PACIFIC JOURNAL OF MATHEMATICS
Vol. 21, No. 2, 1967

CONCERNING NONNEGATIVE MATRICES AND DOUBLY STOCHASTIC MATRICES

RICHARD SINKHORN AND PAUL KNOPP

This paper is concerned with the condition for the convergence to a doubly stochastic limit of a sequence of matrices obtained from a nonnegative matrix A by alternately scaling the rows and columns of A and with the condition for the existence of diagonal matrices D_1 and D_2 with positive main diagonals such that D_1AD_2 is doubly stochastic.

The result is the following. The sequence of matrices converges to a doubly stochastic limit if and only if the matrix A contains at least one positive diagonal. A necessary and sufficient condition that there exist diagonal matrices D_1 and D_2 with positive main diagonals such that D_1AD_2 is both doubly stochastic and the limit of the iteration is that $A \neq 0$ and each positive entry of A is contained in a positive diagonal. The form D_1AD_2 is unique, and D_1 and D_2 are unique up to a positive scalar multiple if and only if A is fully indecomposable.

Sinkhorn [6] has shown that corresponding to each positive square matrix A there is a unique doubly stochastic matrix of the form D_1AD_2 where D_1 and D_2 are diagonal matrices with positive main diagonals. The matrices D_1 and D_2 are themselves unique up to a scalar factor. The matrix D_1AD_2 can be obtained as a limit of the sequence of matrices generated by alternately normalizing the rows and columns of A. But it was shown by example that for nonnegative matrices the iteration does not always converge, and even when it does, the D_1 and D_2 do not always exist.

Marcus and Newman [4] and Maxfield and Minc [5] gave some consideration to this problem for symmetric matrices.

In a recent communication with H. Schneider, the authors learned that Brualdi, Parter and Schneider [2] have independently obtained some of the results of this paper by employing different techniques.

DEFINITIONS. If A is an $N \times N$ matrix and σ is a permutation of $\{1, \cdots, N\}$, then the sequence of elements $a_{1,\sigma(1)}, \cdots, a_{N,\sigma(N)}$ is called the diagonal of A corresponding to σ. If σ is the identity, the diagonal is called the main diagonal.

If A is a nonnegative square matrix, A is said to have total support if $A \neq 0$ and if every positive element of A lies on a positive diagonal. A nonnegative matrix that contains a positive diagonal is said to have support.

The notation $A[\mu \mid \nu]$, $A(\mu \mid \nu]$, etc. is that of [3, pp. 10-11].

THEOREM. *Let A be a nonnegative $N \times N$ matrix. A necessary and sufficient condition that there exist a doubly stochastic matrix B of the form D_1AD_2 where D_1 and D_2 are diagonal matrices with positive main diagonals is that A has total support. If B exists then it is unique. Also D_1 and D_2 are unique up to a scalar multiple if and only if A is fully indecomposable.*

A necessary and sufficient condition that the iterative process of alternately normalizing the rows and columns of A will converge to a doubly stochastic limit is that A has support. If A has total support, this limit is the described matrix D_1AD_2. If A has support which is not total, this limit cannot be of the form D_1AD_2.

Proof. We first demonstrate uniqueness. Suppose $B = D_1AD_2$ and $B' = D_1'AD_2'$ are doubly stochastic where $D_1 = \operatorname{diag}(x_1, \cdots, x_N)$, $D_2 = \operatorname{diag}(y_1, \cdots, y_N)$, $D_1' = \operatorname{diag}(x_1', \cdots, x_N')$, and $D_2' = \operatorname{diag}(y_1', \cdots, y_N')$. If $p_i = x_i'/x_i$, $q_j = y_j'/y_j$,

$$\sum_i x_i a_{ij} y_j = 1 \; ; \qquad \sum_j x_i a_{ij} y_j = 1$$
$$\sum_i p_i x_i a_{ij} q_j y_j = 1 \; ; \qquad \sum_j p_i x_i a_{ij} q_j y_j = 1 \; .$$

