MATHEMATICS OF OPERATIONS RESEARCH
Vol. 7, No. 2, May 1982
Printed in U.S.A.

THE CONTRACTION MAPPING APPROACH TO THE PERRON-FROBENIUS THEORY: WHY HILBERT'S METRIC?*†

ELON KOHLBERG AND JOHN W. PRATT

Harvard University

The Perron-Frobenius Theorem says that if A is a nonnegative square matrix some power of which is positive, then there exists an x_0 such that $A^n x / \|A^n x\|$ converges to x_0 for all $x \geqslant 0$. There are many classical proofs of this theorem, all depending on a connection between positivity of a matrix and properties of its eigenvalues. A more modern proof, due to Garrett Birkhoff, is based on the observation that every linear transformation with a positive matrix may be viewed as a contraction mapping on the nonnegative orthant. This observation turns the Perron-Frobenius theorem into a special case of the Banach contraction mapping theorem. Furthermore, it applies equally to linear transformations which are positive in a much more general sense.

The metric which Birkhoff used to show that positive linear transformations correspond to contraction mappings is known as Hilbert's projective metric. The definition of this metric is rather complicated. It is therefore natural to try to define another, less complicated metric, which would also turn positive matrices into contractions. The main result of this paper is that, essentially, this is impossible.

The paper also gives some other results of possible interest in themselves, as well as enough background to make the presentation self-contained.

1. Introduction. The famous Perron Theorem may be viewed as saying that if A is a positive linear tranformation on R^m, then there exists an $x_0 > 0$ such that[1]

$$\text{for all } x \geqslant 0,\ A^n x \text{ converges in direction to } x_0, \tag{1.1}$$

i.e.,

$$\frac{A^n x}{\|A^n x\|} \to \frac{x_0}{\|x_0\|}.$$

The Banach contraction mapping theorem says that if A is a contraction on a complete metric space (X, D), that is,

$$A \text{ maps } X \text{ into } X \tag{1.2}$$

and

$$\text{for some } k < 1,\ D(Ax, Ay) \leqslant kD(x, y) \text{ for all } x, y \in X, \tag{1.3}$$

then there exists an $x_0 \in X$ such that $A^n x \to x_0$ for all $x \in X$.

It is curious that the condition for (1.1) is positivity but the conclusion is essentially that of the contraction mapping theorem. Birkhoff (1957) has shown that this is no

* Received January 30, 1979; revised May 28, 1980.
AMS 1980 subject classification. Primary 15A48.
IAOR 1973 subject classification. Main: Matrices.
OR/MS Index 1978 subject classification. Primary 435 Mathematics, matrices.
Key words. Perron-Frobenius Theorem, nonnegative matrices, convergence in direction, Hilbert's metric, contractions.
† This research was supported by the Associates of the Harvard Business School.

[1] As usual, $x > 0$ ($\geqq 0$) means all coordinates of x are positive (nonnegative) and $x \geqslant 0$ means $x \geqq 0$ but $x \neq 0$. A linear transformation A is positive (nonnegative), denoted $A > 0$ ($\geqq 0$), if $Ax > 0$ ($\geqq 0$) for all $x \geqslant 0$, that is, if its matrix has all positive (nonnegative) elements.

0364-765X/82/0702/0198$01.25

coincidence, and in fact that the Perron Theorem can be considered a special case of the contraction mapping theorem. Specifically, if a positive linear transformation A is viewed as a transformation on the set X of rays in the nonnegative orthant of R^m, it satisfies (1.2). Birkhoff's key observation was that there is a metric D on X such that all positive linear transformations A also satisfy (1.3). Applying the contraction mapping theorem then gives the Perron Theorem, since convergence of rays is convergence in direction. This natural proof circumvents clever but ad hoc arguments usually used to prove (1.1) by relating positivity to conditions on eigenvalues (e.g., Karlin and Taylor (1975)).[2] More importantly, Birkhoff's approach simplifies and unifies several kinds of generalization of Perron's Theorem. First, it provides an immediate proof of Frobenius's extension from positive matrices to nonnegative matrices some power of which is positive, because the contraction mapping theorem extends trivially to mappings some power of which is a contraction. Second, for a sequence of different positive linear transformations A_n with contraction constants k_n in (1.3), it allows us to conclude that $\prod_1^N A_n x$ and $\prod_1^N A_n y$ become close in direction for all $x, y \geqslant 0$ if $\prod_1^\infty k_n = 0$.[3] Third, Birkhoff's approach is geometric and applies to linear transformations in an arbitrary linear space which map a quite general convex cone K into itself. Thus it allows a reinterpretation of the Perron-Frobenius theory in which $x \geqq 0$ means $x \in K$ and A nonnegative means $AK \subset K$.

The metric used by Birkhoff was invented by Hilbert for different purposes in non-Euclidean geometry. In the case of R_+^m Hilbert's metric is

$$d(x, y) = \log \frac{\max_i (x_i / y_i)}{\min_i (x_i / y_i)} \tag{1.4}$$

where x_i denotes the ith coordinate of x. We can view $d(x, y)$ as a metric on rays because $d(\lambda x, \mu y) = d(x, y)$ for $\lambda, \mu > 0$.

The formula for Hilbert's metric is surprisingly complicated. One wonders whether there might be a "simpler" metric which could be used to carry out Birkhoff's program. The answer turns out to be "No." Our main result shows that Hilbert's is essentially the only metric that will serve the purpose.

The paper is organized as follows. §2 indicates how Hilbert's metric arises from the problem, in R_+^m. §3 gives the general definition and some basic properties of Hilbert's metric. §4 presents our main result for positive linear transformations in finite-dimensional spaces. Infinite-dimensional spaces and more general nonnegative transformations are considered in §5. §6 explains in more detail how theorems of the Perron-Frobenius type can be derived using Hilbert's metric and the contraction mapping theorem.

