The Annals of Probability
1976, Vol. 4, No. 6, 925-939

SPREADING OF SETS IN PRODUCT SPACES AND HYPERCONTRACTION OF THE MARKOV OPERATOR

BY RUDOLF AHLSWEDE[1] AND PÉTER GÁCS

Ohio State University and Hungarian Academy of Sciences

For a pair of random variables, (X, Y) on the space $\mathscr{X} \times \mathscr{Y}$ and a positive constant, λ, it is an important problem of information theory to look for subsets $\mathscr{A}$ of $\mathscr{X}$ and $\mathscr{B}$ of $\mathscr{Y}$ such that the conditional probability of Y being in $\mathscr{B}$ supposed X is in $\mathscr{A}$ is larger than λ. In many typical situations in order to satisfy this condition, $\mathscr{B}$ must be chosen much larger than $\mathscr{A}$. We shall deal with the most frequently investigated case when $X = (X_1, \cdots, X_n)$, $Y = (Y_1, \cdots, Y_n)$ and (X_i, Y_i) are independent, identically distributed pairs of random variables with a finite range. Suppose that the distribution of (X, Y) is positive for all pairs of values (x, y). We show that if $\mathscr{A}$ and $\mathscr{B}$ satisfy the above condition with a constant λ and the probability of $\mathscr{B}$ goes to 0, then the probability of $\mathscr{A}$ goes even faster to 0. Generalizations and some exact estimates of the exponents of probabilities are given. Our methods reveal an interesting connection with a so-called hypercontraction phenomenon in theoretical physics.

1. Introduction. For a pair of random variables (X, Y) on the space $\mathscr{X} \times \mathscr{Y}$ and positive constant λ, it is an important problem of information theory to look for pairs of sets $\mathscr{A} \subset \mathscr{X}$, $\mathscr{B} \subset \mathscr{Y}$ such that

$$\Pr[Y \in \mathscr{B} \mid X \in \mathscr{A}] \geqq \lambda . \tag{1.1}$$

In many typical situations in order to satisfy (1.1), $\Pr[Y \in \mathscr{B}]$ must be much larger than $\Pr[X \in \mathscr{A}]$. We shall deal with the most frequently investigated case where

$$\mathscr{X} = \mathscr{X}_1 \times \cdots \times \mathscr{X}_n , \qquad \mathscr{Y} = \mathscr{Y}_1 \times \cdots \times \mathscr{Y}_n ,$$
$$X = (X_1, \cdots, X_n) , \qquad Y = (Y_1, \cdots, Y_n) ,$$

and (X_i, Y_i) are independent, identically distributed pairs of random variables with finite ranges $\mathscr{X}_i \times \mathscr{Y}_i$. We use the notations

$$\mathscr{X}^n = \mathscr{X}_1 \times \cdots \times \mathscr{X}_n , \qquad X^n = (X_1, \cdots, X_n) .$$

Thus (1.1) turns into

$$\Pr[Y^n \in \mathscr{B} \mid X^n \in \mathscr{A}] \geqq \lambda . \tag{1.2}$$

Suppose for a moment that the distribution of (X_i, Y_i) is fixed in such a way that for all $x \in \mathscr{X}_i$, $y \in \mathscr{Y}_i$ we have

$$\Pr[X_i = x, Y_i = y] > 0 .$$

Received August 22, 1975.

[1] Research of this author was supported by the Deutsche Forschungsgemeinschaft.

AMS 1970 *subject classifications.* 94A15, 60J35.

Key words and phrases. Multiuser communication, Markov operator, Kullback *I*-divergence, common information, maximal correlation, hypercontraction.

Let us fix λ independently of n. We shall show that if $\mathscr{A}$ and $\mathscr{B}$ satisfy (1.2) and $\Pr[Y^n \in \mathscr{B}] \to 0$ then $\Pr[X^n \in \mathscr{A}]/\Pr[Y^n \in \mathscr{B}] \to 0$ uniformly in n. Another formulation of the same result is that for every $\lambda_1 > 0$ and $\lambda_2 > 0$ there is a $\lambda_3 > 0$ (independent of n) such that $\Pr[Y^n \in \mathscr{B} | X^n \in \mathscr{A}] \geqq \lambda_1$, $\Pr[X^n \in \mathscr{A} | Y^n \in \mathscr{B}] \geqq \lambda_2$ implies $\Pr[X^n \in \mathscr{A}, Y^n \in \mathscr{B}] \geqq \lambda_3$. This says that in the product spaces $\mathscr{X}^n$, $\mathscr{Y}^n$ there are no pairs of small sets going into each other with a large (i.e., constant) probability.

Actually, our results are sharper than this. Under the above conditions, we show that there is an $r > 1$ such that with some function $c(\lambda) > 0$, (1.2) implies

$$\Pr[Y^n \in \mathscr{B}] \geqq c(\lambda) \Pr[X^n \in \mathscr{A}]^r \tag{1.3}$$

for all n. This result can be interpreted as follows. If we know by chance that the random sequence X^n is in the set $\mathscr{A}$ so that we have some "information" of the amount $-\log \Pr[X^n \in \mathscr{A}]$ about X^n, our information, in any reasonable sense, about Y^n, will be only $-r \log \Pr[X^n \in \mathscr{A}]$ i.e., a constant times less.

In [1] we have shown that if $\mathscr{A}$ and $\mathscr{B}$ satisfy (1.2), then $n^{-1} \log \Pr[Y^n \in \mathscr{B} | X^n \in \mathscr{A}] \to 0$ and $n^{-1} \log \Pr[X^n \in \mathscr{A} | Y^n \in \mathscr{B}] \to 0$ implies $n^{-1} \log \Pr[X^n \in \mathscr{A}, Y^n \in \mathscr{B}] \to 0$. That is, we cannot have two exponentially small sets $\mathscr{A}$, $\mathscr{B}$ going into each other with greater than exponentially small probability. Witsenhausen showed that if both conditional probabilities are larger than some constant depending on the distribution of (X_i, Y_i) then $\Pr[X^n \in \mathscr{A}, Y^n \in \mathscr{B}]$ is also larger than some constant $\lambda_3 > 0$.