Let $E_j = \{i \mid a_{ij} > 0\}$, $F_i = \{j \mid a_{ij} > 0\}$ and put

$$m = \{i \mid p_i = \min_i p_i = \underline{p}\},\ M = \{j \mid q_j = \max_j q_j = \bar{q}\} \; .$$

Pick $i_0 \in m$, $j_0 \in M$. Then $q_{j_0} = (\sum_i p_i x_i a_{ij_0} y_{j_0})^{-1} \leqq p_{i_0}^{-1}$ and similarly $p_{i_0} \geqq q_{j_0}^{-1}$, forcing $q_{j_0} = p_{i_0}^{-1} = \underline{p}^{-1}$. But equality is possible only if $p_i = \underline{p}$ when $i \in E_{j_0}$. Whence $p_i = \underline{p}$ when $i \in E_j$ and $j \in M$. Thus $\bigcup_{j \in M} E_j \subseteqq m$ and it follows that $A(m \mid M] = 0$. In the same way $p_{i_0} = q_{j_0}^{-1}$ is possible only if $q_j = \bar{q}$ for all $j \in F_{i_0}$. Whence $q_j = \bar{q}$ when $j \in F_i$ and $i \in m$. Thus $\bigcup_{i \in m} F_i \subseteqq M$ and it follows that $A[m \mid M) = 0$.

On $m \times M$, $p_i q_j = \underline{p}\bar{q} = 1$ and it follows that $B[m \mid M] = B'[m \mid M]$ is doubly stochastic. In particular m and M must have the same size.

If A is fully indecomposable, $A(m \mid M]$ and $A[m \mid M)$ thus cannot exist. In such a case $A = A[m \mid M]$. Thus $D_1AD_2 = D_1'AD_2'$, and D_1 and D_2 are themselves unique up to a scalar multiple.

If $A(m \mid M]$ and $A[m \mid M)$ exist, $B(m \mid M)$ and $B'(m \mid M)$ exist and are each doubly stochastic matrices of order less than N. Furthermore $B(m \mid M) = D_1''A(m \mid M)D_2''$ and $B'(m \mid M) = D_1'''A(m \mid M)D_2'''$ where the D's are diagonal matrices with positive main diagonals. The argument may be repeated on these submatrices until $D_1AD_2 = D_1'AD_2'$ is established. Note that D_1 and D_2 no longer need be unique up to a scalar multiple.

The necessity of total support for the existence of D_1AD_2 is an immediate consequence of the celebrated theorem of G. Birkhoff [1] which states that the set of doubly stochastic matrices of order N is the convex hull of the $N \times N$ permutation matrices.

The sufficiency of the condition and the remarks concerning the iteration will follow in part from the following lemmas.

LEMMA 1. *If A is a row stochastic matrix, and $\beta_1, \cdots, \beta_N$ are the column sums of A, then $\prod_{k=1}^N \beta_k \leqq 1$, with equality only if each $\beta_k = 1$.*

Proof. Let A have column sums $\beta_1, \cdots, \beta_N$. Of course, each $\beta_k \geqq 0$ and $\sum_{k=1}^N \beta_k = N$. By the arithmetic-geometric mean inequality

$$\prod_{k=1}^{N} \beta_k \leqq \left[(1/N) \sum_{k=1}^{N} \beta_k\right]^N = 1$$

with equality occurring only if each $\beta_k = 1$.

LEMMA 2. *Let $A = (a_{ij})$ be an $N \times N$ matrix with total support and suppose that if $1 \leqq i, j \leqq N$, $\{x_{i,n}\}$ and $\{y_{j,n}\}$ are positive sequences such that $x_{i,n}y_{j,n}$ converges to a positive limit for each i, j such that $a_{ij} \neq 0$. Then there exist convergent positive sequences $\{x'_{i,n}\}$, $\{y'_{j,n}\}$ with positive limits such that $x'_{i,n}y'_{j,n} = x_{i,n}y_{j,n}$ for all i, j, n.*