2. Discovering Hilbert's metric in R_+^m. This section is motivational and may be skipped. Its purpose is to indicate how one might discover Hilbert's metric when looking for a metric on rays which contracts under positive linear transformations. To simplify the exposition, we consider instead the related problem of finding a metric which contracts weakly under nonnegative linear transformations. Also, we restrict attention to R_+^m. Accordingly, we first seek a function D on $R_+^m \times R_+^m$ such that

$$D(Ax, Ay) \leqslant D(x, y) \quad \text{for all } x, y \geqslant 0 \text{ and all } A \geqslant 0. \tag{2.1}$$

[2] Note that fixed-point arguments (e.g., Debreu and Herstein, 1953) yield only the existence of a positive eigenvector x_0, not convergence in direction to x_0 as at (1.1). For the latter, one still needs to show how positivity implies that the eigenvalue associated with x_0 dominates all other eigenvalues.

[3] Theorems of this type are called "weak ergodic" theorems. The usual proofs by eigenvalue techniques are often very difficult (e.g., Furstenberg and Kesten (1960)).

Since we want $D(x, y)$ to depend only on the rays through x and y, D must also satisfy

$$D(\lambda x, \mu y) = D(x, y) \quad \text{for all } \lambda,\ \mu > 0 \text{ and all } x, y \geqslant 0. \tag{2.2}$$

In view of these conditions, we start by asking whether, given two pairs of points x, y and x', y' in R^m_+, there is a nonnegative linear transformation A such that $Ax = x'$ and $Ay = \alpha y'$ for some $\alpha > 0$. Of course, for any particular α, the requirements $Ax = x'$ and $Ay = \alpha y'$ determine A on $\mathrm{Sp}\{x, y\}$, the two-dimensional subspace spanned by x and y. The restriction to $\mathrm{Sp}\{x, y\}$ of the nonnegativity requirement $Ax \geqq 0$ is

$$A(x - \lambda y) = x' - \lambda\alpha y' \geqq 0 \quad \text{whenever} \quad x - \lambda y \geqq 0; \tag{2.3}$$

$$A(\lambda y - x) = \lambda\alpha y' - x' \geqq 0 \quad \text{whenever} \quad \lambda y - x \geqq 0. \tag{2.4}$$

This is the same as requiring that $m\alpha \leqslant m'$ and $M\alpha \geqslant M'$, where

$$m = m(x, y) = \sup\{\lambda \geqslant 0 : x - \lambda y \geqq 0\} = \min_i\{x_i/y_i\}, \tag{2.5}$$

$$M = M(x, y) = \inf\{\lambda \geqslant 0 : \lambda y - x \geqq 0\} = \max_i\{x_i/y_i\}, \tag{2.6}$$

$m' = m(x', y')$, and $M' = M(x', y')$. In other words, α must satisfy $M'/M \leqslant \alpha \leqslant m'/m$. The choice of such an α is possible[4] if and only if $M'/m' \leqslant M/m$. We will shortly prove (Lemma 2.12) that a nonnegative linear transformation on $\mathrm{Sp}\{x, y\}$ can always be extended to a nonnegative linear transformation on all of R^m. It follows that

THEOREM 2.7. *Given $x, y, x', y' \geqslant 0$, there exists a nonnegative linear transformation A such that $Ax = x'$ and $Ay = \alpha y'$ for some $\alpha > 0$ if and only if $c(x', y') \leqslant c(x, y)$, where*

$$c(x, y) = \frac{M(x, y)}{m(x, y)} = \frac{\max_i\{x_i/y_i\}}{\min_i\{x_i/y_i\}}. \tag{2.8}$$

THEOREM 2.9. *D satisfies (2.1) and (2.2) if and only if D is a monotone nondecreasing function of c.*

PROOF. If D satisfies (2.1) and (2.2), then by Theorem 2.7,

$$c(x', y')\{\lesseqgtr\}c(x, y) \quad \text{implies} \quad D(x', y')\{\lesseqgtr\}D(x, y). \tag{2.10}$$

It follows that D is a monotonic function of c. Conversely, any monotonic function of c obviously satisfies (2.1) and (2.2) if c itself does. But for c, by straightforward application of the definitions (2.5), (2.6), and (2.8), we have, first, if $A \geqslant 0$, then $M(Ax, Ay) \leqslant M(x, y)$ and $m(Ax, Ay) \geqslant m(x, y)$ whence $c(Ax, Ay) \leqslant c(x, y)$. Second,

$$m(\lambda x, \mu y) = \frac{\lambda}{\mu} m(x, y) \quad \text{and} \quad M(\lambda x, \mu y) = \frac{\lambda}{\mu} M(x, y),$$

whence $c(\lambda x, \mu y) = c(x, y)$. ∎

The question remaining is what functions of c are metrics on rays in R^m_+. It is easy to verify (see also §3) that c is symmetric, that $c(x, y) \leqslant c(x, z)c(z, y)$ for all $x, y, z \geqslant 0$, and that $c(x, y) \geqslant 1$ with equality if and only if x and y lie on the same ray. Thus, if we set

$$d(x, y) \equiv \log c(x, y) = \log \frac{\max_i\{x_i/y_i\}}{\min_i\{x_i/y_i\}}, \tag{2.11}$$

[4]The cases $m = 0$ and $M = \infty$ can be handled by special conventions (Kohlberg and Pratt (1979)) but will be ignored since our purpose here is motivational.

we obtain a function that satisfies all the requirements of a metric except that $d(x, y) = 0$ if and only if $x = \lambda y$ for some $\lambda > 0$ (not, as usual, if and only if $x = y$). The function d is called Hilbert's projective metric on R^m_+.

