Relative Entropy Under Mappings by Stochastic Matrices*

Joel E. Cohen
Rockefeller University
1230 York Avenue
New York, New York 10021-6399

Yoh Iwasa
Department of Biology
Kyushu University 33
Fukuoka 812, Japan

Gh. Rautu
Centre of Mathematical Statistics
Bd. Magheru 22
Ro-70158, Bucuresti, Romania

Mary Beth Ruskai
Department of Mathematics
University of Lowell
Lowell, Massachusetts 01854

Eugene Seneta
Department of Mathematical Statistics
University of Sydney
N.S.W. 2006, Australia

and

Gh. Zbaganu
Centre of Mathematical Statistics
Bd. Magheru 22
Ro-70158, Bucuresti, Romania

Submitted by Richard A. Brualdi

*In honor of John Hajnal and Marius Iosifescu.

655 Avenue of the Americas, New York, NY 10010 0024-3795/93/$5.00

ABSTRACT

The relative g-entropy of two finite, discrete probability distributions $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$ is defined as $H_g(x, y) = \Sigma_k x_k g(y_k/x_k - 1)$, where $g : (-1, \infty) \to \Re$ is convex and $g(0) = 0$. When $g(t) = -\log(1 + t)$, then $H_g(x, y) = \Sigma_k x_k \log(x_k/y_k)$, the usual relative entropy. Let $P_n = \{x \in \Re^n : \Sigma_i x_i = 1, x_i > 0 \ \forall i\}$. Our major result is that, for any $m \times n$ column-stochastic matrix A, the contraction coefficient defined as $\eta_g(A) = \sup\{H_g(Ax, Ay)/H_g(x, y) : x, y \in P_n, x \neq y\}$ satisfies $\eta_g(A) \leqslant 1 - \alpha(A)$, where $\alpha(A) = \min_{j,k} \Sigma_i \min(a_{ij}, a_{ik})$ is Dobrushin's coefficient of ergodicity. Consequently, $\eta_g(A) < 1$ if and only if A is scrambling. Upper and lower bounds on$\eta_g(A)$ are established. Analogous results hold for Markov chains in continuous time.

1. INTRODUCTION

Since Boltzmann (1877) introduced the concept of entropy in classical statistical mechanics, significant applications of entropy and related functionals have been found in many fields, including demography (e.g., White 1986, Goldman and Lord 1986), economics (Theil 1967, Georgescu-Roegen 1971), information theory (e.g., Shannon 1948, Csiszár and Körner 1981), physics (e.g., Jaynes 1957, Lieb 1975, Wehrl 1978, Thirring 1983), population biology (e.g., Demetrius 1985, Iwasa 1988), probability theory (e.g., Kelly 1979, Seneta 1982, Ellis 1985, Liggett 1985), and statistics (e.g., Kullback and Liebler 1951, Ali and Silvey 1966, Kullback 1968, Liese and Vajda 1987, Joe 1989). Many applications concern the changes over time in the difference between two distributions, when this difference is measured by a convex functional which we shall call the relative entropy. So many names have been used by so many authors for what we call relative entropy that we do not even attempt a historical review; see Good (1983) and Wehrl (1978).

This paper provides new information about how the relative entropy changes in a finite-state Markov chain in discrete time and in a finite-state Markov process in continuous time. This information can be used to bound the rates of convergence to equilibrium of ergodic Markov chains and Markov processes; but we shall not develop such applications here. Section 2 gives definitions and background. Sections 3–5 describe the relative entropy of two finite positive probability vectors each multiplied by a single finite stochastic matrix. Our major new result (Section 3) is that $H_g(Ax, Ay) \leqslant [1 - \alpha(A)]H_g(x, y)$, where H_g is the relative-entropy functional defined in terms of the convex function g, α is Dobrushin's coefficient of ergodicity $\alpha(A) = \min_{j,k} \Sigma_i \min(a_{ij}, a_{ik})$, and A is an $m \times n$ column-stochastic matrix. Section 4 shows that the contraction coefficient defined as $\eta_g(A) = \sup\{H_g(Ax, Ay)/H_g(x, y) : x, y \in P_n\}$ satisfies $\eta_g(A) < 1$ if and only if A is scrambling. Section 5 gives and compares upper and lower bounds for $\eta_g(A)$. Section 6 describes the relative entropy of two probability vectors multiplied by successive powers of a single stochastic matrix: under suitable conditions on the eigenstructure of A and the smoothness of g, the asymptotic rate of contraction of the relative g-entropy under repeated multiplication by A is the square of the second largest eigenvalue of A. Section 7 gives analogous results for an exponentiated infinitesimal generator of a finite-state Markov process in continuous time.

2. DEFINITIONS AND BACKGROUND

Let m, n, and d be finite positive integers. All matrices will be $m \times n$ or $d \times d$. All vectors will be $n \times 1$ or $d \times 1$ column vectors. Vectors followed by ^{T} will be transposed, i.e., $1 \times n$ or $1 \times d$ row vectors. As usual, the l_p-norms are defined for a d-vector x and for $1 \leqslant p < \infty$ by $\|x\|_p = (\Sigma|x_i|^p)^{1/p}$.

A matrix or vector in which all elements are nonnegative real numbers will be called nonnegative. A matrix or vector in which all elements are positive real numbers will be called positive. Let N_d be the set of nonnegative probability d-vectors, i.e., $N_d = \{x \in R^d : x_i \geqslant 0, \Sigma_i x_i = 1\}$. Let P_d be the set of positive probability d-vectors, i.e., $P_d = \{x \in N_d : x_i > 0 \text{ for all } i\}$. If $x \in P_d$ with elements x_i, let $1 - x$ denote the d-vector with elements $1 - x_i$. Also, let $e_i \in N_d$ denote the ith unit vector with 1 in the ith position and all other elements 0.

A stochastic $m \times n$ matrix is a matrix each column of which belongs to N_m (i.e., a nonnegative matrix with all column sums 1); such a matrix is sometimes called column-stochastic to distinguish it from a row-stochastic matrix, a nonnegative matrix with all row sums 1. A nonnegative matrix is called row-allowable if each row contains at least one positive element. A column-stochastic matrix need not be row-allowable. A matrix with at least one positive row (i.e., all elements of a row positive) is called row-positive. A column-stochastic row-positive matrix is sometimes called a Markov matrix (e.g., Iosifescu 1980, p. 57). Clearly, a positive stochastic matrix is row-allowable and row-positive. A nonnegative $d \times d$ matrix A is called primitive if A^k is positive for some positive integer k.

A column-stochastic $m \times n$ matrix is called a scrambling matrix (Hajnal 1958, p. 235) if any submatrix consisting of two columns has a row both elements of which are positive; i.e., $A = (a_{ij})$ is scrambling if, for all j and k such that $1 \leqslant j < k \leqslant n$, there exists an i such that $1 \leqslant i \leqslant m$ and $a_{ij}a_{ik} > 0$.

Every row-positive matrix is scrambling, but not conversely. A $d \times d$ scrambling matrix need not be primitive, and a primitive matrix need not be scrambling (Remark 4.3).

As usual, a real-valued function h on some convex subset D of a vector space over the reals is called convex if, for all $p \in [0,1]$ and all $s, t \in D$, $h(ps + [1-p]t) \leqslant ph(s) + (1-p)h(t)$. A convex function h is called strictly convex if the preceding inequality is strict whenever $s \neq t$ and $p(1-p) \neq 0$.

A real-valued function h on some convex cone D of a vector space over the reals is called homogeneous if, for all $x \in D$ and all nonnegative λ, $h(\lambda x) = \lambda h(x)$.

Logarithms are to the base e. The function $h(t) = t \log t$ is defined on $(0, \infty)$ and, by defining $h(0) = 0$, extends continuously to $[0, \infty)$. This $h(t) = t \log t$ is strictly convex on $[0, \infty)$.

DEFINITION 2.1. For any two positive d-vectors $x = (x_i)$ and $y = (y_i)$, whether or not x and y are probability vectors, define the relative entropy $H(x, y)$ by $H(x, y) = \Sigma_i x_i \log(x_i/y_i)$ and the symmetric relative entropy or entropy production by

$$J(x, y) = H(x, y) + H(y, x) = \sum_i (x_i - y_i) \log \frac{x_i}{y_i}.$$

Some authors define the relative entropy with the opposite sign (e.g., Ahlswede and Gács 1976, Donald 1986) or with its arguments in the reverse order, so care is required in relating the results of different papers.