In [2] we have investigated pairs of sets $\mathscr{A}$, $\mathscr{B}$ satisfying (1.2) and having probabilities which are exponentially small in n. We have determined all possible pairs of exponents. From the results proved there we later deduced together with J. Körner and I. Csiszár that if in the condition (1.2) λ goes to 1 as $n \to \infty$ then with an appropriate r we have (1.3) (we have also determined the best r). In this paper we remove the condition $\lambda \to 1$. To formulate the result more generally, we first give a condition weaker than the positivity of $\Pr[X_i = x, Y_i = y]$. The distribution of (X, Y) is called *decomposable* if there exist $\mathscr{A}$, $\mathscr{B}$ such that $0 < \Pr[X \in \mathscr{A}]$, $\Pr[Y \in \mathscr{B}] < 1$, $\Pr[X \in \mathscr{A}$ if and only if $Y \in \mathscr{B}] = 1$. Note that if $\Pr[X = x, Y = y]$ is positive for all pairs then the distribution of (X, Y) is indecomposable.

THEOREM 1. *If the distribution of (X_i, Y_i) is indecomposable then there exist $p > 0$, $r < 1$ such that for all n, $\mathscr{A}$, $\mathscr{B}$*

$$\Pr[Y \in \mathscr{B}] \geqq \Pr[Y^n \in \mathscr{B} | X^n \in \mathscr{A}]^p \cdot \Pr[X^n \in \mathscr{A}]^r .$$

REMARK. A more symmetrical formulation is: there are σ, τ with $0 < \sigma, \tau < 1$, $\sigma + \tau > 1$ such that

$$\Pr[X^n \in \mathscr{A}, Y^n \in \mathscr{B}] \leqq \Pr[X^n \in \mathscr{A}]^\sigma \cdot \Pr[Y^n \in \mathscr{B}]^\tau .$$

Note that if $\sigma + \tau = 1$ then this inequality follows from Hölder's inequality without any conditions on the distribution of (X_i, Y_i).

However, our later more sharp results as well as the interpretation concern the nonsymmetrical form.

In this paper we determine the best constant r as well as the best r which is appropriate for any input distribution $\Pr[X = x]$ at a fixed transition probability matrix

$$\Pr[Y = y \mid X = x] \quad (x \in \mathscr{X}, y \in \mathscr{Y}) .$$

In proving Theorem 1 we shall actually prove more. Let us fix an L_p-norm for the functions defined on $\mathscr{X}^n$. Then there is an L_q-norm with $q/p \leqq r$ for the functions defined on $\mathscr{Y}^n$ such that the Markov operator T^n defined by

$$(T^n g)(x) = E[g(Y^n) \mid X^n = x]$$

takes all functions g with $\|g\|_q \leqq 1$ to functions $T^n g$ satisfying $\|T^n g\|_p \leqq 1$.

As Professor Dobrushin noted, this so-called hypercontracting property of the Markov operator as well as the problem of determining the best q was independently considered in theoretical physics (for Gaussian distributions, see [6]).

At the end we illustrate the results on the case of binary random variables.

We are indebted to I. Csiszár and J. Körner for their contributions to Theorem 4 and Theorem 9 and for several useful discussions about the problem. We also are thankful to J. Komlós and Major for their valuable advice.

2. Statement of the main results.

A. *Hypercontraction of the Markov operator.* In order to keep the notation simple we denote the elements of $\mathscr{X}$ (resp. $\mathscr{Y}$) and also the elements of $\mathscr{X}^n$ (resp. $\mathscr{Y}^n$) by x (resp. y). It will be always clear from the context with which set we are dealing. Let us also use the abbreviations

$$w^n(\mathscr{B} \mid x) = \Pr[Y^n \in \mathscr{B} \mid X^n = x]$$

$$P^n(\mathscr{A}) = \Pr[X^n \in \mathscr{A}], \qquad Q^n(\mathscr{B}) = \Pr[Y^n \in \mathscr{B}]$$

$$P^1 = P, \qquad P^n(x) = P^n(\{x\}) .$$

The transition probability matrix $\{w^n(y \mid x)\}$ is denoted by W^n. We can always assume without loss of generality that

$$P(x), Q(x) > 0 \qquad \text{for all} \quad x \in \mathscr{X}, \quad y \in \mathscr{Y} .$$

The simultaneous distribution of the random variables X, Y will be denoted by (W, P).

We denote by $\mathscr{F}(\mathscr{X})$ the set of all real-valued functions defined on the set $\mathscr{X}$ and define the Markov operator $T: \mathscr{F}(\mathscr{Y}) \to \mathscr{F}(\mathscr{X})$ by

$$(Tg)(x) = \sum_{y \in \mathscr{Y}} w(y \mid x) g(y) = E[g(Y) \mid X = x] . \tag{2.1}$$

The operator $T^n: \mathscr{F}(\mathscr{Y}^n) \to \mathscr{F}(\mathscr{X}^n)$ is then the tensor power of T. Notice that if $f_{\mathscr{B}}(x) = w^n(\mathscr{B} \mid x)$ and $1_{\mathscr{B}}(x)$ is the indicator function of $\mathscr{B}$ then $f_{\mathscr{B}} = T^n 1_{\mathscr{B}}$.

For any $p \geqq 1$ we denote by $s_p(W, P)$ the minimum of those r which satisfy for every $g \in \mathscr{F}(\mathscr{Y})$ the inequality

$$\{E[(Tg)(X)]^p\}^{1/p} \leqq \{Eg(Y)^{rp}\}^{1/rp} . \tag{2.2}$$

If we consider $\mathscr{F}(\mathscr{X})$ and $\mathscr{F}(\mathscr{Y})$ together with the underlying measures P and Q, then (2.2) can be written as

$$\|Tg\|_p \leqq \|g\|_{rp} \tag{2.3}$$

where $\|\cdot\|_p$ is the L_p-norm, integration taken with respect to the underlying measures. (It will turn out that rp is never less than 1.)