In developing Theorem 2.7 we used

LEMMA 2.12. *Let E be a two-dimensional subspace of R^m and let $A : E \to R^m$ be a linear transformation such that $Ax \geqslant 0$ whenever $x \in E$ and $x \geqslant 0$. Then there exists a nonnegative linear transformation $\bar{A}$, defined on all of R^m, such that $\bar{A} = A$ on E.*

PROOF. Consider the intersection of E with R^m_+. This is a two-dimensional cone. Let a and b lie on the two sides of this cone. Each must satisfy a boundary constraint not satisfied by the other; that is, for some indices i and j, $a_i > 0$, $b_i = 0$, and $a_j = 0$, $b_j > 0$. If we let

$$\bar{A}x \equiv \frac{x_i}{a_i} Aa + \frac{x_j}{b_j} Ab$$

then $\bar{A}a = Aa$ and $\bar{A}b = Ab$; hence $\bar{A} = A$ on E. Also note that, if $x \geqslant 0$, then $\bar{A}x$ is a nonnegative linear combination of Aa and Ab, and therefore $\bar{A}x \geqslant 0$. Thus, $\bar{A}$ is nonnegative. ∎

3. Definition and properties of Hilbert's metric. In this section we describe Hilbert's metric and its properties. Much of the development follows Bushell (1973) and Birkhoff (1967) but we have removed some unnecessary conditions. In addition, we give some new results (e.g., Theorem 3.17).

Let K be a convex cone in a real vector space X. That is, if $x, y \in K$ and λ, μ are nonnegative real numbers, then $\lambda x + \mu y \in K$. Define the ordering induced by K in the usual way: $y \geqq x$ if and only if $y - x \in K$ and $y \geqslant x$ if and only if $y - x \in K$ but $y - x \neq 0$. Clearly, $K = \{x \in X : x \geqq 0\}$.

We will restrict attention to cones satisfying the following conditions:

$$K \cap -K = \{0\}, \tag{3.1}$$

and

$$\text{the intersection of } K \text{ with every straight line is closed.} \tag{3.2}$$

Cones satisfying (3.1) are called *pointed.* In finite-dimensional spaces, cones satisfying (3.2) are closed.

Conditions (3.1) and (3.2) may easily be rephrased as conditions on the ordering induced by K.

$$x \geqq 0 \quad \text{and} \quad x \leqq 0 \quad \text{imply} \quad x = 0. \tag{3.1$'$}$$

$$x \geqslant 0 \quad \text{and} \quad \lambda x + y \geqslant 0 \quad \text{for all } \lambda > 0 \quad \text{imply } y \geqslant 0. \tag{3.2$'$}$$

Condition (3.2′) is usually referred to as the Archimedean axiom.

A function $D : K \times K \to R \cup \{\infty\}$ is a *projective metric* on K if, for all $x, y, z \geqq 0$,

$$D(x, y) = D(y, x), \tag{3.3}$$

$$D(x, y) \leqslant D(x, z) + D(z, y), \tag{3.4}$$

$$D(x, y) \geqslant 0, \tag{3.5}$$

$$D(x, y) = 0 \quad \text{if and only if } x = \lambda y \text{ for some } \lambda > 0. \tag{3.6}$$

It follows easily that $D(x, y)$ is constant on rays, that is,

$$D(\lambda x, \mu y) = D(x, y) \quad \text{for } \lambda, \mu > 0. \tag{3.7}$$

We now define *Hilbert's (projective) metric*, d, as follows:[5] $d(0,0)=0$; when $x, y \geqslant 0$, $d(x,0)=d(0,y)=\infty$ and

$$d(x,y) \equiv \log \frac{M(x,y)}{m(x,y)} \tag{3.8}$$

where

$$M(x,y) \equiv \inf\{\lambda \geqslant 0 : x \leqq \lambda y\} \quad \text{and} \quad m(x,y) = \sup\{\lambda \geqslant 0 : x \geqq \lambda y\}. \tag{3.9}$$

Clearly,

$$m(x,y) = \frac{1}{M(y,x)}. \tag{3.10}$$

It is straightforward to verify that d is a projective metric. By (3.10), $d(x,y) = \log[M(x,y)M(y,x)]$ and by (3.9), $M(x,y) \leqslant M(x,z)M(z,x)$. These prove conditions (3.3) and (3.4). Furthermore, by (3.2), the intersection of the line $\{-x+\lambda y : 0 \leqslant \lambda < \infty\}$ with K is closed and therefore, if $M = M(x,y) < \infty$ then $My - x \geqq 0$. Also by (3.2), $x - my \geqq 0$. It follows that

$$\text{if } M < \infty \text{ then } my \leqq x \leqq My. \tag{3.11}$$

Hence, by (3.1′), $m \leqslant M$ with equality if and only if $x = my$. This establishes (3.5) and (3.6).

Note that $m = 0$ if and only if $x - \lambda y \notin K$ for all $\lambda > 0$. Since K is a cone, this is the same as saying that the line from y to x leaves K at x. Thus we have

$$d(x,y) < \infty \qquad \text{if and only if } x \text{ and } y \text{ are interior to the intersection of the line through them with } K. \tag{3.12}$$

The main property of Hilbert's metric in studying convergence in direction is that it contracts under a wide class of linear transformations, as explained below. A mapping $A : X \to X$ is called *nonnegative* if $Ax \geqq 0$ for all $x \geqq 0$. If D is a projective metric, we define the *contraction ratio* of A with respect to D as

$$\begin{aligned} k_D(A) &= \inf\{k : D(Ax,Ay) \leqslant kD(x,y) \quad \forall x,y \geqslant 0\} \\ &= \sup\left\{\frac{D(Ax,Ay)}{D(x,y)} : D(x,y) > 0\right\}, \end{aligned} \tag{3.13}$$

with the conventions $(\infty/\infty) = 1$ and $\infty \leqslant k\infty$ if and only if $k \geqslant 1$. If $k_D(A) < 1$ we say that A is a *contraction* with respect to D. For Hilbert's metric d, we have

PROPOSITION 3.14. (BIRKHOFF, 1957). *A nonnegative linear transformation A is a contraction with respect to Hilbert's metric d if and only if the diameter*

$$\Delta \equiv \sup\{d(Ax,Ay) : x, y \geqslant 0\}$$

is finite. More specifically,

$$k_d(A) = \frac{\sqrt{\Gamma}-1}{\sqrt{\Gamma}+1},$$

[5] The definition of d depends on the cone K under discussion, but we suppress this in the notation.

