SIAM J. APPL. MATH.
Vol. 28, No. 1, January 1975

ON SEQUENCES OF PAIRS OF DEPENDENT RANDOM VARIABLES*

H. S. WITSENHAUSEN†

Abstract. The generalized random variables (x, y) have a given joint distribution. Pairs (x_i, y_i) are drawn independently. The observer of $(x_1, \cdots, x_n)$ and the observer of $(y_1, \cdots, y_n)$ each make a binary decision, of entropy bounded away from zero, with probability of disagreement ε_n. It is shown that ε_n can be made to approach zero as $n \to \infty$ if and only if the maximum correlation of x and y is unity. Under a compactness condition, satisfied in particular when x and/or y takes only finitely many values, this occurs if and only if the joint distribution decomposes, that is when ε_1 can be made to vanish by nontrivial decisions, as had been conjectured.

Results are also obtained for nonidentically distributed pairs, for randomized decisions, for multivalued decisions and for decisions based on the infinite sequences.

The question arose in the transmission of data from two dependent sources to two receivers. The results of Gács and Körner [1] for that problem are sharpened and clarified.

Introduction. If u, v, w are independent random variables, an observer of the pair $x = (u, v)$ and an observer of the pair $y = (u, w)$ can agree, with probability one, as to the value of u, which is sometimes called the "common core" of x and y. If x and y are the outputs of two sources at the same location and are to be transmitted to two receivers at another location, the capacity required is the sum $H(u) + H(v) + H(w)$ of the entropies of u, v and w if a common channel can be provided for the transmission of u, and separate channels for the transmission of v and w. This represents a saving in capacity of $H(u)$ when compared with the transmission of x and y through separate channels not taking advantage of their dependence.

Motivated by this observation, information theorists are presently considering a number of ways of defining a notion of "common core" of two dependent variables.[1] One of the questions arising in this connection is the following. Given the joint distribution of two random variables x and y, with values in finite alphabets X and Y:

(a) Can the observer of x and the observer of y each make a statement about the pair (x, y) which is not trivial (i.e., not constant with probability one) with probability of agreement one?

(b) If this is not the case, suppose n pairs (x_i, y_i), $i = 1, \cdots, n$, have been drawn independently under the given distribution, and that the observer of $(x_1, \cdots, x_n)$ and the observer of $(y_1, \cdots, y_n)$ each make a binary (yes or no) statement about these sequences. Let p_n and q_n be the probabilities of "yes" and ε_n the probability of disagreement. Is it possible that $\lim_{n\to\infty} \varepsilon_n = 0$ while p_n and q_n remain bounded away from 0 and 1?

* Received by the editors July 5, 1973.

† Bell Laboratories, Murray Hill, New Jersey 07974.

[1] A. Wyner [8] has shown that a meaningful measure of common information is the infimum of the mutual information between (x, y) and w, where w is any random variable such that x and y are conditionally independent given w.

The problem was stated in this form by D. Slepian, who observed that the answer to (a) is positive if and only if X and Y can each be decomposed into $k \geqq 2$ disjoint sets of positive probability $X_1, \cdots, X_k$; $Y_1, \cdots, Y_k$ such that the probability of $\bigcup_{i=1}^{k} X_i \times Y_i$ is one, which permits agreement on the index i by the two observers. Meanwhile Gács and Körner [1] had obtained results implying a negative answer to a weakened form of (b), when such a decomposition does not occur (see the corollary to Theorem 5 below). Slepian conjectured that when there is no decomposition the answer to (b) is negative, which is proven here (Theorem 4).

The case of general random variables x, y is also considered, and it is shown that a negative answer to (a) implies a negative answer to (b) under a compactness condition. Lower and upper bounds on the minimum error probability are obtained.

1. Problem statement. Let $(X, \mathscr{X})$, $(Y, \mathscr{Y})$ be two measurable spaces and let $(X \times Y, \mathscr{X} \times \mathscr{Y}, \mu)$ be a probability space. Assume that this space is separable, that is, that $\mathscr{X} \times \mathscr{Y}$ contains a countably generated σ-field the μ-completion of which contains $\mathscr{X} \times \mathscr{Y}$. The marginal probabilities induced by μ on $(X, \mathscr{X})$ (resp. $(Y, \mathscr{Y})$) are denoted by ξ (resp. η).