DEFINITION 2.2. Let ϕ be a continuous real-valued function on $(0, \infty) \times (0, \infty)$ that is homogeneous and jointly convex in its arguments and satisfies $\phi(1, 1) = 0$. For any two positive d-vectors $x = (x_i)$ and $y = (y_i)$, whether or not x and y are probability vectors, define the relative ϕ-entropy $H_\phi(x, y)$ by $H_\phi(x, y) = \Sigma_i \phi(x_i, y_i)$ and a symmetric relative ϕ-entropy by

$$J_\phi(x, y) = H_\phi(x, y) + H_\phi(y, x).$$

Since $\tilde{\phi}(a, b) \equiv \phi(a, b) + \phi(b, a)$ satisfies the hypotheses of Definition 2.2 if ϕ does, and since $J_\phi(x, y) = H_{\tilde{\phi}}(x, y)$, there is no further need to speak separately of J_ϕ.

It follows from homogeneity and convexity at $\frac{1}{2}$ that ϕ is subadditive, i.e.,

$$\phi(x_1 + x_2, y_1 + y_2) \leqslant \phi(x_1, y_1) + \phi(x_2, y_2) \qquad \text{for all} \quad x_1, x_2, y_1, y_2 \in (0, \infty).$$

It will occasionally be useful to define $\phi(0,0) = 0$ and correspondingly to extend the definition of $H_\phi(x, y)$ to suitable pairs of nonnegative vectors.

Henceforth H_ϕ will denote any relative ϕ-entropy where ϕ is assumed to satisfy the hypotheses of Definition 2.2. This generalization of relative entropy has been widely studied under various names and notations (e.g., Csiszár 1963, 1967, Ali and Silvey 1966, Abrahams 1982, Petz 1986a, b). We distinguish the logarithmic special case (Definition 2.1) because this case is usually considered in most applications.

It is easily proved that a continuous real-valued homogeneous function ϕ on $(0,\infty) \times (0,\infty)$ is jointly convex in both arguments if and only if $g(t) \equiv \phi(1, 1+t)$ is convex for $t \in (-1, \infty)$. It follows that any continuous real-valued convex function $g(t)$ on $(-1,\infty)$ such that $g(0) = 0$ defines a relative ϕ-entropy via the assumptions that $\phi(1, 1+t) = g(t)$ and ϕ is homogeneous. Hence the relative ϕ-entropy and related quantities can be indexed by ϕ or by g, i.e., $H_\phi(x, y) = \Sigma_i \phi(x_i, y_i)$ if and only if $H_g(x, y) = \Sigma_i x_i g(y_i/x_i - 1)$. However, in all cases $H_{\log}$ denotes the relative entropy in Definition 2.1; $H_{\log}$ is the special case of H_g when $g(t) = -\log(1+t)$.

The following properties of relative entropy are readily established, so we omit proofs. First, H_ϕ is a continuous, real-valued function that is homogeneous, jointly convex in (x, y) for any positive d-vectors x and y, subadditive, and such that $H_\phi(x, x) = 0$. Second, for any $x, y \in P_d$, $H_\phi(x, y) \geqslant 0$; and if $\phi(1, t)$ is strictly convex for $t \in (0, \infty)$, then $H_\phi(x, y) = 0$ if and only if $x = y$. Third, for any positive d-vectors x, y and any $d \times d$ permutation matrix Q, $H_\phi(x, y) = H_\phi(Qx, Qy)$. For any positive n-vectors x, y, any permutation matrices Q_1, Q_2 of size $m \times m$ and $n \times n$, respectively, and any row-allowable $m \times n$ matrix A, there exist positive n-vectors x', y' such that $H_\phi(Q_1 A Q_2 x, Q_1 A Q_2 y)/H_\phi(x, y) = H_\phi(Ax', Ay')/H_\phi(x', y')$.

It has been proved many times and in more general settings that if A is a column-stochastic, row-allowable $m \times n$ matrix and x and y are positive n-vectors (x and y need not be normalized to be probability n-vectors), then $H_\phi(Ax, Ay) \leqslant H_\phi(x, y)$ (e.g., Moran 1961, Csiszár 1963 [p. 90, his Theorem 1], Morimoto 1963). Further, the inequality is strict if $\phi(1, \cdot)$ is strictly convex and A is scrambling and $x \neq y$ (e.g., Ahlswede and Gács 1976).

Definition 2.3. For any $m \times n$ matrix A, Dobrushin's (1956) coefficient of ergodicity is

$$\alpha(A) = \min_{j,k} \sum_{i=1}^{m} \min(a_{ij}, a_{ik}).$$

A column-stochastic, row-allowable matrix A is scrambling if and only if $\alpha(A) > 0$ (see Iosifescu 1980, pp. 56–57). The complement $1 - \alpha(A)$ will be written

$$\bar{\alpha}(A) \equiv 1 - \alpha(A) = \frac{1}{2} \max_{j,k} \sum_{i=1}^{m} |a_{ij} - a_{ik}| \tag{2.3.1}$$

and satisfies (Dobrushin 1956, pp. 69–70)

$$\bar{\alpha}(A) = \sup\left\{ \frac{\| A(x-y)\|_1}{\|x-y\|_1} : x \text{ and } y \text{ are positive } n\text{-vectors}, \right.$$

$$\left. x \neq y, \|x\|_1 = \|y\|_1 \right\}. \tag{2.3.2}$$

3. RELATIVE ENTROPY UNDER THE ACTION OF A STOCHASTIC MATRIX

THEOREM 3.1. *Let A be a column-stochastic, row-allowable $m \times n$ matrix, and let $x, y \in P_n$. Then*

$$H_\phi(Ax, Ay) \leqslant \bar{\alpha}(A) H_\phi(x, y).$$

The proof of this theorem is the goal of the remainder of this section.

LEMMA 3.2. *Let A be a column-stochastic $m \times n$ matrix. For any real n-vector $s = (s_j)$,*

$$\sum_{i=1}^{m} |(As)_i| \leqslant \bar{\alpha}(A) \sum_{j=1}^{n} |s_j| + \alpha(A) \left| \sum_{j=1}^{n} s_j \right|.$$

Proof. Define $\alpha = \alpha(A)$, $\bar{\alpha} = 1 - \alpha$, and

$$J_+ = \{j : s_j \geqslant 0\}, \qquad J_- = \{j : s_j < 0\}, \qquad u_\pm = \sum_{J_\pm} |s_j|.$$

If $s \in N_n$ or $-s \in N_n$, the result is true. Hence assume $\min(u_-, u_+) > 0$. Then, since $\Sigma_{J_-} |s_k| / u_- = \Sigma_{J_+} s_j / u_+ = 1$ by definition,

$$\sum_i |(As)_i| = \sum_i \left| \sum_j a_{ij} s_j \right| = \sum_i \left| \sum_{j \in J_+} a_{ij} s_j - \sum_{k \in J_-} a_{ik} |s_k| \right|$$

$$= \sum_i \left| u_+ \sum_{j\in J_+} a_{ij}\frac{s_j}{u_+} - u_- \sum_{k\in J_-} a_{ik}\frac{|s_k|}{u_-} \right|$$

$$= \sum_i \left| u_+ \left(\sum_{J_-} \frac{|s_k|}{u_-} \right) \sum_{J_+} a_{ij}\frac{s_j}{u_+} - u_- \left(\sum_{J_+} \frac{s_j}{u_+} \right) \sum_{J_-} a_{ik}\frac{|s_k|}{u_-} \right|$$

$$\leqslant \sum_{j\in J_+} \sum_{k\in J_-} \frac{s_j}{u_+}\frac{|s_k|}{u_-} \sum_i |u_+ a_{ij} - u_- a_{ik}|$$

$$\leqslant \max_{j,k\,:\,j\neq k} \sum_i |u_+ a_{ij} - u_- a_{ik}|$$

$$= \max_{j\neq k} \sum_i \left| u_+(a_{ij} - a_{ik}) - (u_- - u_+)a_{ik} \right|.$$

