpha_1\chi_1 + \sum_{i=2}^{k} \sum_{j=1}^{m_i} \alpha_{ij}\, \chi_{ij}, \; \begin{array}{l} |\alpha_{ij}| \leqq \alpha_1 \\ \alpha_{ij} = \bar{\alpha}_{pq} \text{ if } \chi_{ij} = \bar{\chi}_{pq} \end{array} \right\}$$

Thus, if $\chi \in K$, then

$$A\chi = \beta_1\chi_1 + \sum_{i=1}^{k} \sum_{j=1}^{m_i} \beta_{ij}\, \chi_{ij}$$

where

$$\beta_1 = \lambda_1 \alpha_1$$

$$\beta_{ij} = \begin{cases} \alpha_{ij}\lambda_i + \mathcal{E}\, \alpha_{i,j+1} & j < m_i \\ \alpha_{ij}\lambda_i & j = m_i \end{cases} \qquad i = 2,\dots,k$$

and $\mathcal{E} < \lambda_1 - |\lambda_2|$. Hence

$$j < m_i : |\beta_{ij}| = |\alpha_{ij}\lambda_j + \mathcal{E}\, \alpha_{i,j+1}| < \alpha_1(\lambda_1 - \mathcal{E}) + \mathcal{E}\alpha_1 = \beta_1$$
$$j = m_i : |\beta_{ij}| = |\alpha_{ij}\lambda_i| < \alpha_1\lambda_1 = \beta_1$$

so clearly, $A\chi$ is in the interior of K.

It follows easily from the definition of irreducibility given in section 1, that if $0 \leqq A \leqq B$ and A is irreducible, then B must also be irreducible. The next theorem generalizes this result to K-irreducible matrices.

Theorem 4.5: If $0 \overset{K}{\leqq} A \overset{K}{\leqq} B$ and A is K-irreducible, then B is also K-irreducible.

Proof: Suppose B were K-reducible. Then B must have an eigenvector $\chi \in \delta K$. Using lemma 2.1, if $y \in F_\chi$ then $y \overset{K}{\leqq} \alpha\chi$, some α, and $Ay \overset{K}{\leqq} \alpha A\chi \overset{K}{\leqq} \alpha B\chi \in F_\chi$ hence $Ay \in F_\chi$. That is, A leaves F_χ invariant, and is therefore K-reducible.

This result allows us to further extend our generalization of the Perron-Frobenius theory.

Theorem 4.6: If $0 \overset{K}{\leqq} A \overset{K}{\leqq} B$ where A is K-irreducible, and $A \neq B$, then $\rho(A) < \rho(B)$.

Proof: By theorem 4.5, B is also K-irreducible, so there exists $y_1 \in K^\circ$ with

$$By_1 = \rho(B)y_1$$

Let $\chi_1 \in K^\circ$ satisfy

$$A\chi_1 = \rho(A)\chi_1.$$

Then, by hypothesis,

$$B\chi_1 \geqq A\chi_1 = \rho(A)\chi_1. \tag{4.3}$$

Let

$$t_o = \inf\{t > 0 \mid ty_1 - \chi_1 \in K\} \tag{4.4}$$

then, if $\rho(A) \neq 0$,

$$\begin{aligned} 0 \leqq B(t_o y_1 - \chi_1) &= t_o By_1 - B\chi_1 \\ &\leqq t_o \rho(B) y_1 - \rho(A)\chi_1 \quad \text{by } (4.3) \\ &= \rho(A)\left[t_o \frac{\rho(B)}{\rho(A)} y_1 - \chi_1\right] \end{aligned}$$

By (4.4) this implies

$$\rho(B) \geqq \rho(A). \tag{4.5}$$

Suppose $\rho(B) = \rho(A) = \lambda$, and let $t_o = \min\{t > 0 \mid ty_1 - \chi_1 \in K\}$.
If

(4.6) $\quad y_1 = \alpha\chi_1 \;,\; \alpha > 0,$

then

$$A(\alpha\chi_1) = \alpha\lambda\chi_1 = \lambda y_1 = By_1 = B(\alpha\chi_1)$$

hence $(B-A)\chi_1 = 0$. But $\chi_1 \in K^o$ so if $z \in K$ is arbitrary, then there exists $\beta > 0$ so that $0 \leqq^K z \leqq^K \beta\chi_1$. But $0 \leqq^K (B-A)z \leqq^K \beta(B-A)\chi_1 = 0$, and z was arbitrary in K, hence $A = B$. This contradiction implies that (4.6) cannot hold, i.e., χ_1 and y_1 are linearly independent, and hence if

$$z = t_o y_1 - \chi_1$$

then $z \neq 0$, z is on the boundary of K, and

$$Az = t_o Ay_1 - A\chi_1 \leqq^K t_o By_1 - A\chi_1 = \lambda t_o y_1 - \lambda\chi_1 = \lambda z.$$

Let F_z be the face given by lemma 2.1. Then F_z is invariant under A because for any $\chi \in F_z$, $\chi \leqq^K z$, some $\gamma > 0$, and $A\chi \leqq^K \gamma Az \leqq^K \gamma\lambda z \in F_z$, hence $A\chi \in F_z$. This contradicts the K-irreducibility of A, and hence we must conclude that $\rho(A) \neq \rho(B)$.