$s_p(n) = s_p(W^n, P^n)$ shows many similarities with the maximal correlation $\rho(W^n, P^n)$. The maximal correlation $\rho(W, P)$ of X and Y is defined as the maximum of

$$Ef(X)g(Y)$$

for those functions f defined on $\mathscr{X}$, g defined on $\mathscr{Y}$ satisfying

$$Ef(X) = Eg(Y) = 0\,, \qquad Ef^2(x) = Eg^2(Y) = 1\,.$$

ρ is 0 iff X and Y are independent and 1 iff the distribution of (X, Y) is decomposable. (See, for example, [6].) If (X_i, Y_i) $(i = 1, \cdots, n)$ are independent but not necessarily equidistributed pairs with distributions (W_i, P_i) then (see [2], [6])

$$\rho(W^n, P^n) = \max_i \rho(W_i, P_i)\,.$$

The following two theorems insure

$$s_p(n) = s_p(1) < 1 \qquad \text{for} \quad p > 1$$

and together with Lemma 1 below give Theorem 1 as an immediate consequence.

THEOREM 2. *Let (X_i, Y_i) $(i = 1, \cdots, n)$ be independent pairs of random variables —not necessarily equidistributed—with distributions (W_i, P_i) and corresponding Markov operators T_i.*

We have

$$s_p(W^n, P^n) = \max_i s_p(W_i, P_i)\,.$$

THEOREM 3.

(a) *$s_p \geqq p^{-1}$ with equality if and only if X and Y are independent. $s_1 = 1$, s_p is monotonically decreasing in p.*

(b) *$s_p \geqq \rho^2 + p^{-1}(1 - \rho^2)$ where ρ is the maximal correlation.*

(c) *If (W, P) is indecomposable then s_p is strictly decreasing in p.*

LEMMA 1. *For all n, $\mathscr{A} \subset \mathscr{X}^n$, $\mathscr{B} \subset \mathscr{Y}^n$, $p \geqq 1$ we have* (denoting s_p by r)

$$Q^n(\mathscr{B}) \geqq \Pr[Y^n \in \mathscr{B} \mid X^n \in \mathscr{A}]^{pr} \cdot P^n(\mathscr{A})^r\,. \tag{2.4}$$

PROOF. We have by Hölder's inequality

$$\begin{aligned}\Pr[X^n \in \mathscr{A}, Y^n \in \mathscr{B}] &= E1_{\mathscr{A}}(X^n) f_{\mathscr{B}}(X^n) \\ &\leqq P(\mathscr{A})^{1-1/p} \cdot \{Ef_{\mathscr{B}}(X^n)^p\}^{1/p} \leqq P(\mathscr{A})^{1-1/p} \cdot Q(\mathscr{B})^{1/pr}\,.\end{aligned}$$

Rearrangement gives (2.4).

B. *λ-kernels, connection between the L_p-norm and the I-divergence.* In order to

state our next results we need the following definitions. The "λ-kernel" $\Psi_\lambda(\mathscr{B})$ of a set $\mathscr{B} \subset \mathscr{Y}^n$ is defined by

$$\Psi_\lambda(\mathscr{B}) = \{x \in \mathscr{X}^n \mid w^n(\mathscr{B} \mid x) \geqq \lambda\} . \tag{2.5}$$

(This notation is different from the one used in [1]: the set denoted here by $\Psi_\lambda(\mathscr{B})$ was denoted there by $\Psi_{1-\lambda}(\mathscr{B})$.)

For a finite set $\mathscr{X}$ and probability distributions R, S on $\mathscr{X}$ we define the relative entropy of R (it is the negative of the I-divergence of Kullback [4]) with respect to S by

$$H_S(R) = \sum_z R(z) \log \frac{S(z)}{R(z)} . \tag{2.6}$$

It is known that $H_S(R) \leqq 0$ and equality holds if and only if $R = S$.

For a distribution R over $\mathscr{X}$ we define the distribution T^*R over $\mathscr{Y}$ by

$$(T^*R)(\mathscr{B}) = \sum_x w(\mathscr{B} \mid x) R(x) . \tag{2.7}$$

Note that $Q = T^*P$. The quantity

$$\underline{s} = \underline{s}(W, P) = \sup_{R: R \neq P} \frac{H_Q(T^*R)}{H_P(R)} \tag{2.8}$$

will play an important role in the sequel. As was shown in [1], the behaviour of the function

$$D_n(\lambda, \delta, W, P) = \max_{\mathscr{B}: Q^n(\mathscr{B}) < \delta} \frac{\log Q^n(\mathscr{B})}{\log P^n(\Psi_\lambda(\mathscr{B}))} \tag{2.9}$$

is of particular interest in multiuser communication theory. We use the abbreviation

$$D_n(\lambda, \delta) = D_n(\lambda, \delta, W, P) .$$

The function is monotone in λ, n and δ. Therefore the following limits exist:

$$D(\lambda, \delta) = \lim_{n \to \infty} D_n(\lambda, \delta) , \qquad D(\lambda) = \lim_{\delta \to 0} D(\lambda, \delta) . \tag{2.10}$$

From [1] we derived together with I. Csiszár and J. Körner

THEOREM 4.

(a) $\lim_{\lambda \to 1} D(\lambda) = \underline{s}$.

(b) *If* (W, P) *is indecomposable then* $\underline{s} < 1$.

It is easy to show that $\underline{s}$ is 0 iff X and Y are independent and $\underline{s} = 1$ if (W, P) is decomposable. (b) of Theorem 4 is also of independent interest. It says that if (W, P) is indecomposable then the Kullback I-divergence-$H_Q(T^*R)$ of the output distribution T^*R from Q is by a constant multiple less than the I-divergence of the input distribution R from P.

Our main task is to investigate the behaviour of $D(\lambda, \delta)$, in the case when λ does not converge to 1. For this we need in addition to the results stated in so far a further theorem which relates the quantities $s_p = s_p(W, P)$ and $\underline{s} = s(W, P)$.

Since by Theorem 2, $s_p(n) = s_p(1)$, one obtains the best estimate in (2.4) by replacing $s_p(n)$ by s, where

$$s = s(W, P) = \inf_{p \geq 1} s_p(W, P) . \tag{2.11}$$

THEOREM 5.

(a) $\underline{s} = s$ *and* s *is the minimum of those* r *satisfying*

$$E \prod_y g(y)^{w(y|X)} \leqq \{Eg(Y)^r\}^{1/r} \tag{2.12}$$

for every $g \in \mathscr{F}(\mathscr{Y})$, $g \geqq 0$.

(b) *There is a constant* $c(W)$ *depending only on* W *such that*

$$s_p(W, P) - s(W, P) \leqq c(W) \cdot p^{-1} .$$

Notice that we have as a consequence of Theorems 2.5 and Definition (2.11) that

$$\underline{s}(W^n, P^n) = \underline{s}(W, P) .$$

This can also be derived from results of [1].