We say that μ *decomposes* when there are sets $A \in \mathscr{X}$, $B \in \mathscr{Y}$ such that $\xi(A)$, $\xi(X - A), \eta(B), \eta(Y - B)$ are positive while $\mu(A \times (Y - B)) = \mu((X - A) \times B) = 0$. Note that this implies

$$\mu(A \times B) = \xi(A) = \eta(B), \quad \mu((X - A) \times (Y - B)) = \xi(X - A) = \eta(Y - B).$$

Let $\sigma \in (0, 1/2)$ and for any fixed n, let n independent drawings of pairs $(x_1, y_1), \cdots, (x_n, y_n)$ be made under the distribution μ for each drawing. Consider the set of all pairs of measurable Boolean functions (φ, ψ), $\varphi : X^n \to \{0, 1\}$, $\psi : Y^n \to \{0, 1\}$ satisfying $\sigma \leqq \Pr\{\varphi = 1\} \leqq 1 - \sigma$, $\sigma \leqq \Pr\{\psi = 1\} \leqq 1 - \sigma$. Let $\varepsilon_n^*(\sigma)$ be the infimum over this set of $\Pr\{\varphi \neq \psi\}$.[2]

For fixed σ, $\varepsilon_n^*(\sigma)$ is nonincreasing in n, as the functions admissible for n remain admissible for $n + 1$. Let $\varepsilon_\infty^*(\sigma) = \inf_n \varepsilon_n^*(\sigma)$. The Slepian conjecture is that if X and Y are finite sets and μ does not decompose, then $\varepsilon_\infty^*(\sigma) > 0$.

In this paper it is shown that $\varepsilon_\infty^*(\sigma) = 0$ if and only if the Hirschfeld–Gebelein–Rényi maximal correlation $S(X, Y, \mu) = 1$, which is the case when μ decomposes but only implies decomposition under a compactness condition. Since this compactness condition holds in particular when at least one of the sets X, Y is finite, the Slepian conjecture is correct.

2. The maximum correlation. The maximum correlation [2], [3], [5] associated with $(X \times Y, \mathscr{X} \times \mathscr{Y}, \mu)$ is the number

$$S(X, Y, \mu) = \sup E\{f(x)g(y)\}$$

where the supremum is over all pairs of real-valued measurable functions $f(x)$, $g(y)$ sbtisfying $E\{f\} = E\{g\} = 0$, $E\{f^2\} = E\{g^2\} = 1$. It has the following geometric and operator-theoretic interpretation.

---

[2] If there is no pair (φ, ψ) satisfying the constraints, set $\varepsilon_n^*(\sigma) = 1$.

Let H be the separable real Hilbert space $L_2(X \times Y, \mathscr{X} \times \mathscr{Y}, \mu)$. Let H_x (resp. H_y) be the subspace of H consisting of the equivalence classes which contain a function measurable on the cylindrical extension of $\mathscr{X}$ (resp. $\mathscr{Y}$) to $X \times Y$, in short, the functions of x (resp. y) only. Let γ be the one-dimensional subspace of H consisting of the a.s. constant functions and let Z be its orthogonal complement: the functions of zero mean. Note that $\gamma \subset H_x \cap H_y$. Let $H_x^0 = H_x \cap Z$ and $H_y^0 = H_y \cap Z$. Let $\theta \in [0, \pi/2]$ be the angle between H_x^0 and H_y^0, that is, the infimum of the angle between lines in the two subspaces. Then

$$\cos\theta = S(X, Y, \mu). \tag{1}$$

Now let π_x (resp. π_y) be the orthogonal projection of H onto H_x (resp. H_y). These are the operators of conditional expectation with respect to x (resp. y).