Now define $u = \min\{u_+, u_-\}$. Then, by definition, $|\sum_j s_j| = |u_+ - u_-|$. It is readily seen that $2u = u_+ + u_- - |u_+ - u_-|$. Without loss of generality, assume that $u_+ \leqslant u_-$; if, on the contrary, $u_+ > u_-$, then exchange u_- and u_+ and exchange j and k. Continuing the previous equation, we have, since $\sum_i a_{ih} = 1$ for any h, e.g., $h = j$ or k,

$$\leqslant \max_{j\neq k} \left\{ u \sum_i |a_{ij} - a_{ik}| \right\} + |u_- - u_+|$$

$$= 2u\bar{\alpha} + |u_+ - u_-|$$

$$= (u_+ + u_-)\bar{\alpha} - |u_+ - u_-|\bar{\alpha} + |u_+ - u_-|$$

$$= \bar{\alpha}\sum_j |s_j| + \alpha\left|\sum_j s_j\right|. \qquad \blacksquare$$

LEMMA 3.3. *Let μ and ν be measures on the measure space $(\Re, \mathsf{B}(\Re))$ with bounded variation and compact support such that*

(i) *$\int g\,d\mu = \int g\,d\nu$ if $g(x) = ax + b$, and*

(ii) *$\int g\,d\mu \leqslant \int g\,d\nu$ if $g(x) = |x - c|$ for all $c \in \sigma = supp(\mu) \cup supp(\nu)$.*

Then $\int g\,d\mu \leqslant \int g\,d\nu$ for any convex, continuous function $g : \sigma \to \Re$.

Proof. (i) and (ii) imply $\int g\,d\mu \leqslant \int g\,d\nu$ if $g(x) = (x - c)_+ = \frac{1}{2}[(x - c) + |x - c|]$. Any convex continuous $g : \sigma \to \Re$ can be approximated uniformly (Rockafellar 1970, p. 91) by a sequence $\{g_N\}$ of functions, $N \to \infty$, defined by

$$g_N(x) = g(x_0) + m_0(x - x_0) + \sum_{i=1}^{N} (m_i - m_{i-1})(x - x_i)_+$$

where $\{x_0, \ldots, x_N\}$ is a partition of σ and

$$m_i = \frac{g(x_{i+1}) - g(x_i)}{x_{i+1} - x_i}.$$

Thus the terms of g_N satisfy the desired inequality; hence

$$\int g_N \, d\mu \leqslant \int g_N \, d\nu,$$

and letting $N \to \infty$ gives, in view of the uniform convergence of g_N,

$$\int g \, d\mu \leqslant \int g \, d\nu. \qquad \blacksquare$$

LEMMA 3.4. *Let $x, y \in P_n$, and let $g : [-1, \infty) \to \mathfrak{R}$ be convex and continuous. Then for any column-stochastic, row-allowable $m \times n$ matrix A, with $\alpha = \alpha(A)$, $\bar{\alpha} = \bar{\alpha}(A)$,*

$$\sum_i (Ax)_i g\left(\frac{(Ay)_i - (Ax)_i}{(Ax)_i}\right) \leqslant \bar{\alpha} \sum_k x_k g\left(\frac{y_k - x_k}{x_k}\right) + \alpha g(0).$$

Proof. Let δ denote the Dirac needle function, and define

$$\mu = \sum_i (Ax)_i \delta\left(\frac{(Ay)_i - (Ax)_i}{(Ax)_i}\right),$$

$$\nu = \bar{\alpha} \sum_k x_k \delta\left(\frac{y_k - x_k}{x_k}\right) + \alpha \delta(0).$$

Then for any function g,

$$\int g \, d\mu = \sum_i (Ax)_i g\left(\frac{(Ay)_i - (Ax)_i}{(Ax)_i}\right),$$

$$\int g \, d\nu = \bar{\alpha} \sum_k x_k g\left(\frac{y_k - x_k}{x_k}\right) + \alpha g(0).$$

We now verify that hypotheses (i) and (ii) of Lemma 3.3 are satisfied. As for (i), if $g(x) = ax + b$, then $g(0) = b$ and

$$\int g\,d\mu = \sum_i \{a[(Ay)_i - (Ax)_i] + b(Ax)_i\} = b,$$

$$\int g\,d\nu = \bar{\alpha}\sum_k [a(y_k - x_k) + bx_k] + \alpha b$$

$$= \bar{\alpha}b + \alpha b = b.$$

This proves that μ and ν satisfy (i) of Lemma 3.3.

As for (ii), if $g = |x - c|$, where $c \geqslant -1$, then

$$\int g\,d\mu = \sum_i |(Ay)_i - (A(c+1)x)_i|$$

$$\int g\,d\nu = \bar{\alpha}\sum_k |y_k - x_k - cx_k| + \alpha|-c|$$

$$= \bar{\alpha}\sum_k |y_k - (c+1)x_k| + \alpha|c|.$$

Now apply Lemma 3.2 with $s_k = y_k - (c+1)x_k$. Then, since $\Sigma_k s_k = -c$ if $x, y \in P_n$,

$$\int g\,d\mu = \sum_i |(As)_i| \leqslant \bar{\alpha}\sum_k |s_k| + \alpha\left|\sum_k s_k\right| = \int g\,d\nu.$$

This proves that μ and ν satisfy (ii) of Lemma 3.3. Hence

$$\int g\,d\mu \leqslant \int g\,d\nu. \qquad \blacksquare$$

Proof of Theorem 3.1.

$$H_\phi(x, y) = \sum_k x_k \phi\left(1, \frac{y_k}{x_k}\right) = \sum_k x_k \phi\left(1, 1 + \frac{y_k - x_k}{x_k}\right) = \sum_k x_k g\left(\frac{y_k - x_k}{x_k}\right),$$

where $g : [-1, \infty) \to \Re$ defined by $g(t) = \phi(1, 1 + t)$ is convex and $g(0) = \phi(1, 1) = 0$. The desired conclusion follows immediately from Lemma 3.4. ■

4. THE RELATIVE-ENTROPY CONTRACTION COEFFICIENT

Assume in this section that $\phi(1, \cdot)$ is strictly convex on $(0, \infty)$. Hence, for any $x \in P_n$, $y \in P_n$, whenever $x \neq y$, $H_\phi(x, y) > 0$. Also assume throughout this section that A is a column-stochastic, row-allowable $m \times n$ matrix. For such a matrix A, define the relative ϕ-entropy contraction coefficient

$$\eta_\phi(A) = \sup\left\{ \frac{H_\phi(Ax, Ay)}{H_\phi(x, y)} : x \in P_n, \ y \in P_n, \ x \neq y \right\}.$$

The requirement $x, y \in P_n$ can be replaced by the requirements that x, y are positive and $\|x\|_1 = \|y\|_1$.

The special case when $\phi(a, b) = a \log(a/b)$ was extensively investigated by Ahlswede and Gács (1976).

The following properties of the relative ϕ-entropy contraction coefficient are elementary to prove, so we omit the details:

(i) If A and B are column-stochastic, row-allowable matrices of size $m \times n$ and $n \times q$, respectively, then $\eta_\phi(AB) \leqslant \eta_\phi(A)\eta_\phi(B)$.

(ii) For any permutation matrix P, $\eta_\phi(P) = 1$.

(iii) $\eta_\phi(A)$ is invariant under arbitrary independent permutations of the rows and columns of A. (However, see Remark 4.3.)

(iv) $\eta_\phi(A)$ is convex in A, i.e., if A and B are column-stochastic, row-allowable $m \times n$ matrices and $0 \leqslant p \leqslant 1$, then $\eta_\phi(pA + (1-p)B) \leqslant p\eta_\phi(A) + (1-p)\eta_\phi(B)$.

(v) If A is a positive column-stochastic matrix, then $\eta_\phi(A) = 0$ if and only if A has rank 1.

The major result of this section, which follows, is an immediate consequence of Theorem 3.1.

THEOREM 4.1. $0 \leqslant \eta_\phi(A) \leqslant \bar{\alpha}(A) \leqslant 1$.