Proof. Consider first the case in which A is fully indecomposable. Let $E^{(1)} = \{1\}$, $F^{(1)} = \{j \mid a_{1j} > 0\}$, and $E^{(s)} = \{i \notin \bigcup_{k=1}^{s-1} E^{(k)} \mid$ for some $j \in F^{(s-1)}, a_{ij} > 0\}$, $F^{(s)} = \{j \notin \bigcup_{k=1}^{s-1} F^{(k)} \mid$ for some $i \in E^{(s)}, a_{ij} > 0\}$ when $s > 1$. The sets $E^{(s)}$ and $F^{(s)}$ are void for sufficiently large s, e.g., for $s > N$. Define $E = \bigcup_k E^{(k)}$ and $F = \bigcup_k F^{(k)}$. Since A has total support, the first row of A contains a nonzero element; thus $F^{(1)}$ is nonvoid. Since $F^{(1)} \subseteqq F$, F is nonvoid. Also since $\{1\} = E^{(1)} \subseteqq E$, E is nonvoid.

Suppose E is a proper subset of $\{1, \cdots, N\}$. Pick $i \notin E, j \in F$. Then $j \in F^{(s)}$ for some s. Since $i \notin E$, certainly $i \notin \bigcup_{k=1}^{s} E^{(k)}$. Certainly then it could not be that $a_{ij} > 0$ for then $i \in E^{(s+1)} \subseteqq E$, a contradiction. Whence $i \notin E, j \in F \Rightarrow a_{ij} = 0$, i.e., $A(E \mid F] = 0$. In the same way it follows that if $F \neq \{1, \cdots, N\}$, $A[E \mid F) = 0$.

Define an $N \times N$ matrix $H = (h_{ij})$ as follows. If $a_{ij} = 0$, set $h_{ij} = 0$. If $a_{ij} \neq 0$ and a_{ij} lies on t positive diagonals in A, set $h_{ij} = t/\tau$ where τ is the total number of positive diagonals in A. Then H is doubly stochastic and $h_{ij} = 0$ if and only if $a_{ij} = 0$. Suppose E contains u elements and F contains v elements. Since $H(E \mid F] = 0$, $\sum_{i \in E} \sum_{j \in F} h_{ij} = v$, and since either $F = \{1, \cdots, N\}$ or $H[E \mid F) = 0$, $\sum_{i \in E} \sum_{j \in F} h_{ij} = u$. Thus E and F have the same number of elements.

But E and F cannot be proper subsets of $\{1, \cdots, N\}$ if A is assumed to be fully indecomposable. Thus $E = F = \{1, \cdots, N\}$.

Define $x'_{i,n} = x_{1,n}^{-1} x_{i,n}$ and $y'_{j,n} = x_{1,n} y_{j,n}$ for all i, j, n. Then $x'_{i,n} y'_{j,n} = x_{i,n} y_{j,n}$ for all i, j, n. Since $x'_{1,n} = 1$ for all n, certainly $x'_{1,n} \to 1$. For $j \in F^{(1)}$, $y'_{j,n} = x'_{1,n} y'_{j,n} = x_{1,n} y_{j,n}$ has a positive limit.

Inductively suppose that it is known that $x'_{i,n}$ and $y'_{j,n}$ converge to positive limits when $i \in \bigcup_{k=1}^{s-1} E^{(k)}$ and $j \in \bigcup_{k=1}^{s-1} F^{(k)}$. For $i \in E^{(s)}$ there is a $j_{s-1} \in F^{(s-1)}$ such that $a_{ij_{s-1}} > 0$. Thus $x'_{i,n} = x'_{i,n} y'_{j_{s-1},n} / y'_{j_{s-1},n} = x_{i,n} y_{j_{s-1},n} / y'_{j_{s-1},n}$ has a positive limit. Then for $j \in F^{(s)}$ there is a $i_s \in E^{(s)}$ such that $a_{i_s j} > 0$. Whence $y'_{j,n} = x'_{i_s,n} y'_{j,n} / x'_{i_s,n} = x_{i_s,n} y_{j,n} / x'_{i_s,n}$ has a positive limit. This completes the proof in case A is fully indecomposable.