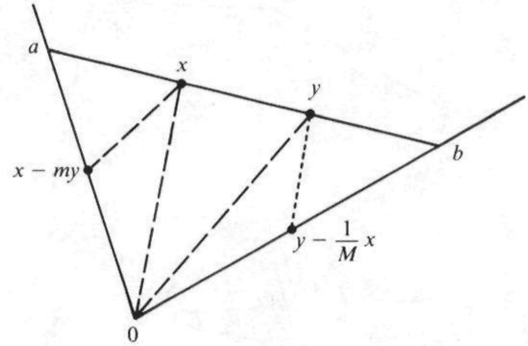

FIGURE 1

where

$$\Gamma = e^{\Delta} = \sup\left\{ \frac{M(Ax,Ay)}{m(Ax,Ay)} : x, y \geqslant 0 \right\}.$$

Geometrically, m, M, and d may be depicted as follows. Replacing x by λx for a suitable $\lambda > 0$, if necessary, will insure that the line through x and y leaves K at two points, a and b, in the two-dimensional subspace spanned by x and y, as shown in Figure 1. By (3.9), the point $x - my$ is obtained by moving from x in the $-y$ direction until the nonnegativity constraint is violated. By similar triangles, we see that $m = \overline{ax}/\overline{ay}$ and that $M = \overline{xb}/\overline{yb}$. Thus

$$d(x, y) = \log \frac{M(x, y)}{m(x, y)} = \log \frac{\overline{ay}\,\overline{xb}}{\overline{ax}\,\overline{yb}}, \tag{3.15}$$

i.e., d is the logarithm of what is known in projective geometry as the cross ratio[6] of (a, x, y, b).

An important property of Hilbert's metric which we shall have occasion to use is that it is additive on straight lines, that is,

LEMMA 3.16. *If z is a positive combination of x and y, then $d(x, y) = d(x,z) + d(z,x)$.*

PROOF. By (3.7), it suffices to consider convex combinations, for which, by (3.15),

$$d(x,z) + d(z, y) = \log\left(\frac{\overline{az}\,\overline{xb}}{\overline{ax}\,\overline{zb}}\;\frac{\overline{ay}\,\overline{zb}}{\overline{az}\,\overline{yb}} \right) = \log\left(\frac{\overline{ay}\,\overline{xb}}{\overline{ax}\,\overline{yb}} \right) = d(x, y). \quad ∎$$

In passing, we mention that the converse of Lemma 3.16 holds for K strictly convex, but not for other K. Necessary and sufficient conditions for additivity are given by the following

THEOREM 3.17. *Let $x, y, z \geqslant 0$. Then $d(x, y) = d(x,z) + d(z, y)$ if and only if both $a(x, y), a(x,z), a(z, y)$ are coplanar and $b(x, y), b(x,z), b(z, y)$ are coplanar, where for any $x, y \geqslant 0$, $a(x, y)$, $b(x, y)$ are the boundary rays of the cone* $\mathrm{Sp}\{x, y\} \cap K$ *(labelled so that x is between a and y).*

[6]The cross ratio of any four points a, x, y, b lying in that order on a straight line in any linear space is defined as $R(a, x, y, b) = \overline{ay}\,\overline{xb}/\overline{ax}\,\overline{yb}$, where $\overline{ay}$, $\overline{xb}$, etc., are distances along this line. More precisely, $R(a, x, y, b) = t_y(1 - t_x)/t_x(1 - t_y)$, where $x = a + t_x(b - a)$ and $y = a + t_y(b - a)$. The fundamental property of the cross ratio is that it is invariant under projections (distances, and even ratios of distances, are not), that is, $R(a', x', y', b') = R(a, x, y, b)$ whenever a', x', y', b' are the intersections of a straight line with the rays through a, x, y, and b, respectively. One proof of this invariance is that Hilbert's metric satisfies (3.7).

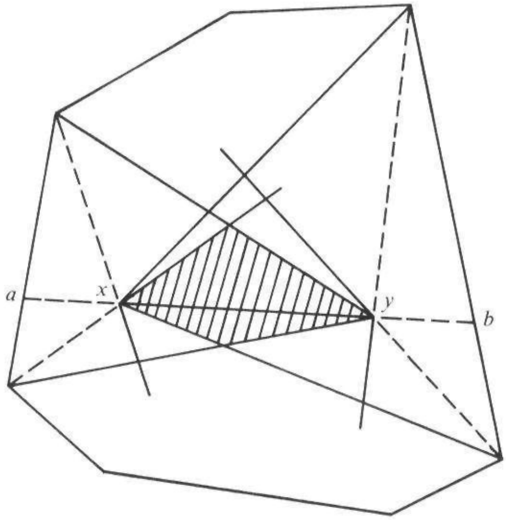

FIGURE 2

To illustrate the theorem, consider the intersection of a cone in R^3 with an appropriate plane, as shown in Figure 2. Then $d(x, y) = d(x,z) + d(z,x)$ if and only if z lies in the shaded region.

PROOF. The definitions (3.9) of $M(x, y)$ and $m(x, y)$ imply that $a(x, y)$ and $b(x, y)$ are the rays through $x - m(x, y)y$ and $M(x, y)y - x$, respectively. Thus if we choose the vectors x, y, z as the basis of a coordinate system, the condition of the theorem is simply that both the vectors $(1, -m(x, y), 0)$, $(1, 0, -m(x, z))$, $(0, -m(z, y), 1)$ and the vectors $(-1, M(x, y), 0)$, $(-1, 0, M(x, z))$, $(0, M(z, y), -1)$ are linearly dependent. It is easily seen that this is equivalent to $m(x, y) = m(x, z)m(z, y)$ and $M(x, y) = M(x, z)M(z, y)$, and that this in turn is equivalent to $d(x, y) = d(x, z) + d(z, y)$. ∎

REMARK 3.18. Note that the conditions on K assumed in defining Hilbert's metric, (3.1) and (3.2), do not require either a topology on X or that K have a nonempty radial kernel.[7] In fact, a necessary and sufficient condition on K for the d defined by (3.8) to be a projective metric is that K contain no lines, i.e., for all $x, y \geqslant 0$, the boundary rays of $\mathrm{Sp}\{x, y\} \cap K$ make an angle less than 180°. (If the angle is 180°, $m(x, y) = M(x, y)$ and $d(x, y) = 0$ even though x and y lie on different rays.)