THEOREM 4.2. $\eta_\phi(A) < 1$ *if and only if A is scrambling if and only if* $\bar{\alpha}(A) < 1$.

Proof. If A is scrambling, then $\bar{\alpha}(A) < 1$, so $\eta_\phi(A) < 1$ by Theorem 4.1.

If A is not scrambling, then there exist indices j and k such that for all indices i, $a_{ij}a_{ik} = 0$. Thus if $J = \{i : a_{ij} > 0\}$ and $K = \{i : a_{ik} > 0\}$, then $J \cap K = \varnothing$ and $\Sigma_{i \in J} a_{ij} = 1 = \Sigma_{i \in K} a_{ik}$. Choose s, t such that $0 < s, t < 1$, and define $x \in N_n$, $y \in N_n$ by $x_j = s$, $x_k = 1 - s$, $y_j = t$, $y_k = 1 - t$; let all other elements of x and y be 0. Then $x \neq y$ if and only if $s \neq t$, and

$\|x\|_1 = \|y\|_1 = 1$. Then

$$H_\phi(x, y) = \phi(s,t) + \phi(1-s, 1-t) \neq 0 \qquad \text{if} \quad s \neq t.$$

Now $(Ax)_i = a_{ij}s$ and $(Ay)_i = a_{ij}t$ for $i \in J$; $(Ax)_i = a_{ik}(1-s)$ and $(Ay)_i = a_{ik}(1-t)$ for $i \in K$; and $(Ax)_i = 0$, $(Ay)_i = 0$ otherwise. Thus

$$\begin{aligned} H_\phi(Ax, Ay) &= \sum_{i\in J} \phi(a_{ij}s, a_{ij}t) + \sum_{i\in K} \phi(a_{ik}(1-s), a_{ik}(1-t)) \\ &= \sum_{i\in J} a_{ij}\phi(s,t) + \sum_{i\in K} a_{ik}\phi(1-s, 1-t) \\ &= \phi(s,t) + \phi(1-s,1-t) = H_\phi(x,y). \end{aligned}$$

By replacing 0 with $\varepsilon > 0$ and taking the limit as $\varepsilon \to 0$, one can find a set of strictly positive vectors $\{x_\varepsilon, y_\varepsilon\}$ such that

$$\lim_{\varepsilon\to 0} \frac{H_\phi(Ax_\varepsilon, Ay_\varepsilon)}{H_\phi(x_\varepsilon, y_\varepsilon)} = 1;$$

hence $\eta_\phi(A) = 1$. ■

REMARK 4.3. It is not true that if a column-stochastic, row-allowable $d \times d$ matrix A is primitive, then $\eta_\phi(A) < 1$, because many primitive matrices are not scrambling.

EXAMPLE. Let

$$A = \begin{pmatrix} 0 & s & s' \\ 1 & 0 & 0 \\ 0 & 1-s & 1-s' \end{pmatrix}, \qquad x = \begin{pmatrix} 1-(a+b) \\ a \\ b \end{pmatrix},$$

$$y = \begin{pmatrix} 1-\lambda(a+b) \\ \lambda a \\ \lambda b \end{pmatrix},$$

where λ is any positive real number and $s, s' \in (0,1)$. Since A^3 is positive, A is primitive. Elementary calculation shows that $H_\phi(Ax, Ay) = H_\phi(x, y)$, regardless of λ; hence $\eta_\phi(A) = 1$. This example generalizes an example kindly given us by Gerald S. Goodman (personal communication, 8 June 1989).

Since $\eta_\phi(A^2) < 1$, we have $\eta_\phi(A^2) < [\eta_\phi(A)]^2$ in this case. This example also shows that, in general, $\eta_\phi(A^2)$ is not invariant under permutations of the rows of A alone. For by exchanging the first and second rows of A in the example, one obtains a block-diagonal, nonscrambling matrix (call it B) for which $\eta_\phi(B) = 1$ and $\eta_\phi(B^2) = 1$.

PROBLEM 4.4. Let ϕ satisfy the assumptions of Definition 2.2, and assume that $\phi(1,\cdot)$ is strictly convex on $(0,\infty)$. Suppose ϕ' is another function with the same properties. Does $\eta_\phi(A) = \eta_{\phi'}(A)$ for all column-stochastic, row-allowable A imply $\phi = c\phi'$ for some $c > 0$?

5. BOUNDS RELATING DIFFERENT COEFFICIENTS OF CONTRACTION AND ERGODICITY

Assume throughout this section that A is a column-stochastic, row-allowable $m \times n$ matrix.

DEFINITION 5.1. Doeblin's coefficient of ergodicity δ is

$$\delta(A) = \sum_{i=1}^{m} \min\{a_{ij} : j = 1, \dots, n\}.$$

It is known that $\delta(A) \leqslant \alpha(A)$ with equality if $n = 2$.

Define $\bar{\delta}(A) = 1 - \delta(A)$.

DEFINITION 5.2. If x and y are positive d-vectors, the Hilbert projective pseudometric d is (e.g., Seneta 1981, pp. 80–81)

$$d(x, y) = \log \frac{\max_i(x_i/y_i)}{\min_j(x_j/y_j)} = \max_{i,j} \log \frac{x_i y_j}{x_j y_i}.$$

DEFINITION 5.3. Birkhoff's contraction coefficient is

$$\tau_B(A) = \sup\left\{ \frac{d(Ax, Ay)}{d(x, y)} : x \in P_d,\ y \in P_d,\ x \neq y \right\},$$

where $d(x, y)$ is the Hilbert projective pseudometric just defined.

It is easy to prove that for all positive probability d-vectors x, y, one has $0 \leqslant J(x, y) \leqslant d(x, y)$, and $0 = J(x, y)$ if and only if $J(x, y) = d(x, y)$ if and only if $x = y$. It is already known, moreover, that for any column-stochastic $d \times d$ matrix A, one has $\bar{\alpha}(A) \leqslant \tau_B(A)$ (Bauer, Deutsch, and Stoer 1969; see e.g. Seneta 1981, p. 110).

For $g(t) = \phi(1, 1 + t)$, the relative entropy, and the corresponding contraction coefficients, can be indexed by ϕ or by g. It will now be convenient to write H_g rather than H_ϕ. We will simply write $\eta_{\log}(A)$ instead of $\eta_{-\log(1+t)}(A)$. It follows from Theorem 4.1 that for all A and for all strictly convex functions g, one has $\eta_g(A) \leqslant \eta_{|w|}(A) = \bar{\alpha}(A)$, $\eta_g(A) \leqslant \bar{\delta}(A)$, and $\eta_g(A) \leqslant \tau_B(A)$. Because there exist matrices A such that $\bar{\alpha}(A) < \bar{\delta}(A)$ and $\bar{\alpha}(A) < \tau_B(A)$ (e.g., Example 5.7 and Example 5.8), there is no strictly convex function g for which either $\bar{\delta}(A) = \eta_g(A)$ or $\tau_B(A) = \eta_g(A)$, i.e., neither Doeblin's nor Birkhoff's coefficient of ergodicity can be realized as the contraction coefficient of some relative ϕ-entropy.

We now study $\eta_g(A)$ for $g(w) = w^2$ and $g(w) = -\log(1 + w)$. Since g is always assumed to be convex, the point of Theorem 5.4 is the strict positivity of $g''(0)$.

THEOREM 5.4. *If $g(w)$ is thrice differentiable in a neighborhood of 0 and $g''(0) > 0$, then $\eta_{w^2}(A) \leqslant \eta_g(A)$; in particular, $\eta_{w^2}(A) \leqslant \eta_{\log}(A)$.*

Proof. Whenever g is homogeneous of even degree, it can be extended to all of $\Re$. In particular, because $g(w) = w^2$ is homogeneous of degree two,

$$\eta_{w^2}(A) = \sup_{\substack{x \neq y \\ x, y \in P_n}} \frac{H_{w^2}(Ax, Ay)}{H_{w^2}(x, y)}$$

$$= \sup_{\substack{x \in P_n \\ v^T 1 = 0}} \frac{H_{w^2}(Ax, Ax + Av)}{H_{w^2}(x, x + v)} \equiv \sup_{\substack{x \in P_n \\ v^T 1 = 0}} \frac{\Phi(Ax, Av)}{\Phi(x, v)}, \quad (5.4.1)$$

where $\Phi(x, v) = H_{w^2}(x, x + v) = \Sigma_j v_j^2 / x_j$ and $v \neq 0$ is arbitrary except that $v^T 1 = 0$; henceforth, when we write $v^T 1 = 0$, we always assume $v \neq 0$.