If A is not K-irreducible, then the above theorem is weakened slightly.

<u>Corollary</u>: If $0 \leqq^K A \leqq^K B$ then $\rho(A) \leqq \rho(B)$.

<u>Proof</u>: Let $C >>^K 0$, and define $A_t = A+tC$, $B_t = B+tC$, $t > 0$. Then clearly $A_t >> 0$, hence A_t is K-irreducible. By the previous theorem

$$\rho(A_t) < \rho(B_t)$$

and letting $t \to 0$ gives

$$\rho(A) \leqq \rho(B).$$

In the classical setting, it follows directly from the definitions that if a matrix A is positive, non-negative, or irreducible, then the same is true of A^t. Because of the spectral characterizations given in theorems 3.1, 4.3 and 4.4, the same type of statement can be made about $A \geqq^{K} 0$, $A >>^{K} 0$, or $A \geqq^{K} 0$ and K-irreducible. The cone which is left invariant by A, however, may not be the same cone which is invariant under A^t. For example, if

$$A = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix}$$

Then $A \geqq^{K} 0$ where $K = \{(\chi, y) \mid \chi \geqq 0,\ 2|y| \leqq \chi\}$.
But, K is not invariant with respect to A^t. The result that can be proven is:

<u>Theorem 4.7</u>: If $A \geqq^{K} 0$ then there exists a solid cone $\tilde{\tilde{K}}$ such that $A^t \geqq^{\tilde{K}} 0$. The same type of statement holds for $A \geqq^{K} 0$ and K-irreducible, or $A >>^{K} 0$.

<u>Proof</u>: Using the fact that A and A^t have the same eigenvalues, this theorem follows from 3.1, 4.3, 4.4.

We conclude this section by showing the connection between K-irreducibility and two related concepts.

Krasnoselski [5] has proved a result similar to theorem 4.3 using U_o-positivity in place of K-irreducibility, where a matrix $A \geqq^{K} 0$ is called <u>U_o-positive</u> if for some $U_o \in K^o$ and any $\chi \in K$ there are constants $\alpha(\chi) > 0$, $\beta(\chi) > 0$, and an integer $k(\chi) > 0$ such that

$$\alpha(\chi) U_o \leqq A^k \chi \leqq \beta(\chi) U_o. \tag{4.7}$$

Krasnoselski proves that a U_o-positive matrix has a unique eigenvector χ_1 in K. Since $A^k \chi_1 = \lambda_1^k \chi_1$, and $U_o \in K^o$, (4.7) implies $\chi_1 \in K^o$. Hence, by theorem 4.2, every U_o-positive matrix is K-irreducible. The converse is false, as shown by the example

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

which is non-negative and irreducible, but is not U_o-positive.

Schaefer [8] has proven a result, which is similar to theorem 4.3, for quasi-interior operators on a Banach space. For matrices which leave a solid cone invariant, this property can be defined as follows:

$A \overset{K}{\geqq} 0$ is <u>quasi-interior</u> if, for some $\lambda > \rho(A)$, $A(\lambda I-A)^{-1} \overset{K}{>>} 0$.

Before proving that this is equivalent to K-irreducibility, we need another basic fact about K-irreducible matrices.

<u>Lemma 4.2</u>: If $A \overset{K}{\geqq} 0$ is K-irreducible, then $(I+A)^{n-1} \overset{K}{>>} 0$. (n is the dimension of the space.)

<u>Proof</u>: Let y be an arbitrary non-zero element on the boundary of K, and let F_y be the face given by lemma 2.1. Then by K-irreducibility, $Ay \not\in F_y$ and hence $(I+A)y \not\in H_{F_y}$. In fact, if $y_1 = (I+A)y$ is not in K^o, then it must be in a face F_1, and the dimension of H_{F_1} must be greater than the dimension of H_F. Repeating this argument shows that $(I+A)^k y \in K^o$ for some $k \leqq n-1$, hence $(I+A)^{n-1} \overset{K}{>>} 0$.

<u>Theorem 4.8</u>: $A \overset{K}{\geqq} 0$ is K-irreducible if and only if A is quasi-interior.