If A is not fully indecomposable, then neither is the corresponding doubly stochastic matrix H. This means that there exist permutations P and Q such that $PHQ = H_1 \oplus \cdots \oplus H_g$ where each H_k is doubly stochastic and fully indecomposable. Thus also $PAQ = A_1 \oplus \cdots \oplus A_g$ where each A_k has total support and is fully indecomposable. The above argument may be repeated on each of the A_k.

Now we return to the theorem. Suppose A has support. Define an iteration on A as follows.

Let $x_{i,0} \equiv 1$, $y_{j,0} \equiv (\sum_{i=1}^N a_{ij})^{-1}$ and set $x_{i,n+1} = \alpha_{i,n}^{-1} x_{i,n}$, $y_{j,n+1} = \beta_{j,n}^{-1} y_{j,n}$ where

$$\alpha_{i,n} = \sum_{j=1}^{N} x_{i,n} a_{ij} y_{j,n} \ ; \qquad \beta_{j,n} = \sum_{i=1}^{N} \alpha_{i,n}^{-1} x_{i,n} a_{ij} y_{j,n} \ ,$$

$i = 1, \cdots, N$, $j = 1, \cdots, N$, $n = 0, 1, \cdots$. Note that $(x_{i,n} a_{ij} y_{j,n})$ is column stochastic and $(x_{i,n+1} a_{ij} y_{j,n})$ is row stochastic. Then in particular

$$y_{j,n} = \left(\sum_{i=1}^{N} x_{i,n} a_{ij} \right)^{-1} \leqq x_{i_0,n}^{-1} a_{i_0 j}^{-1} \leqq x_{i_0,n}^{-1} a^{-1}$$

where i_0 is such that $a_{i_0 j} > 0$ and a is the minimal positive a_{ij}. Thus $x_{i,n} y_{j,n} \leqq a^{-1}$ if $a_{ij} > 0$.

Let A have a positive diagonal corresponding to a permutation σ, and set $s_n = \prod_{i=1}^N x_{i,n} y_{\sigma(i),n}$ and $s'_n = \prod_{i=1}^N x_{i,n+1} y_{\sigma(i),n}$. By Lemma 1 and the preceding remark, $s_n \leqq s'_n \leqq s_{n+1} \leqq a^{-N}$. Thus $s_n \to L$ and $s'_n \to L$ where $0 < L \leqq a^{-N}$. Whence $\prod_{j=1}^N \beta_{j,n} = s'_n / s_{n+1} \to 1$. This is impossible unless each $\beta_{j,n} \to 1$ since $\prod_{k=1}^N \beta_k$ has a unique maximal value of 1 only when $\beta_1 = \cdots = \beta_N = 1$. Similarly each $\alpha_{i,n} \to 1$.

Thus if A has a positive diagonal, the limit points of the sequence of matrices generated by the iteration are doubly stochastic. However, two such limit points are diagonally equivalent. Suppose that A_n is the nth matrix in the iteration and that $A_{n_k} \to B$ and $A_{m_k} \to C$. Observe that for any given pair i, j $b_{ij} \neq 0 \Leftrightarrow c_{ij} \neq 0$. For any permutation σ, $\prod_{i=1}^{N} b_{i,\sigma(i)} = \prod_{i=1}^{N} c_{i,\sigma(i)} = L \prod_{i=1}^{N} a_{i,\sigma(i)}$. Then certainly $b_{ij} \neq 0 \Rightarrow c_{ij} \neq 0$, for suppose $b_{i_0 j_0} \neq 0$. Then $b_{i_0 j_0}$ lies on a positive diagonal. The corresponding diagonal in C would have a positive product. Thus $c_{i_0 j_0} \neq 0$. In the same way $c_{ij} \neq 0 \Rightarrow b_{ij} \neq 0$. If in addition A has total support then $a_{ij} \neq 0 \Leftrightarrow b_{ij} \neq 0 \Leftrightarrow c_{ij} \neq 0$.