4. The main theorem.

We now give our main result in finite-dimensional space, indicating later how it can be generalized. In stating this result, we use the following natural definition of positivity in finite-dimensional space: $x > 0$ if x is in the relative interior of the nonnegative cone K and a mapping A is *positive* if $Ax > 0$ for all $x \geqslant 0$.

THEOREM 4.1. *Let $\geqslant$ be the ordering induced by a closed pointed cone K in R^m and let d be Hilbert's metric on K. Then every positive linear transformation is a contraction with respect to d. Conversely, if D is a projective metric on K such that every positive linear transformation is a contraction with respect to D, then there exists a continuous strictly increasing function $f: R_+ \to R_+$ such that $D(x, y) = f(d(x, y))$ for all $x, y > 0$. Furthermore, the contraction ratios satisfy $k_d(A) \leqslant k_D(A)$ for all positive A.*

[7] A point $z \in K$ is in the radial kernel if, for every $x \in X$, $\exists \epsilon > 0$ such that $z + \epsilon x \in K$. The assumption that such a point z exists would rule out many interesting examples, such as the cone of nonnegative functions in an L^p space $(1 \leqslant p < \infty)$ and the cone of nonnegative functions in the space of Lebesgue measurable functions.

PROOF. If A is positive, then by (3.12), $d(Ax, Ay) < \infty$ for all $x, y \geqslant 0$, and hence by a simple compactness argument, $\sup\{d(Ax, Ay) : x, y \geqslant 0\} < \infty$. It follows, by Proposition 3.14, that A is a contraction with respect to d.

We give the proof of the rest of the theorem in the form of a series of lemmas. Whenever D is mentioned, it is assumed to be a projective metric that contracts under positive linear transformations.

LEMMA 4.2. *If $x', y' > 0$ and $d(x', y') < d(x, y)$, then there exists a positive linear transformation A such that $Ax = x'$ and $Ay = \alpha y'$ for some $\alpha > 0$.*

PROOF. Let $M = M(x, y)$, $m = m(x, y)$, $M' = M(x', y')$, and $m' = m(x', y')$ be as defined in (3.9). Since $d(x', y') < d(x, y)$ we may assume, replacing $\alpha y'$ by y' for a suitable $\alpha > 0$, that

$$M' < M \quad \text{and} \quad m' > m. \tag{4.3}$$

Define a linear transformation A on $\mathrm{Sp}\{x, y\}$ by $Ax = x'$, $Ay = y'$. The points $u = M'y - x$ and $v = x - m'y$, which [by (4.3)] lie outside K in the x and y directions respectively, are mapped into $Au = M'y' - x'$ and $Av = x' - m'y'$ which lie on the two boundary rays of $\mathrm{Sp}\{x', y'\} \cap K$.

To keep track of the notation the reader may find Figure 3 helpful. In this figure, x and y designate the intersections of the rays through x and y with the line through u and v and similarly for x', y', Au, and Av.

We can extend the definition of A so as to obtain a positive linear transformation on all of R^m as follows. A standard separation argument shows that there exist linear functionals g and h such that

$$g(u) < 0 < g(z) \quad \text{and} \quad h(v) < 0 < h(z) \quad \text{for all } z \geqslant 0. \tag{4.4}$$

Define u_0 and v_0 as the points on the line segment (u, v) such that $g(u_0) = h(v_0) = 0$. Clearly u_0 is between u and K and hence on the opposite side of K from v. Therefore $h(u_0) > 0$. Also $Au_0 > 0$ since Au_0 is a positive combination of $Au \geqslant 0$ and $x' > 0$. Similarly $g(v_0) > 0$ and $Av_0 > 0$.

Define a linear transformation $\bar{A}$ on all of R^m by

$$\bar{A}z = \frac{g(z)}{g(v_0)} Av_0 + \frac{h(z)}{h(u_0)} Au_0. \tag{4.5}$$

Then $\bar{A}$ is an extension of A since it coincides with A at u_0 and v_0 and $\mathrm{Sp}\{u_0, v_0\} = \mathrm{Sp}\{x, y\}$. Furthermore, $\bar{A}$ is positive since, by (4.4), for every $z \geqslant 0$, $\bar{A}z$ is a positive combination of the positive points Au_0 and Av_0. ∎

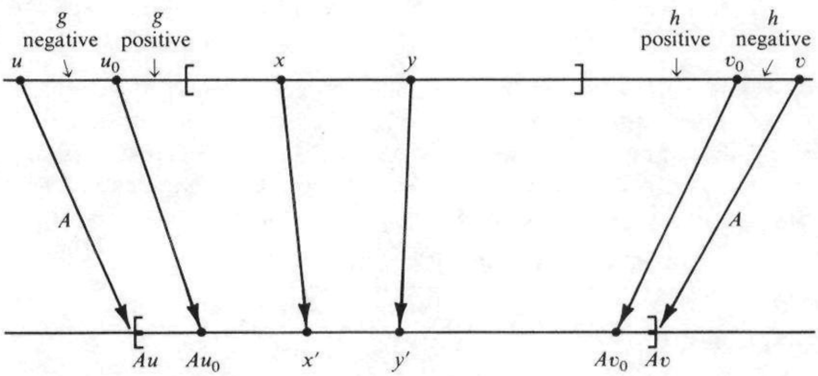

FIGURE 3

Arrows indicate mapping A. Ratios of distances are preserved. [,] indicate points where lines leave the cone K.