<u>Proof</u>: If $\tilde{A} = A(\lambda I-A)^{-1} \overset{K}{>>} 0$ for some λ, then $\tilde{A}$ is K-irreducible. But, A and $\tilde{A}$ have the same eigenvectors so, by theorem 4.2, A is also K-irreducible. Conversely, if A is K-irreducible, then for $\lambda > \rho(A)$,

$$A(\lambda I-A)^{-1} = \sum_{k=1}^{\infty} \frac{A^k}{\lambda^k} \overset{K}{\geqq} \frac{A}{\lambda}\left(I + \frac{A}{\lambda} + \frac{A^2}{\lambda^2} + \ldots + \frac{A^{n-1}}{\lambda^{n-1}}\right)$$
$$\overset{K}{\geqq} \alpha A\,(I+A)^{n-1}$$

where α is some positive constant. By lemma 4.2, $(I+A)^{n-1} >> 0$ so the proof is complete provided $A\chi \overset{K}{>>} 0$ whenever $\chi \overset{K}{>>} 0$. But, if for some $\chi_o \overset{K}{>>} 0$, $A\chi_o$ were on the boundary of K, then $\chi_1 \leqq t\chi_o$, where $\chi_1 \in K^o$ is the eigenvector corresponding to $\rho(A)$, $t > 0$, and

$$\rho(A)\chi_1 = A\chi_1 \leqq tA\chi_o$$

which implies χ_1 is on the boundary of K. This contradiction proves the theorem.

5. Applications

In this section, we will show how the preceding results can be used to generalize some well known theorems which can be used to compare the rates of convergence of iterative processes. We give a detailed proof of the important Stein-Rosenberg theorem, in order to verify that our generalized Perron-Frobenius theory is as complete as the classical theory. The proof we use follows that given in [2]. Because of the comments at the end of section 4, the theorems of this section contain the results of Marek [7] for U_o-positive matrices.

Theorem 5.1: Let $B = B_1 + B_2$, be $n\chi n$ matrices, with $B_1 \neq 0$, $\rho(B_1) < 1$, and $B_i \overset{K}{\geqq} 0$ where K is a solid cone, and assume B is K-irreducible. Then, the matrix $H = (I-B_1)^{-1}B_2$ exists, and exactly one of the following holds:

$$\rho(H) < \rho(B) < 1$$

$$\rho(H) = \rho(B) = 1$$

$$\rho(H) > \rho(B) > 1$$

Proof: Since $\rho(B_1) < 1$, the series $\sum_{k=0}^{\infty} B_1^k$ converges. This shows that H exists, and $H \overset{K}{\geqq} 0$. Let $\gamma = \rho(H)$ and assume, for now, that $\gamma \neq 0$. By 3.1, there exists an $\chi_o \in K$, such that $H\chi_o = \gamma\chi_o$

Thus $$(I-B_1)^{-1}B_2\chi_o = \gamma\chi_o$$

$$B_2\chi_o = \gamma(I-B_1)\chi_o$$

5.1 $$(\gamma B_1 + B_2)\chi_o = \gamma\chi_o$$

5.2 $$(B_1 + \frac{1}{\gamma}B_2)\chi_2 = \chi_o$$

By theorem 4.5, γB_1+B_2 and $B_1+\gamma B_2$ are irreducible and if

$$\varphi_1(t) = \rho(tB_1 + B_2)$$

$$\varphi_2(t) = \rho(B_1 + \frac{1}{t}B_2)$$

then by theorem 4.3, there is a unique vector $y_o \in K^o$, with

$$(\gamma B_1 + B_2)y_o = \varphi_1(\gamma)y_o$$

Hence, by (5.1) and the uniqueness,

$$\varphi_1(\gamma) = \gamma$$

Similarly,

$$\varphi_2(\frac{1}{\gamma}) = 1$$

Now suppose $\gamma = 1$, then

$$\rho(B) = \rho(\gamma B_1 + B_2) = \varphi_1(\gamma) = \gamma = \rho(H)$$

If $\gamma > 1$, then

$$\gamma B_1 + B_2 \geqq^K B_1 + B_2 \geqq^K 0, \qquad \gamma B_1 + B_2 \neq B_1 + B_2$$

$$0 \leqq^K B_1 + \frac{1}{\gamma} B_2 \leqq^K B_1 + B_2 \quad , \qquad B_1 + \frac{1}{\gamma} B_2 \neq B_1 + B_2$$

so, by theorem 4.6,

$$\rho(\gamma B_1 + B_2) > \rho(B)$$

$$\rho(B_1 + \frac{1}{\gamma} B_2) < \rho(B)$$

and hence

$$\gamma = \varphi_1(\gamma) > \rho(B) \quad \text{and} \quad 1 = \varphi_2(\frac{1}{\gamma}) < \rho(B),$$

i.e.,

$$1 < \rho(B) < \gamma = \rho(H).$$

Finally, if $0 < \gamma < 1$, then

$$0 \leqq^K \gamma B_1 + B_2 \leqq^K B, \qquad \gamma B_1 + B_2 \neq B$$

$$B_1 + \frac{1}{\gamma} B_2 \geqq^K B \geqq^K 0, \quad B_1 + \frac{1}{\gamma} B_2 \neq B,$$

so, as before,

$$1 = \varphi_2(\frac{1}{\gamma}) = \rho(B_1 + \frac{1}{\gamma} B_2) > \rho(B) > \rho(\gamma B_1 + B_2) = \gamma = \rho(H).$$