Now we are ready to state our main result about $D(\lambda, \delta)$, which goes beyond Theorem 4. New in this result is not only the fact that $D(\lambda)$ equals $\underline{s}$ but also that it is less than one.

THEOREM 6.

(a) $D(\lambda) = \underline{s}$ *for all* λ *with* $0 < \lambda < 1$.

(b) *For* $\delta < \lambda^2$

$$D(\lambda, \delta) \leqq \underline{s} + O\left(\frac{\log \lambda}{\log \delta}\right)^{\frac{1}{2}} .$$

PROOF. Since evidently

$$\Pr [Y^n \in \mathscr{B} \mid X^n \in \Psi_\lambda(\mathscr{B})] \geqq \lambda ,$$

Lemma 1 yields

$$\begin{aligned} \frac{\log Q^n(\mathscr{B})}{\log P^n(\Psi_\lambda(\mathscr{B}))} &\leqq s_p(n) + \frac{s_p(n) \cdot p \log \lambda}{\log P^n(\Psi_\lambda(\mathscr{B}))} \\ &\leqq s_p(n) + \frac{p \log \lambda}{\log Q^n(\mathscr{B}) - \log \lambda} . \end{aligned} \tag{2.13}$$

This and Theorem 2 imply

$$D_n(\lambda, \delta) \leqq s_p + \frac{p \log \lambda}{\log \delta - \log \lambda} \tag{2.14}$$

for all n and therefore also

$$D(\lambda, \delta) \leqq s_p + \frac{p \log \lambda}{\log \delta - \log \lambda} . \tag{2.15}$$

The right-hand side tends to s_p as δ goes to 0. Therefore

$$D(\lambda) \leqq s = \inf_{p \geq 1} s_p , \qquad 0 < \lambda < 1 . \tag{2.16}$$

Since $s = \underline{s}$ (by Theorem 5) and since $D(\lambda)$ is decreasing in λ, (a) follows from (2.16) and Theorem 4. Using $\delta < \lambda^2$ and (b) of Theorem 5 we obtain from (2.15)

$$D(\lambda, \delta) \leqq s + O\left(p^{-1} + p\frac{\log \lambda}{\log \delta}\right). \tag{2.17}$$

Choosing $p = (\log \lambda/\log \delta)^{-\frac{1}{2}}$ one gets (b). □

Notice that (2.13) is equivalent to

$$P^n(\Psi_\lambda(\mathscr{B})) \leqq \lambda^{-p} Q^n(\mathscr{B})^{1/r} \tag{2.18}$$

where $r = s_p(n)$.

By Theorems 2, 3 if (W, P) is indecomposable, then $s_p(n) = s_p < 1$,

and hence we have the

COROLLARY 1. *If (W, P) is indecomposable, then there is a constant $r < 1$, and for all λ $(0 < \lambda < 1)$ a $c(\lambda) > 0$ such that for all n and $\mathscr{B}$,*

$$P^n(\Psi_\lambda(\mathscr{B})) \leqq c(\lambda) Q^n(\mathscr{B})^{1/r}. \tag{2.19}$$

C. *Relation to maximal correlation.* We mentioned already earlier that Theorem 2 expresses a property of s_p and hence by Theorem 5 also of s and $\underline{s}$, which is very familiar for the maximal correlation. The following result establishes a connection between $s = \underline{s}$ and

THEOREM 7. *The following properties of (X, Y) are equivalent and imply $s = \rho^2$:*

(i) *The inequality*

$$H_Q(T^*R)/H_P(R) \leqq s \tag{2.20}$$

is always strict unless $R = P$.

(ii) *The inequality*

$$E \prod_y g(y)^{w(y|X)} \leqq \{Eg(Y)^s\}^{1/s} \tag{2.21}$$

is always strict unless g is a constant.

D. *Bounds on $\bar{D}_n(\lambda, \delta, W) = \max_P D_n(\lambda, \delta, W, P)$.* Let us now turn to the problem of finding an estimate on $D_n(\lambda, \delta, W, P)$ which is independent of the distribution P. This problem has arisen in [1]. We define

$$\bar{D}(\lambda, \delta) = \lim_{n\to\infty} \max_P D_n(\lambda, \delta, W, P) \tag{2.22}$$

and

$$\bar{\rho}(W) = \max_P \rho(W, P). \tag{2.23}$$

THEOREM 8.

$$\max_P s(W, P) = \bar{\rho}^2(W).$$

(For the definition of $s(W, P)$ see (2.11).)

Notice that $c(W)$ in Theorem 4 is independent of P. We therefore obtain as a consequence of Theorem 6 and Theorem 8

COROLLARY 2.

(a) $$\bar{D}(\lambda) = \lim_{\delta \to 0} \bar{D}(\lambda, \delta) = \bar{\rho}^2(W) .$$

(b) *For* $\delta < \lambda^2$ *we have*

$$\bar{D}(\lambda, \delta) \leqq \rho^2(W) + O\left(\frac{\log \lambda}{\log \delta}\right)^{\frac{1}{2}} .$$

REMARK. A simple necessary and sufficient condition for $\bar{\rho}(W) < 1$ can be given: for every pair $x_1, x_2 \in \mathscr{X}$ there exists a $y \in \mathscr{Y}$ such that

$$w(y \mid x_1) \cdot w(y \mid x_2) > 0 .$$

E. *Illustration of the behaviour of* $s(W, P)$ *in the binary case.* Suppose that

$$\mathscr{X} = \mathscr{Y} = \{0, 1\} \qquad \text{and}$$
$$W_{\alpha\beta} = \begin{pmatrix} 1-\alpha & \alpha \\ \beta & 1-\beta \end{pmatrix} \quad (0 < \alpha, \beta < \tfrac{1}{2}) .$$

We denote by $P_{\alpha\beta}$ the stationary input probability distribution:

$$Q_{\alpha\beta} = T^* P_{\alpha\beta} = P_{\alpha\beta} .$$

Evidently,

$$P_{\alpha\alpha}(0) = P_{\alpha\alpha}(1) = \tfrac{1}{2} .$$

THEOREM 9.