By the hypotheses, we can expand g in a Taylor's series about $w = 0$, and since $g(0) = 0$,

$$\phi(s, t) = sg\left(\frac{t}{s} - 1\right) = (t - s) g'(0) + \frac{(t - s)^2}{s} \frac{g''(0)}{2} + \frac{1}{s^2} O\left((t - s)^3\right).$$

Now let $x \in P_n$, and let $v \neq 0$ satisfy $v^T 1 = 0$. For sufficiently small ε, $y_\varepsilon = x + \varepsilon v \in P_n$. Then $H_g(x, y_\varepsilon)$ is well defined and

$$H_g(x, y_\varepsilon) = \frac{g''(0)}{2}\varepsilon^2 \Phi(x, v) + O(\varepsilon^3). \tag{5.4.2}$$

Since A is linear, $H_g(Ax, Ay_\varepsilon) = [g''(0)/2]\varepsilon^2\Phi(Ax, Av) + O(\varepsilon^2)$. Therefore

$$\eta_g(A) \geqslant \frac{H_g(Ax, Ay_\varepsilon)}{H_g(x, y_\varepsilon)} = \frac{\Phi(Ax, Av)}{\Phi(x, v)} + O(\varepsilon). \tag{5.4.3}$$

It follows from (5.4.1) that, by choosing x, v, ε appropriately, the right side of (5.4.3) can be made arbitrarily close to $\eta_{w^2}(A)$. ■

THEOREM 5.5.

$$\eta_{w^2}(A) \geqslant \zeta(A) \equiv \frac{1}{2}\max_{j,k:j\neq k}\sum_{i=1}^{n}\frac{(a_{ij} - a_{ik})^2}{a_{ij} + a_{ik}}.$$

Proof. Let k, l be the indices of the columns of A for which the maximum above is attained. Choose x, v so that $x_k = x_l = \frac{1}{2}$ and $x_j = \varepsilon$ for $j \neq k$, $j \neq l$; $v_k = +1$, $v_l = -1$, $v_j = 0$ for $j \neq k$, $j \neq l$. Then it can readily be verified that

$$\lim_{\varepsilon\to 0}\frac{\Phi(Ax, Av)}{\Phi(x, v)} = \frac{1}{2}\sum_{i=1}^{m}\frac{(a_{ik} - a_{il})^2}{a_{ik} + a_{il}} = \zeta(A). \qquad ■$$

If A is not scrambling, then there exist indices j and k such that for all indices i, $|a_{ij} - a_{ik}| = a_{ij} + a_{ik} = \max\{a_{ij}, a_{ik}\}$, so that $\zeta(A) = 1$. Therefore

$$\zeta(A) = \eta_{w^2}(A) = \eta_{\log}(A) = \bar{\alpha}(A) = 1.$$

THEOREM 5.6. *Let A be a $d \times d$ primitive column-stochastic matrix. Let $\lambda_2(A)$ be an eigenvalue of A second largest in modulus. Then* $\eta_{w^2}(A) \geqslant [\eta_{|w|}(A)]^2 = [\bar{\alpha}(A)]^2 \geqslant |\lambda_2(A)|^2$.

Proof. Schwarz's inequality implies that, for all $x \in P_n$, $\Phi(x, v) \geqslant [\Sigma_k |v_k|]^2 = \|v\|_1^2$, so that

$$\eta_{w^2}(A) = \sup_{\substack{x\in P_n\\ v^T 1 = 0}}\frac{\Phi(Ax, Av)}{\Phi(x, v)} \geqslant \sup_{\substack{x\in P_n\\ v^T 1 = 0}}\frac{\|Av\|_1^2}{\Phi(x, v)}.$$

Now choose $x_k = |v_k|/\|v\|_1$; then $x \in P_d$ and $\Phi(x, v) = \|v\|_1^2$. Thus

$$\eta_{w^2}(A) \geqslant \sup_{v^T 1 = 0} \frac{\|Av\|_1^2}{\|v\|_1^2} = \left[\sup_{v^T 1 = 0} \frac{\|Av\|_1}{\|v\|_1} \right]^2 = \left[\eta_{|w|}(A)\right]^2.$$

The final inequality of the theorem is already known (e.g., Seneta 1979). ■

Example 5.7. Let A be the matrix with elements $a_{ij} = (1 - \delta_{ij})/(d - 1)$, i.e.,

$$a_{ij} = \begin{cases} \dfrac{1}{d-1} & \text{if} \quad i \neq j, \\ 0 & \text{if} \quad i = j. \end{cases}$$

The following remarks are readily verified:

(i) $\zeta(A) = \eta_{w^2}(A) = \eta_{\log}(A) = \bar{\alpha}(A) = 1/(d - 1)$.

(ii) $\delta(A) = 0$ and $\tau_B(A) = 1$. Therefore, if $d > 2$ then $\bar{\alpha}(A) < \bar{\delta}(A) = \tau_B(A)$.

(iii) $Ax = (1 - x)/(d - 1)$ for all $x \in P_d$. Combining this with (i), Theorem 3.1, and the homogeneity of H_ϕ yields

$$H_\phi(Ax, Ay) = \frac{1}{d-1} H_\phi((1 - x), (1 - y)) \leqslant \frac{1}{d-1} H_\phi(x, y).$$

Equivalently, if $x, y \in P_d$, then $H_\phi((1 - x), (1 - y)) \leqslant H_\phi(x, y)$. Similarly, $\Phi(1 - x, v) \leqslant \Phi(x, v)$.

(iv) Since both $\eta_{\log}(\cdot)$ and $\delta(\cdot)$ are continuous functions of the elements of the matrix argument, A can be perturbed slightly to a matrix A' with all elements positive while retaining $\eta_{\log}(A') < 1 - \delta(A')$.

Example 5.8. Combining Theorem 5.4 and Theorem 5.5 with Theorem 4.1 and the remark following Definition 5.3 gives

$$\zeta(A) \leqslant \eta_{w^2}(A) \leqslant \eta_{\log}(A) \leqslant \bar{\alpha}(A) \leqslant \tau_B(A).$$

Various combinations of equality and strict inequality can hold; to demonstrate this we summarize in Table 1 the simple example

$$A = \begin{pmatrix} p & 1 - q \\ 1 - p & q \end{pmatrix}.$$

TABLE 1

CONTRACTION AND ERGODICITY COEFFICIENTS IN 2 × 2 MATRICES

q	$\zeta(A)$	$\eta_{w^2}(A) = \eta_{\log}(A)$	$\bar{\alpha}(A) = \bar{\delta}(A)$	$\tau_B(A)$
1	$\frac{p}{2-p}$	p	p	p
0	$\frac{1-p}{1+p}$	$1-p$	$1-p$	$1-p$
p	$(1-2p)^2$	$(1-2p)^2$	$\|1-2p\|$	$\|1-2p\|$
Arbitrary	$\frac{[1-(p+q)]^2}{1-(p-q)^2}$	$\frac{[1-(p+q)]^2}{\left[\sqrt{pq}+\sqrt{(1-p)(1-q)}\right]^2}$	$\|1-(p+q)\|$	$\frac{1-(p+q)}{\left[\sqrt{pq}+\sqrt{(1-p)(1-q)}\right]^2}$
$1-p$	0	0	0	0

The expressions for arbitrary q and the proof that $\eta_{w^2}(A) = \eta_{\log}(A)$ for all 2×2 matrices require elementary, but tedious, computations. We omit the computations, which can be verified by using a symbolic manipulation computer program. We distinguish four situations:

(a) $\zeta(A) = \eta_{w^2}(A) = \eta_{\log}(A) = \bar{\alpha}(A) = \tau_B(A)$. This occurs if $p = q = 1$ or $p = q = 0$ [in which cases A is a permutation matrix and $\eta_g(A) = 1$] or $p = 1 - q$ [in which case A is a projection matrix of rank 1 and $\eta_g(A) = 0$].