Now suppose $\gamma = 0$. We must show that $\rho(B) < 1$. But, if $B_2 = 0$, then $\rho(B) = \rho(B_1) < 1$ by assumption. If $B_2 \neq 0$, then

$$\varphi_2(0) = \rho(B_1) < 1.$$

If $\rho(B) \geqq 1$ then $\exists\ t_o \in [0,1]$ with $\varphi_2(t_o) = 1$, but then

$$1 = \varphi_2(t_o) = \rho(B_1 + t_o B_2)$$

so $\exists\ \chi > >^K 0$, with $(B_1 + t_o B_2)\chi = \chi$

i.e., $$(I-B_1)^{-1} B_2 \chi = \frac{1}{t_o}\chi$$

so $$\rho(H) \geqq \frac{1}{t_o} > 0$$

which contradicts $\gamma = 0$. Hence, we must conclude $\rho(B) < 1$.

If B is not K-irreducible, we can no longer prove the strict inequalities. In fact, the theorem becomes;

<u>Theorem 5.2</u>: If $B = B_1 + B_2$, where $B_i \geqq^K 0$, $\rho(B_1) < 1$, $B_1 \neq 0$. Then $H = (I-B_1)^{-1}B_2$ exists, and either $\rho(H) \leqq \rho(B) \leqq 1$ or $\rho(H) \geqq \rho(B) \geqq 1$.

As a corollary to theorem 5.1, we have the following generalization of a theorem due to Fiedler and Ptak [3].

<u>Theorem 5.3</u>: Let $B = B_1 + B_2$, where $B_i \geqq^K 0$, B is K-irreducible, $\rho(B) < 1$, and $B_1 \neq 0$. Suppose P is another matrix which satisfies $0 \leqq^K P \leqq^K B_2$, $P \neq 0$, $P \neq B_2$. Then $H_P = \left(I-(B_1 + P)\right)^{-1} (B_2 - P)$ exists, and

$$0 < \rho(H_P) < \rho(H_o)$$

where $$H_o = (I-B_1)^{-1}B_2 .$$

The proof follows exactly as in the case where K is the positive hyper-octant.

As an indication that these results are indeed more useful than the classical theorems, consider a matrix B with elements $\{b_{ij}\}$ which satisfy

$$(-1)^{i+j}\, b_{ij} \geqq 0.$$

If B_1 and B_2 are upper and lower triangular, respectively, then we have $B_i \geqq^K 0$ where

$$K = \{(x_1, \ldots, x_n) \mid (-1)^i\, x_i \leqq 0\}.$$

Hence, the above theorems may be applicable although B is not non-negative as required by the standard theorems.

Finally, we point out that, in a similar manner, Varga's theory of regular splittings [2] can also be extended to matrices with invariant cones.

References

1. Birkhoff, G., Linear Transformations With Invariant Cones, American Math. Monthly, 74 No. 3, 1967, 274-276.

2. Varga, R. S., Matrix Iterative Analysis, Prentice-Hall, New Jersey, 1963.

3. Fiedler, M., Ptak, V., Über die Konvergenz des Verallgemeinerten Seidelschen Verfahrens zur Lösung von Systemen Linearer Gleichungen, Math. Nach., 15, 1956, 31-38.

4. Gantmacher, F. R., The Theory of Matrices, Vol. 2, Chelsea, New York, 1959.

5. Krasnoselskii, M. A., Positive Solutions of Operator Equations, Trans. R. E. Flaherty, Noordhatt, Groningen, 1964.

6. Krein, M. G., Rutman, M. A., Linear Operators Leaving Invariant a Cone in a Banach Space, American Math. Society, Translation No. 26, 1950.

7. Marek, I., U_0-Positive Operators and Some of their Applications, SIAM, J. Appl. Math., 15, No. 3, 1967, 484-494.

8. Schaefer, H., Some Spectral Properties of Positive Linear Operators, Pacific J. Math., 10, 1960, 1009-1019.

9. Vandergraft, J. S., Newton's Method for Convex Operators in Partially Ordered Spaces, SIAM J. Numerical Analysis, 4, No. 3, 1967, 406-432.