(a) $$s(W_{\alpha\alpha}, P_{\alpha\alpha}) = \rho^2(W_{\alpha\alpha}, P_{\alpha\alpha}) = \bar{\rho}^2(W_{\alpha\alpha}) = (1 - 2\alpha)^2 .$$

(b) *If* $\alpha \neq \beta$ *then*

$$s(W_{\alpha\beta}, P_{\alpha\beta}) > \rho^2(W_{\alpha\beta}, P_{\alpha\beta}) .$$

3. Norm improvement of the Markov operator, proof of Theorems 2 and 3.

PROOF OF THEOREM 2. It is enough to prove the statement for $n = 2$. The general case is then proved by induction.

(1) $$s_p(W^2, P^2) \geqq \max_{i=1,2} s_p(W_i, P_i) .$$

Let us define the operator

$$\Pi_1 : \mathscr{F}(\mathscr{Y}_1) \to \mathscr{F}(\mathscr{Y}^2) \qquad \text{by} \quad (\Pi_1 g)(y_1, y_2) = g(y_1) .$$

Then we have for any $g \in \mathscr{F}(\mathscr{Y}_1)$, $\|\Pi_1 g\|_p = \|g\|_p$, $\|T^2 \Pi_1 g\|_p = \|T_1 g\|_p$. Hence $s_p(W^2, P^2) \geqq s_p(W_1, P_1)$. The same holds for $s_p(W_2, P_2)$.

(2) $$s_p(W^2, P^2) \leqq \max_{i=1,2} s_p(W_i, P_i) .$$

For a function h in some $\mathscr{F}(\mathscr{X}_1 \times \mathscr{X}_2)$ and any $z_1 \in \mathscr{X}_1$ we define $h_{z_1} \in \mathscr{F}(\mathscr{X}_2)$ by

$$h_{z_1}(z_2) = h(z_1, z_2) .$$

Let us write, for a moment,

$$r = \max_{i=1,2} s_p(W_i, P_i) .$$

Then we have

$$\begin{aligned} ||T^2g||_p &= \{\textstyle\sum_{x_1x_2} P_1(x_1)P_2(x_2)[(T^2g)(x_1, x_2)]^p\}^{1/p} \\ &= \{\textstyle\sum_{x_1} P(x_1)E[(T^2g)_{x_1}(X_2)]^p\}^{1/p} \\ &= \{E||(T^2g)_{X_1}||_p^p\}^{1/p} . \end{aligned}$$

Now

$$(T^2g)_{x_1} = \textstyle\sum_{y_1} w(y_1 \mid x_1)T_2g_{y_1} .$$

By Minkowski's inequality

$$||(T^2g)_{x_1}||_p \leqq \textstyle\sum_{y_1} w(y_1 \mid x_1)||T_2g_{y_1}||_p .$$

If we define $h(y_1) = ||T_2g_{y_1}||_p$, then the right-hand side is equal to $(T_1h)(x_1)$. Thus we have by the definition of r:

$$||T^2g||_p \leqq \{E[(T_1h)(X_1)]^p\}^{1/p} \leqq \{Eh(Y_1)^{rp}\}^{1/rp} = \{E||T_2g_{Y_1}||_p^{rp}\}^{1/rp} \leqq ||g||_{rp} ,$$

which completes the proof.

PROOF OF THEOREM 3. We shall use a simple fact, well known from elementary calculus.

Fact 1. Let Q be a distribution on $\mathscr{Y}$, $g \in \mathscr{F}(\mathscr{Y})$, $g \geqq 0$. Then the expression

$$||g||_q = ||\textstyle\sum_y Q(y)g(y)^q||^{1/q}$$

is increasing in q ($q > 0$) and

$$\lim_{q\to 0}||g||_q = \textstyle\prod_y g(y)^{Q(y)} \qquad \text{where} \quad 0^0 = 1 .$$

If $\exists\, y_0, y_1$ with $g(y_0) \neq g(y_1)$ and $Q(y_0), Q(y_1) > 0$, then $||g||_q$ is strictly increasing.

Proof of (a). It is easy to see that (2.3) holds for all $g \in F(y)$ iff it holds for all $g \geqq 0$. Choosing now $g^{1/rp}$ instead of g, (2.2) can be written as

$$\{E[\textstyle\sum_y w(y \mid X)g(y)^{1/rp}]^{rp\cdot 1/r}\}^r \leqq Eg(Y) . \tag{3.1}$$

It follows from Fact 1 that the left-hand side of (3.1) is decreasing in p. But then s_p is also decreasing, which was to be proved. The other statements of (a) are trivial.

(b) Since all our inequalities are homogeneous, we can normalize them. Let us define $\mathscr{F}_+^1 = \mathscr{F}_+^1(\mathscr{Y})$ by

$$\mathscr{F}_+^1(\mathscr{Y}) = \{g \mid g \in \mathscr{F}(\mathscr{Y}), g \geqq 0, Eg(Y) = 1\} . \tag{3.2}$$

Set

$$F_{p,r}(g) = E[(Tg^{1/rp})(X)]^p . \tag{3.3}$$

(2.2) is then equivalent to

$$F_{p,r}(g) \leqq 1 \qquad \text{for all} \quad g \in \mathscr{F}_+^1 . \tag{3.4}$$

Finally, we define g_0 by

$$g_0(y) = 1 \qquad \text{for all} \quad y \in \mathscr{Y}, \tag{3.5}$$

and r_p by

$$r_p = \rho^2 + p^{-1}(1 - \rho^2)\,. \tag{3.6}$$

We complete the proof of (b) and show that in a small neighbourhood of g_0, r_p can be substituted by a value close to r_p.

LEMMA 2.

(a) $s_p \geqq r_p$.

(b) *For every* $\varepsilon > 0$ *there exists a neighbourhood* $U_\varepsilon(g_0)$ *of* g_0 *in* $\mathscr{F}_+^1$ *such that for all* $p \geqq 1$, $r \geqq r_p + \varepsilon$ *implies* $F_{p,r}(g) \leqq 1$ *for all* $g \in U_\varepsilon(g_0)$.