COROLLARY 4.6. *If $x', y' > 0$ and $d(x', y') < d(x, y)$, then $D(x', y') < D(x, y)$.*

PROOF. Define A as in Lemma 4.2. Since A is positive, it is a contraction with respect to D. It follows that $D(x', y') = D(Ax, Ay) < D(x, y)$. ∎

LEMMA 4.7. *If $x, y > 0$, then $D(x, x + y/n) \to 0$ as $n \to \infty$.*

PROOF. Let $x_n = x + y/n$. It follows easily from (3.9) that $\lim d(x, x_n) = 0$. If $l = \limsup D(x, x_n) > 0$, let A be a positive linear transformation that maps x and y to different rays. Then, for every N, $d(Ax, Ax_N) > 0 = \lim d(x, x_n)$. Therefore, by Corollary 4.6, $D(Ax, Ax_N) \geqslant \limsup D(x, x_n) = l$, whence

$$\limsup \frac{D(Ax, Ax_n)}{D(x, x_n)} \geqslant \frac{l}{l} = 1.$$

This contradicts the assumption that every positive linear transformation is a contraction with respect to D. ∎

LEMMA 4.8. *If $x, y, x', y' > 0$ and $d(x, y) = d(x', y')$, then $D(x, y) = D(x', y')$.*

PROOF. Since $d(x + y/n, y) < d(x, y) = d(x', y')$, we have by Corollary 4.6 $D(x + y/n, y) < D(x', y')$. It follows that $D(x, y) < D(x, x + y/n) + D(x', y')$. Therefore $D(x, y) \leqslant D(x', y')$ by Lemma 4.7. ∎

LEMMA 4.9. *$D(x, y) = f(d(x, y))$ for all $x, y > 0$, where $f: R_+ \to R_+$, $f(0) = 0$, and f is strictly increasing, continuous, and subadditive, that is,*

$$f(s + t) \leqslant f(s) + f(t) \quad \textit{for all } s, t > 0. \tag{4.10}$$

PROOF. The existence and strict monotonicity of f follow from Corollary 4.6 and Lemma 4.8, while $f(0) = 0$ because both D and d vanish when $x = y$. To prove the subadditivity, let $x, y, z > 0$ lie on a straight line and satisfy $d(x, z) = s$ and $d(z, y) = t$. By Lemma 3.16, $d(x, y) = s + t$. It follows that $f(s + t) = D(x, y) \leqslant D(x, z) + D(z, y) = f(s) + f(t)$. Continuity at 0 follows from Lemma 4.7. Continuity elsewhere then follows from subadditivity and monotonicity by

$$f(s) - f(\epsilon) \leqslant f(s - \epsilon) \leqslant f(s) \leqslant f(s + \epsilon) \leqslant f(s) + f(\epsilon) \quad \text{for } \epsilon > 0. \quad ∎$$

The following lemma says that for Hilbert's metric, the small distances $d(x, y)$ are those which contract least.

LEMMA 4.11. *For all $\epsilon > 0$ and all positive A,*

$$k_d(A) = \sup\left\{ \frac{d(Ax, Ay)}{d(x, y)} \ : \ x, y > 0 \quad \textit{and} \quad 0 < d(x, y) < \epsilon \right\}.$$

PROOF. The case $d(x, y) = \infty$ can be ignored because $d(Ax, Ay) < \infty$ for A positive. A simple continuity argument shows that adding the condition $x, y > 0$ has no effect on the supremum in the definition (3.13) of $k_d(A)$. To show that adding the condition $d(x, y) < \epsilon$ also has no effect, let z lie on the line segment between x and y. Since Hilbert's metric is additive on straight lines, $d(x, y) = d(x, z) + d(z, y)$. Thus

$$\frac{d(Ax, Ay)}{d(x, y)} \leqslant \frac{d(Ax, Az) + d(Az, Ay)}{d(x, z) + d(z, y)}$$

$$\leqslant \max\left\{ \frac{d(Ax, Az)}{d(x, z)}, \frac{d(Az, Ay)}{d(z, y)} \right\}.$$

Since z can be chosen so that $d(x, z) = d(z, y) = \frac{1}{2} d(x, y) < \infty$, the lemma follows. ∎

Since $D = f(d)$ as in Lemma 4.9, Lemma 4.11 implies that, for positive A,

$$k_D(A) \geqslant \sup\left\{ \frac{f(d(Ax, Ay))}{f(d(x, y))} : x, y > 0 \quad \text{and} \quad d(x, y) > 0 \right\}$$

$$\geqslant \limsup_{t \to 0+} \frac{f(k_d(A)t)}{f(t)}. \tag{4.12}$$

$k_D(A) \geqslant k_d(A)$ now follows from

LEMMA 4.13. *If f is a positive, nondecreasing, subadditive function on $(0, \infty)$, then $\limsup_{t \to 0+} (f(kt)/f(t)) \geqslant k$ for all $k \leqslant 1$.*

PROOF. If not, then $\exists \delta > 0$, $h < k < 1$ such that $f(kt) \leqslant hf(t)$ for all $t \leqslant \delta$. Let $0 < t \leqslant \delta$. For some integer m, $k^m\delta \leqslant t \leqslant k^{m-1}\delta$. By our assumption and the monotonicity of f,

$$f(t) \leqslant f(k^{m-1}\delta) \leqslant h^{m-1}f(\delta) \leqslant \frac{t}{k\delta}\left(\frac{h}{k}\right)^{m-1} f(\delta).$$

As $t \to 0$, $m \to \infty$ so that $\limsup_{t \to 0+} (f(t)/t) = 0$.