By construction there exist matrices $\tilde{D}_{1,k} = \operatorname{diag}(w_{1,k}, \cdots, w_{N,k})$ and $\tilde{D}_{2,k} = \operatorname{diag}(z_{1,k}, \cdots, z_{N,k})$ with positive main diagonals such that $A_{m_k} = \tilde{D}_{1,k} A_{n_k} \tilde{D}_{2,k}$. For $b_{ij} > 0$, $w_{i,k} z_{j,k} \to c_{ij} b_{ij}^{-1}$. By Lemma 2 there exist positive sequences $\{w'_{i,k}\}$ and $\{z'_{j,k}\}$ converging to positive limits such that $w'_{i,k} z'_{j,k} = w_{i,k} z_{j,k}$, for all i, j, k. If

$$D_1 = \lim_{k\to\infty} \operatorname{diag}(w'_{1,k}, \cdots, w'_{N,k}) \quad \text{and} \quad D_2 = \lim_{k\to\infty} \operatorname{diag}(z'_{1,k}, \cdots, z'_{N,k}) ,$$

then $C = D_1 B D_2$. By the uniqueness part of the theorem, $B = C$. It follows that the iteration converges. It is clear from Birkhoff's theorem that no limit to the iteration is possible without at least one positive diagonal.

Suppose A has total support. Let $D_{1,n} = \operatorname{diag}(x_{1,n}, \cdots, x_{N,n})$ and $D_{2,n} = \operatorname{diag}(y_{1,n}, \cdots, y_{N,n})$. Then $B = \lim_{n\to\infty} D_{1,n} A D_{2,n}$ exists and $b_{ij} \neq 0 \Leftrightarrow a_{ij} \neq 0$. When $a_{ij} > 0$, $x_{i,n} y_{j,n} \to b_{ij} a_{ij}^{-1}$. By Lemma 2 there are convergent positive sequences $\{x'_{i,n}\}$, $\{y'_{j,n}\}$ with positive limits such that $x'_{i,n} y'_{j,n} = x_{i,n} y_{j,n}$ for all i, j, n. Let $D_1 = \lim_{n\to\infty} \operatorname{diag}(x'_{1,n}, \cdots, x'_{N,n})$ and $D_2 = \lim_{n\to\infty} \operatorname{diag}(y'_{1,n}, \cdots, y'_{N,n})$. Then $B = D_1 A D_2$.

Finally we observe that if A has support which is not total, then by Birkhoff's theorem, there is a nonzero element of A which tends to zero in the iteration. In fact every nonzero element of A which is not on a positive diagonal must do so. If the limit matrix could be put in the form $D_1 A D_2$ then some term $x_i a_{ij} y_j = 0$ where $a_{ij} > 0$. But then either $x_i = 0$ or $y_j = 0$. The former leads to a row of zeros and the latter to a column of zeros in $D_1 A D_2$. In either case $D_1 A D_2$ could not be doubly stochastic.

The authors wish to express their gratitude to Professor Albert Newhouse for his help on numerical examples connected with this problem. They also are indebted to Professors M. Marcus and H. Schneider and the referee for suggestions which simplified some of the proofs.

REFERENCES

1. G. Birkhoff, *Tres observaciones sobre el algebra lineal,* Univ. Nac. Tucumán Rev. Ser. (A) **5** (1946), 147-150.

2. R. Brualdi, S. Parter and H. Schneider, *The diagonal equivalence of a nonnegative matrix to a stochastic matrix* (to appear).

3. M. Marcus and H. Minc, *A survey of matrix theory and matrix inequalities,* Allyn

and Bacon, Boston, 1964.
4. M. Marcus and M. Newman, unpublished paper.
5. J. Maxfield and H. Minc, *A doubly stochastic matrix equivalent to a given matrix,* Notices Amer. Math. Soc. **9** (1962), 309.
6. R. Sinkhorn, *A relationship between arbitrary positive matrices and doubly stochastic matrices,* Ann. Math. Statist. **35** (1964), 876-879.

Received February 20, 1966.

UNIVERSITY OF HOUSTON
HOUSTON, TEXAS