(b) $\zeta(A) < \eta_{w^2}(A) = \eta_{\log}(A) = \bar{\alpha}(A) = \tau_B(A)$. This occurs if $q \neq p, 1 - p$ and $q = 1, 0$.

(c) $\zeta(A) = \eta_{w^2}(A) = \eta_{\log}(A) < \bar{\alpha}(A) = \tau_B(A)$. This occurs if $q = p$ but $q \neq 0, \frac{1}{2}, 1$.

(d) $\zeta(A) < \eta_{w^2}(A) = \eta_{\log}(A) < \bar{\alpha}(A) < \tau_B(A)$. This occurs if $q \neq 0, p, 1 - p, 1$.

These examples illustrate that any combination of equality and strict inequality can occur in $\zeta(A) \leqslant \eta_{w^2}(A) \leqslant \bar{\alpha}(A)$. That either equality or strict inequality can hold in $\bar{\alpha}(A) \leqslant \tau_B(A)$ and $\bar{\alpha}(A) \leqslant \bar{\delta}(A)$ is well known and is illustrated by Example 5.8(d) and Example 5.7(ii), although the latter requires $d > 2$. It is an open question whether $\eta_{w^2}(A) < \eta_{\log}(A)$ is possible in higher dimensions.

6. ASYMPTOTIC BEHAVIOR OF RELATIVE ENTROPY

As usual, a $d \times d$ matrix A is said to be simple if it is similar to a diagonal matrix, i.e., if there exist $d \times d$ matrices C and Λ, where Λ is diagonal, $\Lambda = \operatorname{diag}(\lambda_i)$, such that $A = C\Lambda C^{-1}$. If c_i denotes the ith column of C, then $Ac_i = \lambda_i c_i$, so the column c_i is a right eigenvector of A corresponding to the eigenvalue λ_i. If $V = C^{-1}$, and the ith row of V is written v_i^T, then $v_i^T A = \lambda_i v_i^T$. Thus v_i^T is a left eigenvector of A correspoding to λ_i. The usual spectral decomposition or spectral resolution of a simple matrix A is $A = \sum_i \lambda_i c_i v_i^T$. It follows that $A^k = \sum_i \lambda_i^k c_i v_i^T$ for any positive integer k. Assume the eigenvalues of a simple matrix A are labeled so that $|\lambda_1| \geqslant |\lambda_2| \geqslant \cdots$.

THEOREM 6.1. *Let A be a primitive, simple column-stochastic $d \times d$ matrix with spectral decomposition $A = \sum_i \lambda_i c_i v_i^T$. Fix probability vectors $x \in N_d$, $y \in N_d$, $x \neq y$. Let i_0 be the least positive integer i for which $v_i^T(x - y) \neq 0$. Assume A is such that λ_{i_0} is real and $|\lambda_{i_0}| > |\lambda_{i_0 + 1}|$. Assume that $g(w) \equiv \phi(1, 1 + w)$ is thrice differentiable in a neighborhood of 0 and* *that $g''(0) > 0$. Then*

$$\lim_{t\to\infty} \frac{H_\phi(A^{t+1}x,\, A^{t+1}y)}{H_\phi(A^t x,\, A^t y)} = \left(\lambda_{i_0}\right)^2.$$

In words, the asymptotic rate of contraction of the relative ϕ-entropy is the square of the largest real eigenvalue for which the left eigenvector is not orthogonal to $x - y$, when such an eigenvalue exists and exceeds in modulus the remaining eigenvalues.

Before giving the proof, we observe that, since v_1^T is a positive constant vector and the sums of both x and y are 1, we have $i_0 \geqslant 2$.

Proof. Let $x(t) = A^t x$, $y(t) = A^t y$. By the Perron-Frobenius theorem for primitive matrices (e.g., Seneta 1981) applied to a column-stochastic matrix, $\lambda_1 = 1$, c_1 is a positive vector, which we name $\pi = (\pi_i)$, and v_1 is a positive constant vector, which we may take to be $\mathbf{1}^T$. Thus

$$x(t) = \pi + \sum_{i=2}^{d} \lambda_i^t c_i v_i^T x = \pi + \sum_{i=2}^{d} \lambda_i^t c_i \alpha_i \qquad \left(\alpha_i \equiv v_i^T x\right),$$

$$y(t) = \pi + \sum_{i=2}^{d} \lambda_i^t c_i v_i^T y = \pi + \sum_{i=2}^{d} \lambda_i^t c_i \beta_i \qquad \left(\beta_i \equiv v_i^T y\right).$$

If $c_i(j)$ is the jth element of c_i, then

$$\varepsilon_j(t) \equiv y_j(t) - x_j(t) = \sum_{i=2}^{d} \lambda_i^t(\beta_i - \alpha_i)c_i(j) = \sum_{i=i_0}^{d} \lambda_i^t(\beta_i - \alpha_i)c_i(j).$$

In the summation on the right, all terms (if any) with $2 \leqslant i < i_0$ vanish, because for them $\alpha_i - \beta_i = v_i^T(x - y) = 0$. The Perron-Frobenius theorem guarantees that $1 > |\lambda_2|$. So for large t, $\varepsilon_j(t) \to 0$ and therefore $|\varepsilon_j(t)/x_j(t)| \leqslant 1$ and $\varepsilon_j(t)/x_j(t) \neq -1$. Using a Taylor-series argument similar to that at (5.4.2), it follows that for large t,

$$H_\phi(x(t), y(t)) = -\sum_j x_j(t)\left(g'(0)\frac{\varepsilon_j(t)}{x_j(t)} - \frac{g''(0)}{2}\left[\frac{\varepsilon_j(t)}{x_j(t)}\right]^2\right) + O\left(\max_i |\varepsilon_i|^3\right).$$

Now $\Sigma_j \varepsilon_j(t) = 0$, so, neglecting all but the first nonzero term,

$$H_\phi(x(t), y(t)) = \frac{g''(0)}{2} \sum_j \frac{[\varepsilon_j(t)]^2}{x_j(t)} \sim \frac{g''(0)}{2} \sum_j \frac{\left[\Sigma_{i=i_0}^d \lambda_i^t(\beta_i - \alpha_i)c_i(j)\right]^2}{2\pi_j}$$

where $a(t) \sim b(t)$ means $\lim_{t\to\infty} a(t)/b(t) = 1$. Thus

$$\lim_{t\to\infty} \frac{H_\phi(x(t+1), y(t+1))}{H_\phi(x(t), y(t))}$$

$$= \lim_{t\to\infty} \frac{\lambda_{i_0}^{2t+2} \sum_j \frac{1}{2\pi_j}\left(\sum_{i=i_0}^{d} \lambda_i^{t+1}\lambda_{i_0}^{-(t+1)}(\beta_i - \alpha_i)c_i(j)\right)^2}{\lambda_{i_0}^{2t} \sum_j \frac{1}{2\pi_j}\left(\sum_{i=i_0}^{d} \lambda_i^{t}\lambda_{i_0}^{-t}(\beta_i - \alpha_i)c_i(j)\right)^2} = \lambda_{i_0}^2. \quad ■$$

The assumption that g is smooth cannot be eliminated from Theorem 6.1; e.g., for the nonsmooth function $\phi(s,t) = |s - t|$, an analogous argument shows that the asymptotic rate of contraction is $|\lambda_{i_0}|$ rather than $(\lambda_{i_0})^2$.

A column-stochastic matrix A is defined to be reversible if there exist d positive numbers π_i, $i = 1, 2, \dots, d$, such that $a_{ij}\pi_j = a_{ji}\pi_i$ for all $i, j = 1, 2, \dots, d$.