Let $s > 0$. Since f is positive and subadditive,

$$0 < f(s) \leqslant nf\left(\frac{s}{n}\right) = s\,\frac{f(s/n)}{s/n} \quad \text{for all } n.$$

Hence

$$0 < f(s) \leqslant s \cdot \limsup_{t \to 0+} \frac{f(t)}{t} = 0,$$

a contradiction. ∎

5. Extensions of the main theorem. In infinite-dimensional space, defining $x > 0$ to mean $x \in \operatorname{rel\,int} K$ may make no sense because $\operatorname{rel\,int} K$ may be empty. Even when $\operatorname{rel\,int} K$ is not empty, the naive generalization of the Perron-Frobenius theorem, in which $A > 0$ is interpreted as $Ax > 0$ whenever $x \geqslant 0$, is not true. Thus we cannot usefully generalize the definition of $A > 0$ to infinite-dimensional spaces by considering the images of single points. The condition that Ax and Ay belong to the relative interior of $\operatorname{Sp}\{Ax, Ay\} \cap K$ whenever $x, y \geqslant 0$, which holds for $A > 0$ in R^m, can be used in infinite-dimensional spaces as well, however. It is equivalent, by (3.12), to $d(Ax, Ay) < \infty$ for all $x, y \geqslant 0$. To obtain the contraction property, we need this uniformly. Accordingly, we define A to be *d-positive* if $A \geqslant 0$ and $\sup\{d(Ax, Ay): x, y \geqslant 0\} < \infty$. The following analogue of Theorem 4.1 implies that there is no projective metric which contracts for more linear transformations than d.

THEOREM 5.1. *In an arbitrary real vector space X, let $\geqslant$ be the ordering induced by a cone K satisfying:*

(5.2) *K is closed in some locally convex topology on X;*

(5.3) *there exists a linear functional w such that $w(x) > 0$ for all $x \geqslant 0$.*

Let d be Hilbert's metric on K. Then every d-positive linear transformation is a contraction with respect to d. Conversely, if D is a projective metric on K such that every d-positive linear transformation is a contraction with respect to D, then there exists a continuous, strictly increasing function $f: R_+ \to R_+$ such that $D(x, y) = f(d(x, y))$ whenever $d(x, y) < \infty$. Furthermore, the contraction ratios satisfy $k_d(A) \leqslant k_D(A)$ for all nonnegative A.

We remark that (5.3) is equivalent to the existence of a hyperplane intersecting every ray of K in exactly one point.[8] We also note that, if relint K is not empty, then (5.2) reduces to (3.2) (Kelley and Namioka, (1963, 5H)), and that, if X is normed and separable, then (5.3) reduces to pointedness of K (Klee, (1955)). Indeed, (5.2) and (5.3) hold for most of the interesting cones which satisfy (3.1) and (3.2).[9]

PROOF. The direct part is now merely a restatement of Proposition 3.14. The proof of the converse part of Theorem 4.1 remains correct when positive is replaced by d-positive and $x, y > 0$ by $d(x, y) < \infty$, but needs the following amplifications.

To establish (4.4), let ϕ be a linear functional such that $\phi(u) < 0 \leqslant \phi(z)$ for all $z \geqslant 0$.[10] Let w be as in (5.3) and let $g = \phi + \lambda w$ where $\lambda > 0$ is small. Then $g(u) < 0 < g(z)$ for all $z \in K$. Define h similarly.

To show that the linear transformation $\bar{A}$ defined by (4.5) is d-positive we observe that $d(Au_0, Av_0) < \infty$ (see Figure 3) and, for every $z, z' \geqslant 0$, $\bar{A}z$ and $\bar{A}z'$ are positive linear combinations of Au_0 and Av_0. Hence $d(\bar{A}z, \bar{A}z') \leqslant d(Au_0, Av_0) < \infty$.

This proves all of Theorem 5.1 except that it proves $k_d(A) \leqslant k_D(A)$ only for d-positive A. To extend the proof to nonnegative A, we need to replace Lemma 4.11 by

LEMMA 5.4. *For all $\epsilon > 0$ and all nonnegative A,*

$$k_d(A) = \sup\left\{ \frac{d(Ax, Ay)}{d(x, y)} : 0 < d(x, y) < \epsilon \right\}.$$

PROOF. The proof is the same as that of Lemma 4.11 except that the case $d(x, y) = \infty$ can no longer be ignored because $d(Ax, Ay)$ might also be infinite. Using (3.15), however, one can easily see that if $d(x, y) = d(Ax, Ay) = \infty$, then

$$x_n = \frac{nx + y}{n + 1} \quad \text{and} \quad y_n = \frac{ny + x}{n + 1}$$

satisfy

$$d(x_n, y_n) < \infty \quad \text{and} \quad \frac{d(Ax_n, Ay_n)}{d(x_n, y_n)} \to 1 = \frac{\infty}{\infty}. \quad \blacksquare$$

To clarify Theorem 5.1, we note that in R^m, d-positive is a weaker condition than positive. Specifically, a linear transformation mapping the nonnegative cone into the relative interior of one of its faces is d-positive but not positive. Theorem 5.1 therefore does not specialize to Theorem 4.1 when $X = R^m$.

Note also that Theorem 5.1 does not say that all transformations for which convergence in direction (1.1) occurs are contractions with respect to Hilbert's metric, nor does it preclude transformations which are contractions with respect to some

[8] When such a hyperplane exists, or alternatively, when the space is normed, it is natural to normalize the nonnegative vectors to the hyperplane or the unit ball. It is then possible to discuss ordinary metrics on the normalized vectors instead of projective metrics. We did not adopt this approach in §3 because there are interesting cones for which neither normalization is possible.

[9] For instance, for the first example mentioned in footnote 7, the cone of nonnegative functions in an L^p space ($1 \leqslant p < \infty$), condition (5.2) is obvious, and (5.3) holds with $w(x) = \int xy$ where y is any positive function in L^q ($1/p + 1/q = 1$). For the second example, however, the cone of nonnegative measurable functions, neither (5.2) nor (5.3) holds. This follows from the fact that there exists no nonnegative functional on this cone (Kelley and Namioka (1963, 18H)) and the separation theorem in footnote 10.