PROOF. We consider $F_{p,r}(g)$ as a function of the vector g and differentiate it partially.

$$\frac{\partial F_{p,r}}{\partial g(y)}(g) = r^{-1}g(y)^{1/rp-1}Ew(y\,|\,X)[(Tg^{1/rp})(X)]^{p-1}\,,$$

$$\frac{\partial^2 F_{p,r}}{\partial g(y_0)\,\partial g(y_1)}(g) = r^{-2}(1 - p^{-1})[g(y_0)g(y_1)]^{1/rp-1}Ew(y_0\,|\,X)w(y_1\,|\,X)[(Tg^{1/rp})(X)]^{p-2}$$
$$+ \delta_{y_0y_1}r^{-1}(1/rp - 1)g(y_0)^{1/rp-2}Ew(y_0\,|\,X)[(Tg^{1/rp})(X)]^{p-1}\,.$$

Here $\delta_{y_0y_1}$ is the Kronecker symbol. It is easy to see that these expressions converge uniformly in p to their values at g_0 as $g \to g_0$. Now we have

$$\frac{\partial F_{p,r}}{\partial g(y)}(g_0) = r^{-1}Q(y)\,,$$

$$\frac{\partial^2 F_{p,r}}{\partial g(y_0)\,\partial g(y_1)}(g_0) = r^{-2}(1 - p^{-1})\left\{Ew(y_0\,|\,X)w(y_1\,|\,X) - \frac{\delta_{y_0y_1}Q(y_0)(rp - 1)}{p - 1}\right\}.$$

Define now

$$\mathscr{F}^0 = \mathscr{F}^0(\mathscr{Y}) = \{h \in \mathscr{F}(\mathscr{Y}) \,|\, Eh(Y) = 0\}\,. \tag{3.7}$$

It is known (see [6]) that the quadratic form

$$\sum_{y_0,y_1} \frac{\partial^2 F_{p,r}}{\partial g(y_0)\,\partial g(y_1)}(g_0)h(y_0)h(y_1)$$

is negative semidefinite in the space $\mathscr{F}^0$ iff

$$\rho^2 \leqq \frac{rp - 1}{p - 1}\,, \tag{3.8}$$

and is negative definite iff strict inequality holds in (3.8). This fact, together with the uniform convergence, proves Lemma 1 completely, because (3.8) is equivalent to (a) of Lemma 2.

(c) Suppose $p < p'$, $\varepsilon = (r_p - r_{p'})/2$. Then, with $r' = r_p + \varepsilon$, $F_{p',r'}(g) \leqq 1$ for all $g \in U_\varepsilon(g_0)$. Put $r = s_p$. We claim that

$$F_{p',r}(g) < F_{p,r}(g) \qquad \text{for all} \quad g \neq g_0\,. \tag{3.9}$$

Let us suppose that $g \neq g_0$.

We denote

$$\mathscr{B} = \{y_0 \in \mathscr{Y} \mid g(y_0) = \max_y g(y)\},$$
$$\mathscr{A} = \{x \in \mathscr{X} \mid \exists\, y \in \mathscr{B} \quad w(y \mid x) > 0\}.$$

Since $g \neq g_0$ we have $0 < P(\mathscr{B}) < 1$. We have assumed that the distribution of (X, Y) is indecomposable. Thus there is a pair (x, y) such that $w(y \mid x) > 0$ and either $x \in \mathscr{A}$, $y \notin \mathscr{B}$ or $x \notin \mathscr{A}$, $y \in \mathscr{B}$. The later possibility is ruled out by the construction of $\mathscr{A}$. Therefore we have an $x \in \mathscr{X}$ such that there are $y_0 \in \mathscr{B}$, $y_1 \notin \mathscr{B}$ with $w(y_i \mid x) > 0$ $(i = 0, 1)$. Hence by Fact 1

$$[\textstyle\sum_y w(y \mid x) g(y)^{1/rp}]^{rp} > [\sum_y w(y \mid x) g(y)^{1/rp'}]^{rp'}, \tag{3.10}$$

and this implies (3.9).

Now, the set $\mathscr{T}_+^1 - U_\varepsilon(g_0)$ is compact and on it, (3.9) holds. By the continuity of $F_{p,r}(g)$ we can choose an $r'' < r = s_p$ such that

$$F_{p',r''}(g) \leqq 1 \qquad \text{for all} \quad g \notin U_\varepsilon(g_0) \tag{3.11}$$

and we obtain

$$s_{p'} \leqq \max(r', r'') < s_p, \tag{3.12}$$

which was to be proved.

4. λ-kernels, proof of Theorem 4. We prove Theorem 4 by proving the following three statements:

(a_1) $$D(\lambda) \geqq \underline{s} \qquad \text{for all} \quad \lambda < 1,$$

(a_2) $$\lim_{\lambda \to 1} D(\lambda) \leqq \underline{s},$$

(b) If (W, P) is indecomposable then $\underline{s} < 1$.

Even though they are provable independently (a_2) follows from (2.16) and Theorem 5 and (b) follows from Theorem 3 and Theorem 5. Thus we have to prove only (a_1).

In [1] we have pointed out that (in our notation) for every distribution $R \neq P$ over $\mathscr{X}$ there exists a sequence $\mathscr{B}_n$ of subsets of $\mathscr{Y}^n$ such that for any $\lambda > 0$,

$$\lim_n n^{-1} \log Q^n(\mathscr{B}_n) = H_Q(T^*R),$$
$$\liminf_n n^{-1} \log P^n(\Psi_\lambda(\mathscr{B}_n)) \geqq H_P(R).$$

Now fix a distribution $R \neq P$ such that $H_Q(T^*R) \neq 0$. Denote by $n(\delta)$ the least integer n such that $Q^n(\mathscr{B}_n) \leqq \delta$. Then we have

$$\liminf_{\delta \to 0} D_{n(\delta)}(\lambda, \delta) \geqq H_Q(T^*R)/H_P(R).$$

The left side is clearly not greater than $D(\lambda)$ which completes the proof of (a_1).

5. Connections between the L_p-norm, the I-divergences and the maximal correlation. Proofs of Theorems 5, 7 and 8. We use the function

$$G_r(g) = E \textstyle\prod_y g(y)^{w(y \mid X)/r}. \tag{5.1}$$

A. *Proof of Theorems* 5 *and* 7. We denote by s^* the minimum of those r for which $G_r(g) \leqq 1$ for all $g \in \mathscr{T}_+^1$, that is the minimal r for which (2.12) holds.