COROLLARY 6.2. *Let A be a primitive, reversible column-stochastic $d \times d$ matrix. Then A is simple and all the eigenvalues $\lambda_1 = 1, \lambda_2, \dots$ of A are necessarily real. Assume the second and third eigenvalues have different moduli, i.e., $|\lambda_2| > |\lambda_3|$. Let $x \in N_d$, $y \in N_d$ be such that $x \neq y$ and $v_2^T(x - y) \neq 0$, where $v_2^T A = \lambda_2 v_2^T$. Assume that $g(w) \equiv \phi(1, 1 + w)$ is thrice differentiable in a neighborhood of 0 and that $g''(0) > 0$. Then*

$$\lim_{t\to\infty} \frac{H_\phi(A^{t+1}x, A^{t+1}y)}{H_\phi(A^t x, A^t y)} = \lambda_2^2.$$

Proof. Because A is primitive, there is a unique $v \in P_d$ such that $Av = v$, by the Perron-Frobenius theorem. But if $a_{ij}\pi_j = a_{ji}\pi_i$ for all i, j, then $\Sigma_j a_{ij}\pi_j = \Sigma_j a_{ji}\pi_i = \pi_i$, i.e., $A\pi = \pi$. Therefore $\pi = v$. Because A is reversible, A is similar to a symmetric matrix (e.g., Cohen et al. 1982, proof of Lemma 1) and therefore has all real eigenvalues. The remaining assumptions of the corollary now meet the conditions of the previous theorem with $i_0 = 2$. ■

Significant classes of column-stochastic matrices do not satisfy the assumptions of Theorem 6.1. For example, if A has rank one, say $A = v\mathbf{1}^T$, where $v \in P_d$, then for any $x \in N_d$, $y \in N_d$, we have $Ax = v = Ay$. Therefore $H_\phi(Ax, Ay) = 0$, and the ratio $H_\phi(A^{t+1}x, A^{t+1}y)/H_\phi(A^t x, A^t y)$ is undefined for $t > 0$. As another example, suppose A is a 3×3 column-stochastic matrix with two conjugate complex eigenvalues in addition to the Perron-Frobenius root 1. Analysis along the lines of the previous proof suggests, and numerical computation verifies, that the ratio $H_\phi(A^{t+1}x, A^{t+1}y)/H_\phi(A^t x, A^t y)$ can be a nonmonotonic function of t.

7. CONTINUOUS-TIME MARKOV CHAINS AND RELATIVE ENTROPY

Results analogous to those for discrete-time Markov chains hold also for continuous-time Markov chains. For background, see Alberti and Uhlmann (1982, pp. 35–36).

Now assume all matrices are $d \times d$ and real. A matrix in which all off-diagonal elements are nonnegative and the sum of every column is zero is called an intensity matrix; such matrices have zero or negative elements on the main diagonal. If B is an intensity matrix, it is well known that for all nonnegative real t, e^{Bt} is column-stochastic. An intensity matrix B is the infinitesimal generator of a continuous-time Markov chain with transition probabilities from j to i given by the (i, j) element of e^{Bt}. For $x(0) \in P_d$, $y(0) \in P_d$, and $t \geqslant 0$, define $x(t) = e^{Bt}x(0)$ and $y(t) = e^{Bt}y(0)$.

THEOREM 7.1. *If B is an intensity matrix, $\omega \geqslant \max_i |b_{ii}|$, and*

$$\eta \equiv \eta_\phi(\omega^{-1}B + I),$$

then

$$\frac{d}{dt}\log H_\phi(x(t), y(t)) \leqslant \omega(\eta - 1).$$

Proof. Because $(B + \omega I)_{ij} \geqslant 0$ for all i, j, $(B + \omega I)/\omega$ is column-stochastic; hence η is meaningful. Let $\tilde{B} = B + \omega I$ and $\tilde{x}(t) = e^{\tilde{B}t}x_0 = e^{\omega t}e^{Bt}x_0 = e^{\omega t}x(t)$. Then

$$H_\phi[\tilde{x}(t), \tilde{y}(t)] = e^{\omega t}H_\phi[x(t), y(t)],$$

so

$$\frac{d}{dt}H_\phi[\tilde{x}(t), \tilde{y}(t)] = \omega e^{\omega t}H_\phi[x(t), y(t)] + e^{\omega t}\frac{d}{dt}H_\phi[x(t), y(t)]. \tag{7.1.1}$$

Now for very small real ε, $e^{\tilde{B}(t+\varepsilon)} = e^{\tilde{B}t}e^{\varepsilon\tilde{B}} \approx (I + \varepsilon\tilde{B})e^{\tilde{B}t}$; hence

$$\begin{aligned}\frac{d}{dt}H_\phi\left[e^{\tilde{B}t}x_0, e^{\tilde{B}t}y_0\right] &= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\Big(H_\phi\left[e^{\tilde{B}(t+\varepsilon)}x_0, e^{\tilde{B}(t+\varepsilon)}y_0\right]\\ &\qquad -H_\phi\left[e^{\tilde{B}t}x_0, e^{\tilde{B}t}y_0\right]\Big)\\ &= \lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\Big\{H_\phi\left[(I + \varepsilon\tilde{B})e^{\tilde{B}t}x_0, (I + \varepsilon B)e^{\tilde{B}t}y_0\right]\\ &\qquad -H_\phi\left[e^{\tilde{B}t}x_0, e^{\tilde{B}t}y_0\right]\Big\}. \end{aligned}\tag{7.1.2}$$

Now it is well known (e.g., Rockafellar 1970, p. 235) that if $f : D \to \Re$ is a convex and homogeneous (of degree one) function on any convex cone D of a real vector space, then $\overline{\lim}_{\varepsilon\to 0}[f(x + \varepsilon b) - f(x)]/\varepsilon \leqslant f(b)$. We apply this fact to the last expression with f replaced by H_ϕ, x replaced by $(\tilde{x}(t), \tilde{y}(t)) = (e^{\tilde{B}t}x_0, e^{\tilde{B}t}y_0)$, and b replaced by $(\tilde{B}\tilde{x}(t), \tilde{B}\tilde{y}(t))$. Then

$$\begin{aligned}\frac{d}{dt}H_\phi\left[e^{\tilde{B}t}x_0, e^{\tilde{B}t}y_0\right] &\leqslant H_\phi\left[\tilde{B}\tilde{x}(t), \tilde{B}\tilde{y}(t)\right]\\ &= e^{\omega t}H_\phi[(B + \omega I)x(t), (B + \omega I)y(t)]. \end{aligned}\tag{7.1.3}$$

Now combining (7.1.1) and (7.1.3) and multiplying by $e^{-\omega t'}$ gives

$$\begin{aligned}&\frac{d}{dt}H_\phi[x(t), y(t)] + \omega H_\phi[x(t), y(t)]\\ &\quad\leqslant H_\phi[(B + \omega I)x(t), (B + \omega I)y(t)]\\ &\quad= \omega H_\phi\left[(\omega^{-1}B + I)x(t), (\omega^{-1}B + I)y(t)\right]\\ &\quad\leqslant \omega\eta_\phi(\omega^{-1}B + I)H_\phi[x(t), y(t)].\end{aligned}$$

Hence

$$\frac{d}{dt}H_\phi[x(t), y(t)] \leqslant \omega(\eta - 1)H_\phi[x(t), y(t)]. \quad ■ \tag{7.1.4}$$

COROLLARY 7.2. *If B is an intensity matrix, then $dH_\phi(x(t), y(t))/dt \leqslant 0$. If, in addition, $\omega^{-1}B + I$ is scrambling, then $dH_\phi(x(t), y(t))/dt < 0$.*

Proof. This is immediate from Theorem 7.1 and Theorem 4.1. ■

The first part of this corollary contains, as a special case, a result quoted and proved, using a different method, by Shigesada and Teramoto (1975, p. 82) and Iwasa (1988). They proved, in effect, that $dH(x(t), y(t))/dt \leqslant 0$ in the special case when $y(t)$ is constant and equal to a positive right eigenvector of B corresponding to the eigenvalue 0.

Define $\xi_\phi(B) = \sup\{d \log H_\phi(x(t), y(t))/dt : t \geqslant 0,\ x(0) \in P_d,\ y(0) \in P_d,\ x(0) \neq y(0)\}$.