[10] If K is a closed convex cone in a locally convex space and $u \notin K$ then there exists a linear functional ϕ such that $\phi(x) \geqslant 0$ for all $x \in K$ and $\phi(u) < 0$. This is a standard application of the separating hyperplane theorem.

metric but not Hilbert's. For example, if $K = R^2_+$ and

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

then $A^n x$ converges in direction to $\binom{0}{1}$ for all $x \geqslant 0$. However, A carries the unit vectors into points between which the Hilbert distance is infinite. Hence, by Proposition 3.14, A is not a contraction with respect to Hilbert's metric. On the other hand, A is a contraction on R^m_+ with respect to the metric $D(x, y) = \sum_i |\bar{x}_i - \bar{y}_i|$ where $\bar{x}_i = x_i / \sum_j x_j$. It is well known, in fact, that D contracts under all nonnegative matrices having at least one row in which all entries are positive. (Such matrices are called Markov matrices.) Thus, D can give us many convergence theorems which Hilbert's metric cannot. Theorem 5.1 implies, however, that D cannot cover all situations covered by Hilbert's metric.

6. Remarks on Birkhoff's program. We now discuss in more detail how convergence in direction of $A^n x$ can be proved using Hilbert's metric. We consider only normal cones in normed linear spaces. (A cone K is called normal if $\exists \epsilon > 0$ such that $\|x + y\| \geqslant \epsilon$ whenever $x, y \in K$ and $\|x\| = \|y\| = 1$. In R^m, a closed cone is normal if and only if it is pointed.) On such cones it is easily verified that Hilbert's metric is complete and that convergence in Hilbert's metric is the same as convergence in direction (e.g., Kohlberg and Neyman, 1979). By Proposition 3.14 therefore

THEOREM 6.1. *Let $\geqslant$ be the ordering on a normed linear space induced by a closed normal cone K and let d be Hilbert's metric on K. If A is a nonnegative linear transformation and $d(Ax, Ay)$ is bounded for $x, y \in K$, then $\exists x_0 \geqslant 0$ such that*

$$A^n x / \|A^n x\| \to x_0 \quad \textit{for all } x \geqslant 0.$$

For many purposes, it is useful to have a method of calculating the diameter $\Delta = \sup\{d(Ax, Ay) : x, y \in K\}$ in Proposition 3.14. In a finite dimensional situation, Δ is obtained by maximizing $d(Ax, Ay)$ when x and y range over the extreme rays of K. This follows from

LEMMA 6.2. *$d(x, y)$ is quasi-convex in each of its arguments.*

PROOF. By (3.11), $(m(x, y) + m(z, y))y \leqslant x + z \leqslant (M(x, y) + M(z, y))y$ (when the right-hand side is finite). Hence,

$$d\left(\frac{x+z}{2}, y\right) = d(x + z, y) \leqslant \log \frac{M(x, y) + M(z, y)}{m(x, y) + m(z, y)}$$

$$\leqslant \log \max\left\{ \frac{M(x, y)}{m(x, y)}, \frac{M(z, y)}{m(z, y)} \right\} = \max\{d(x, y), d(z, y)\}. \quad \blacksquare$$

In particular, when $K = R^m_+$, we have $\Delta = \max_{i,j} d(a_i, a_j)$, where a_i is the image of the ith unit vector, that is, the ith column of the matrix of A, so by Proposition 3.14

$$k_d(A) = \frac{\sqrt{\Gamma} - 1}{\sqrt{\Gamma} + 1}, \quad \text{where } \Gamma = \max_{i,j,k,l} \frac{a_{ki} a_{lj}}{a_{kj} a_{li}}. \tag{6.3}$$

For a sequence of positive matrices A_n, it follows that the condition $\prod_1^\infty k_n = 0$ mentioned in the introduction holds if all elements of all A_n lie between two positive numbers (e.g., Golubitsky, Keeler, and Rothschild, 1975), or more generally if $\sum_1^\infty r_n = \infty$ where r_n is the ratio of the minimum to the maximum element of A_n.

References

Birkhoff, G. (1957). Extensions of Jentzsch's Theorem. *Trans. Amer. Math. Soc.* **85** 219–227.

———. (1967). *Lattice Theory*, 3rd ed. American Mathematical Society, Providence, Rhode Island.

Bushell, P. J. (1973). Hilbert's Metric and Positive Contraction Mappings in a Banach Space. *Arch. Rational Mech. Anal.* **52** 330–338.

Debreu, G. and Herstein, I. N. (1953). Nonnegative Square Matrices. *Econometrica* **21** 597–607.

Furstenburg, H. and Kesten, H. (1960). Products of Random Matrices. *Ann. Math. Statist.* **31** 457–469.

Golubitsky, M., Keeler, E. B. and Rothschild, M. (1975). Convergence of the Age Structure: Applications of the Projective Metric. *Theoret. Population Biology* **7** 84–93.

Karlin, S. and Taylor, H. M. (1975). *A First Course in Stochastic Processes*. Academic Press, New York.

Kelley, J. L. and Namioka, I. (1963). *Linear Topological Spaces.* D. Van Nostrand.

Klee, V. L. (1955). Separation Properties of Convex Cones. *Proc. Amer. Math. Soc.* **6** 313–318.

Kohlberg, E. and Neyman, A. (1979). On the Relationship Between Convergence in Hilbert's Projective Metric and Convergence in Direction. Harvard Business School Working Paper 79-54.

——— and Pratt, J. W. (1979). The Contraction Mapping Approach to the Perron-Frobenius Theory: Why Hilbert's Metric? Harvard Business School Working Paper 79-4.

GRADUATE SCHOOL OF BUSINESS ADMINISTRATION, HARVARD UNIVERSITY, SOLDIERS FIELD, BOSTON, MASSACHUSETTS 02163