(a_1) First we show that $s = s^*$. By Fact 1 we have $G_r \leqq F_{p,r}$. This proves $s^* \leqq s$. Let us fix now an $r > s^*$. Put

$$\mathscr{F}_r = \{g \in \mathscr{F}_+^1 \mid G_r(g) = 1\} .$$

We shall show that $\mathscr{F}_r = \{g_0\}$. If $g_1 \in \mathscr{F}_r$, then clearly for all x

$$\prod_y g_1(y)^{w(y|x)} = 1$$

by Fact 1. Hence $g_1 > 0$. Let us compute the first derivative of $G_r(g)$ in a $g_1 > 0$.

$$\frac{\partial G_r}{\partial g(y_0)} = r^{-1} g_1(y_0)^{-1} E w(y_0 \mid X) \prod_y g_1(y)^{w(y|X)/r} . \tag{5.2}$$

If $g_1 \in \mathscr{F}_r$ then this is equal to $r^{-1} g_1(y_0)^{-1} Q(y_0)$. Now $G_r(g)$ has a maximum at g_1 in $\mathscr{F}_+^1$. Hence, by the theorem on Lagrange's multipliers, $g_1^{-1} Q$ must be proportional to Q. This is possible only for $g_1 = g_0$. Hence

$$\mathscr{F}_r = \{g_0\} .$$

The next step is to show that $s^* \geqq \rho^2$. This follows from the fact that the quadratic form of the second partial derivatives of G_r at g_0 is negative semi-definite in $\mathscr{F}^0$ iff $r \geqq \rho^2$.

For a fixed $r > s^*$, let us choose now $\varepsilon = (r - \rho^2)/2$ and find a p such that

$$r_p \leqq \rho^2 + \varepsilon .$$

Then we have $r \geqq r_p + \varepsilon$, and hence by Lemma 2,

$$F_{p,r}(g) \leqq 1 \qquad \text{for all} \quad g \in U_\varepsilon(g_0) .$$

On the other hand, on the compact set $\mathscr{F}_+^1 - U_\varepsilon(g_0)$

$$\lim_{p \to \infty} [\textstyle\sum_y w(y \mid x) g(y)^{1/p}]^p = \prod_y g(y)^{w(y|x)} \tag{5.3}$$

holds uniformly in g, and thus $\lim_{p \to \infty} F_{p,r}(g) = G_r(g) < 1$ uniformly in $g \in \mathscr{F}_+^1 - U_\varepsilon(g_0)$. Choose a p' such that for all $g \in \mathscr{F}_+^1 - U_\varepsilon(g_0)$ $F_{p',r}(g) \leqq 1$. Then we have

$$r \geqq \min (s_p, s_{p'}) = s_{\max(p,p')} \tag{5.4}$$

which proves $s = s^*$.

(a_2) To complete the proof of Theorem 5 (a) we have to prove $s = \underline{s}$. First we show $s \geqq \underline{s}$. Let us denote for a distribution R over $\mathscr{X}$

$$V_r(R) = V_r(R, W, P) = r H_P(R) - H_Q(T^*R) . \tag{5.5}$$

With this notation, $\underline{s}$ is the minimum of those r satisfying $V_r(R) \leqq 0$ for all R. It is enough to show that

$$\max_R \exp[r^{-1} V_r(R)] \leqq \max_{g \in \mathscr{F}_+^1} G_r(g) \tag{5.6}$$

holds for all r $(0 < r < 1)$.

For every distribution R on $\mathscr{X}$ let us define g_R by

$$g_R(y) = (T^*R)(y)/Q(y)\,. \tag{5.7}$$

We have

$$\begin{aligned} G_r(g_R) &= E\exp[r^{-1}(T\log g_R)(X)] \\ &= \textstyle\sum_x R(x)\exp\left[r^{-1}(T\log g_R)(x) - \log\frac{R(x)}{P(x)}\right]. \end{aligned} \tag{5.8}$$

The last expression is, by the convexity of $\exp t$, larger than

$$\exp\left[r^{-1}\textstyle\sum_x R(x)(T\log g_R)(x) - \sum_x R(x)\log\frac{R(x)}{P(x)}\right] = \exp[r^{-1}V_r(R)]\,.$$

This proves (5.6) and hence $s \geqq \underline{s}$.

The next step is to show that $\underline{s} \geqq \rho^2$. In order to do this one has to differentiate $V_r(R)$ twice partially and establish that if it has a local maximum at $R = P$ then $r \geqq \rho^2$. This is rather straightforward and we write down the derivatives of $V_r(R)$ only for later purposes. We have

$$\frac{\partial V_r}{\partial R(x)}(R) = (1-r) + \left(T\log\frac{T^*R}{Q}\right)(x) - r\log\frac{R(x)}{P(x)}\,, \tag{5.9}$$

$$\frac{\partial^2 V_r}{\partial R(x_0)\,\partial R(x_1)}(R) = \textstyle\sum_y w(y\,|\,x_0)w(y\,|\,x_1)\frac{1}{(T^*R)(y)} - \frac{r\cdot\delta_{x_0x_1}}{R(x_0)}\,. \tag{5.10}$$

We refer to [6] for the proof of the fact that the quadratic form with coefficients as in (5.10) is negative definite in the space of functions $\{f\in\mathscr{F}(\mathscr{X})\,|\,\sum_x f(x)=0\}$ iff $r > \rho^2(W, R)$.

Now we show that

$$\max_R \exp[s^{-1}V_s(R)] = \max_{g\in\mathscr{F}_+^1} G_s(g) = 1\,.$$

Suppose first that for every $g \neq g_0$, $G_s(g) < 1$. Then $\underline{s} = s = \rho^2$. Indeed, for each $r > \rho^2$ in some neighbourhood U of g_0, $G_r(g) \leqq 1$ holds. If $G_s(g) < 1$ everywhere outside U and if $s > r$ then for some r' with $r < r' < s$, $G_{r'}(g) \leqq 1$ for all $g\in\mathscr{F}_+^1$. This contradicts the minimality of s. Suppose now that there is a $g_1\in\mathscr{F}_+^1$, $g_1 \neq g_0$ with $G_s(g_1) = 1$. We define the distribution R_1 by

$$R_1(x) = P(x)\textstyle\prod_y g_1(y)^{w(y|x)/s}\,. \tag{5.11}$$

Clearly $\sum_x R_1(x) = 1$. We shall show that $g_1 = g_{R_1}$. This implies that all members of the weighted sum in (5.8) are equal and hence

$$1 = G_s(g_{R_1}) = \exp[s^{-1}V_s(R_1)]\,, \quad V_s(R_1) = 0\,.$$

s is then the minimum of those r satisfying $V_r(R_1) = 0$ since $g_1 \neq g_0$ implies $R_1 \neq P$, $H_P(R_1) \neq 0$. This will complete the proof of $s = \underline{s}$ and, by the way, also of Theorem 7.