COROLLARY 7.3. *For the $d \times d$ intensity matrix B,*

$$\xi_\phi(B) \leqslant \omega(\eta - 1) \leqslant -\beta,$$

where $\eta \equiv \eta_\phi(\omega^{-1}B + I)$ and $\omega \geqslant max_i|b_{ii}|$, and where

$$\beta \equiv \min_{j,k : j<k} \left\{ \sum_{\substack{i \neq j \\ i \neq k}} \left[\min(b_{ij}, b_{ik})\right] + b_{jk} + b_{kj} \right\}.$$

Proof. It follows from (7.1.4) and Theorem 4.1 that

$$\xi_\phi \leqslant \omega(\eta - 1) \leqslant -\omega\alpha(\omega^{-1}B + I).$$

Now if ω is sufficiently large, then

$$\alpha(\omega^{-1}B + I) = \beta/\omega. \quad ■$$

THEOREM 7.4. *Let B be an intensity matrix such that, for some positive constant c, $A_{,} = I + cB$, $x \in N_d$, and $y \in N_d$ all satisfy the assumptions of Theorem 6.1. Then*

$$\lim_{t \to \infty} \frac{d}{dt} \log H_\phi(e^{Bt}x, e^{Bt}y) = 2\lambda_{i_0}.$$

The proof essentially repeats the proof of Theorem 6.1, and will be omitted.

Notes added in proof: S. R. S. Varadhan (4 March 1992) suggested that Theorems 3.1 and 7.1 provide a new and easier way to prove logarithmic Sobolev inequalities (e.g., Gross, L. 1976. Logarithmic Sobolev inequalities, *Amer. J. Math.* 97:1061–1083).

M. B. Ruskai and E. Seneta (8 March 1992) proved that $\eta_{w^2}(A) = \eta_{\log}(A)$ for all A. This answers in the negative the open questions at the ends of Sections 4 and 5.

Gerald S. Goodman, John Hajnal, Marius Iosifescu, and a referee made helpful comments on previous drafts. Gerald S. Goodman first made us aware of the work of Csiszár (1963) and produced a counterexample on which Remark 4.3 is based. The work of J.E.C. was supported in part by U.S. National Science Foundation grant BSR 87-05047 and the Japan Society for the Promotion of Science, and by the hospitality of Mr. and Mrs. William T. Golden, the Institute for Advanced Study, Princeton, New Jersey, and the Departments of Biophysics and Zoology, Kyoto University, Kyoto, Japan. Most of M.B.R.'s contributions to this work were performed while visiting AT & T Bell Laboratories and the Courant Institute of Mathematical Sciences of New York University; her work was partially supported by these institutions and by National Science Foundation grants DMS 88-08112 and DMS 89-08125.

REFERENCES

Abrahams, J. 1982. On the selection of measures of distance between probability distributions, *Inform. Sci.* 26:109–113.

Ahlswede, R. and Gács, P. 1976. Spreading of sets in product spaces and hypercontraction of the Markov operator, *Ann. Probab.* 4(6):925–939.

Alberti, P. M. and Uhlmann, A. 1982. *Stochasticity and Partial Order: Doubly Stochastic Maps and Unitary Mixing*, VEB Deutscher Verlag der Wissenschaften, Berlin; Reidel, Boston.

Ali, S. M. and Silvey, S. D. 1966. A general class of coefficients of divergence of one distribution from another, *J. Roy. Statist. Soc. Ser. B* 28:131–142.

Bauer, F. L., Deutsch, E., and Stoer, J. 1969. Abschätzungen für die Eigenwerte positiver linearer Operatoren, *Linear Algebra Appl.* 2:275–301.

Boltzmann, L. 1877. Über die Beziehung zwischen dem zweiten Hauptsatz der mechanischen Wärmetheorie und der Wahrscheinlichkeitsrechnung respektiv den Sätzen über das Warmgleichgewicht, *Wiener Ber.* 76:373–435.

Cohen, J. E., Friedland, S., Kato, T., and Kelly, F. P. 1982. Eigenvalue inequalities for products of matrix exponentials, *Linear Algebra Appl.* 45:55–95.

Csiszár, I. 1963. Éine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten, *Magyar Tud. Akad. Mat. Kutató Int. Közl.* 8:85–108; MR 29:333, #1671, 1965.

Csiszár, I. 1967. Information-type measures of difference of probability distributions and indirect observations, *Studia Sci. Math. Hungar.* 2:299–318.

Csiszár, I. and Körner, J. 1981. *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic, New York; Akademiai Kiado, Budapest.

Demetrius, L. 1985. The units of selection and measures of fitness, *Proc. Roy. Soc. London Ser. B* 225:147–159.

Dobrushin, R. L. 1956. Central limit theorem for nonstationary Markov chains. I, *Theory Probab. Appl.* 1:65–80; II, *ibid.* 1:329–383.

Donald, M. J. 1986. On the relative entropy, *Comm. Math. Phys.* 105:13–34.

Ellis, R. S. 1985. *Entropy, Large Deviations, and Statistical Mechanics*, Springer-Verlag, New York.

Georgescu-Roegen, N. 1971. *The Entropy Law and the Economic Process*, Harvard U.P., Cambridge.

Goldman, N. and Lord, G. 1986. A new look at entropy and the life table, *Demography* 23:275–282.

Good, I. J. 1983. *Good Thinking*, Univ. of Minnesota Press, Minneapolis.

Hajnal, J. 1958. Weak ergodicity in non-homogeneous Markov chains, *Proc. Cambridge Philos. Soc.* 54:233–246.

Iosifescu, M. 1980. *Finite Markov Processes and Their Applications*, Wiley, New York; Editura Tehnica, Bucharest.

Iwasa, Y. 1988. Free fitness that always increases in evolution, *J. Theoret. Biol.* 135:265–282.

Jaynes, E. 1957. Information theory and statistical mechanics, *Phys. Rev.* 106:620–630.

Joe, H. 1989. Relative entropy measures of multivariate dependence, *J. Amer. Statist. Assoc.* 84(405):157–164.

Kelly, F. P. 1979. *Reversibility and Stochastic Networks*, Wiley, New York.

Kullback, S. 1968. *Information Theory and Statistics*, Dover, New York.

Kullback, S. and Liebler, R. A. 1951. On information and sufficiency, *Ann. Math. Statist.* 22:79–86.

Lieb, E. H. 1975. Some convexity and subadditivity properties of entropy, *Bull. Amer. Math. Soc.* 81:1–13.

Liese, F. and Vajda, I. 1987. *Convex Statistical Distances*, Teubner, Leipzig.

Liggett, T. M. 1985. *Interacting Particle Systems*, Springer-Verlag, New York.

Moran, P. A. P. 1961. Entropy, Markov processes and Boltzmann's H-theorem, *Proc. Cambridge Philos. Soc.* 57:833–842.

Morimoto, T. 1963. Markov processes and the H-theorem, *J. Phys. Soc. Japan* 18:328–331.

Petz, D. 1986a. Sufficient subalgebras and the relative entropy of states of a von Neumann algebra, *Comm. Math. Phys.* 105:123–131.

_____. 1986b. Quasi-entropies for finite quantum systems, *Rep. Math. Phys.* 23:57–65.

Rockafellar, R. T. 1970. *Convex Analysis*, Princeton U.P., Princeton, N.J.

Seneta, E. 1979. Coefficients of ergodicity: Structure and applications, *Adv. Appl. Probab.* 11:576–590.

_____. 1981. *Non-negative Matrices and Markov Chains*, 2nd ed., Springer-Verlag, New York.

____. 1982. Entropy and martingales in Markov chain models, *J. Appl. Probab.* 19A:367–381.

Shannon, C. 1948. A mathematical theory of communication, *Bell System Tech. J.* 27:379–423, 623–656; see also Shannon, C. and Weaver, W., 1949, *The Mathematical Theory of Communication*, Univ. of Illinois, Urbana.

Shigesada, N. and Teramoto, E. 1975. Mathematics of biological reaction systems (in Japanese), in *Life Viewed with Mathematics*, Iwanami, Tokyo, pp. 69–122.

Theil, H. 1967. *Economics and Information Theory*, North-Holland, Amsterdam.

Thirring, W. 1983. *A Course in Mathematical Physics 4. Quantum Mechanics of Large Systems*, Springer-Verlag, New York.

Wehrl, A. 1978. General properties of entropy, *Rev. Modern Phys.* 50:221–260.

White, M. J. 1986. Segregation and diversity measures in population distribution, *Population Index* 52(2):198–221.

Received 23 December 1991