Let us denote

$$\mathscr{B} = \{y\in\mathscr{Y}\,|\,g_1(y) > 0\}\,, \qquad \mathscr{F}_{\mathscr{B}} = \{g\in\mathscr{F}_+^1\,|\,g(y) = 0 \text{ for all } y\notin\mathscr{B}\}\,.$$

We prove that for $y_0 \notin \mathscr{B}$, $(T^*R_1)(y_0) = 0$. By definition one has

$$(T^*R_1)(y_0) = \sum_x P(x)w(y_0|x) \prod_y g_1(y)^{w(y|x)/s} .$$

Suppose that $w(y_0|x_0) \neq 0$. Then

$$\prod_y g_1(y)^{w(y|x_0)/s} = 0$$

holds since this expression has a factor

$$g_1(y_0)^{w(y_0|x_0)/s} .$$

Hence every term in the above sum is 0. The function $G_s(g)$ has a maximum for $g = g_1$ under the condition $g \in \mathscr{T}_{\mathscr{B}}$. By Lagrange's multiplier theorem it follows that there is a μ such that for all $y \in \mathscr{B}$,

$$\frac{\partial G_s}{\partial g(y)}(g_1) - \mu Q(y) = 0$$

holds. This can be written, because of (5.2), as follows: for all $y \in \mathscr{B}$,

$$\mu g_1(y)Q(y) = s^{-1}(T^*R_1)(y) . \tag{5.12}$$

If $y \notin \mathscr{B}$ then the left-hand side is 0 and as we just showed then also $(T^*R_1)(y) = 0$. Thus (5.12) is true for all y. Let us sum up to determine μ

$$\mu \sum_y Q(y)g_1(y) = s^{-1} \sum_y (T^*R_1)(y) ,$$

hence $\mu = s^{-1}$. Then from (5.12) we have $g_1 = g_{R_1}$. □

We do not prove (b) of Theorem 5 here. It is rather elementary: one has to do the estimates in the proof of $s^* = s$ more carefully. Especially, one needs an appropriate speed of convergence in Fact 1.

B. *Proof of Theorem* 8. Clearly, $\tilde{\rho}^2(W) \leqq \max_P s(W, P)$. On the other hand, put $r = \tilde{\rho}(W)$ and choose P_0 with $r = \rho^2(W, P_0)$. Then—as it was shown in (5.10) and the text thereafter—the function $V_r(R, W, P_0)$ of R is concave when R runs over all possible distributions. It has a local maximum at $R = P_0$, which is also a global maximum. Thus $V_r(R, W, P_0) \leqq 0$ for all R, hence $r \geqq s(W, P_0)$. □

6. Pairs of binary random variables. Proof of Theorem 9.

(a) Let us compute $\rho^2(W_{\alpha\alpha}, P)$ for an arbitrary input distribution P. As was shown in [6], it is the value of a certain determinant:

$$\rho^2(W_{\alpha\alpha}, P) = P(0)P(1)Q(0)^{-1}Q(1)^{-1}(1 - 2\alpha)^2 . \tag{6.1}$$

By Theorem 8 we are done if we show that

$$(1 - 2\alpha)^2 = \rho^2(W_{\alpha\alpha}, P_{\alpha\alpha}) = \tilde{\rho}^2(W_{\alpha\alpha})$$

i.e., $\rho(W_{\alpha\alpha}, P)$ is maximal for $P = P_{\alpha\alpha}$. In our case this is equivalent to

$$P(0)P(1) \leqq Q(0)Q(1) .$$

The last inequality follows from the fact that $Q(0)$ is closer to $\frac{1}{2}$ than $P(0)$.

(b) Let us define an arbitrary distribution by

$$R_t(0) = P_{\alpha\beta}(0) + t$$

and denote $F(t) = H_{P_{\alpha\beta}}(R_t)$. An easy computation shows that

$$H_{P_{\alpha\beta}}(T^* R_t) = F(at)$$
$$F(t) = c_0 t^2 + c_1 t^3 + \sigma(t^3)$$

where $a = 1 - \alpha - \beta$, $c_0, c_1 \neq 0$.

We have to show that

$$\sup_t \frac{F(at)}{F(t)} \neq \lim_{t \to 0} \frac{F(at)}{F(t)} = \rho^2(W_{\alpha\beta}, P_{\alpha\beta}) .$$

It is easily seen that this limit equals a^2. Now

$$\frac{F(at)}{F(t)} = a^2 \frac{c_0 + c_1 at + o(t)}{c_0 + c_1 t + o(t)} .$$

For some t with $c_1 t / c_0 < 0$ this expression is clearly larger than a^2. □

REFERENCES

[1] AHLSWEDE, R., GÁCS, P. and KÖRNER, J. (1976). Bounds on conditional probabilities with applications in multi-user communication. *Z. Wahrscheinlichkeitstheorie und Verw. Gebiete* **34** 157–177.

[2] CSÁKI, P. and FISCHER, J. (1963). On the general notion of maximal correlation. *Publ. Inst. Math. Hung. Acad. Sci.* **8** 27–51.

[3] GÁCS, P. and KÖRNER, J. (1973). Common information is far less than mutual information. *Problems of Control and Information Theory* **2** 149–162.

[4] KULLBACK, S. (1958). *Information Theory and Statistics.* Wiley, New York.

[5] NELSON, E. (1973). The free Markoff field. *J. Functional Anal.* **12** 211–227.

[6] WITSENHAUSEN, H. S. (1975). On sequences of pairs of dependent random variables. *SIAM J. Appl. Math.* **28** 100–113.

DEPT. OF MATHEMATICS
THE OHIO STATE UNIVERSITY
231 W. 18TH AVENUE
COLUMBUS, OHIO 43210

MATHEMATICAL INSTITUTE OF
THE HUNGARIAN ACADEMY OF SCIENCES
H-1053 BUDAPEST, REÁLTANODA U. 13-15