Let $P_x : H_y \to H_x$ be the restriction of π_x to domain H_y and codomain H_x; likewise let $P_y : H_x \to H_y$ be the restriction of π_y. By the orthogonality of the projection, P_x and P_y are mutually adjoint, i.e., for $f \in H_x$, $g \in H_y$,

$$\langle f, P_x g\rangle = \langle P_y f, g\rangle = \langle f, g\rangle. \tag{2}$$

H_x (resp. H_y) is the orthogonal direct sum of γ and H_x^0 (resp. H_y^0) and P_x (resp. P_y) leaves γ fixed and maps H_y^0 (resp. H_x^0) into H_x^0 (resp. H_y^0). Thus one has the orthogonal decompositions

$$P_x = P_x^0 \oplus I_\gamma, \qquad P_y = P_y^0 \oplus I_\gamma, \tag{3}$$

where I_γ is the identity on γ, while $P_x^0 : H_y^0 \to H_x^0$; $P_y^0 : H_x^0 \to H_y^0$ are the restrictions of P_x and P_y. Again P_x^0 and P_y^0 are mutually adjoint. For $f \in H_x^0$, $f \neq 0$, the ratio $\|P_y^0 f\|/\|f\|$ is the cosine of the angle between f and H_y^0. Hence

$$\|P_y^0\| = \|P_x^0\| = \cos\theta. \tag{4}$$

By the adjointness,

$$\|P_x^0 P_y^0\| = \|P_y^0 P_x^0\| = \|P_x^0\|^2 = \|P_y^0\|^2 = \cos^2\theta, \tag{5}$$

where $P_x^0 P_y^0$ (resp. $P_y^0 P_x^0$) is a nonnegative definite self-adjoint operator on H_x^0 (resp. H_y^0).

Note that for the nonnegative definite self-adjoint operators of unit norm $P_x P_y$ on H_x and $P_y P_x$ on H_y, one has the orthogonal decompositions

$$P_x P_y = P_x^0 P_y^0 \oplus I_\gamma, \qquad P_y P_x = P_y^0 P_x^0 \oplus I_\gamma. \tag{6}$$

LEMMA 1. *The subspaces H_x^0 and H_y^0 have a line in common if and only if μ decomposes.*

Proof. If μ is decomposed by the sets $A \subset X$, $B \subset Y$, then one has $\xi(A) = \eta(B) = p \in (0, 1)$. Let $f(x) = 1 - p$ for $x \in A$ and $-p$ otherwise, $g(y) = 1 - p$ for $y \in B$, $-p$ otherwise. Then f and g have zero mean, do not vanish a.s. and one has $f = g$ a.s., that is, the line $\{\lambda f | \lambda \in R\}$ belongs to both H_x^0 and H_y^0. Conversely, if H_x^0 and H_y^0 have a line in common, there is a function φ such that $E\{\varphi\} = 0$, $E\{\varphi^2\} = 1$ and

$$\varphi(x, y) = f(x) = g(y) \quad \text{a.s.}$$

for some measurable functions f, g. Then the values of f and g have the same distribution with unit variance. Hence there exists $a \in R$ such that the sets $A = \{x | f(x) > a\}$ and $B = \{y | g(y) > a\}$ have marginal probability $p \in (0, 1)$. But $|f(x) - g(y)| > 0$ on $A \times (Y - B) \cup (X - A) \times B$ which must therefore have zero measure, that is A, B furnish a decomposition of μ.

3. The compact case. If any one of the operators $P_x P_y$, $P_y P_x$, $P_x^0 P_y^0$, $P_y^0 P_x^0$ is compact, then all four are compact. This will be called the compact case.

If one of the sets X, Y, say X, is finite, then $P_x P_y$ is an operator on the finite-dimensional space H_x, hence compact.

If μ is absolutely continuous with respect to $\xi \times \eta$ with Radon–Nikodym derivative $d(x, y)$, then $g = P_y f$ can be represented by

$$g(y) = \int_X \xi(dx) d(x, y) f(x) \tag{7}$$

and $f = P_x g$ by

$$f(x) = \int_Y \eta(dy) d(x, y) g(y) \tag{8}$$

and $\hat{f} = P_x P_y f$ by

$$\hat{f}(x') = \int_{X \times Y} \xi(dx) \eta(dy) d(x', y) d(x, y) f(x). \tag{9}$$

Then, letting $\| \cdot \|_{\mathrm{HS}}$ denote the Hilbert–Schmidt norm, one has

$$\begin{aligned} \|P_x\|_{\mathrm{HS}}^2 = \|P_y\|_{\mathrm{HS}}^2 &= \operatorname{tr} P_x P_y = \operatorname{tr} P_y P_x \\ &= \int_{X \times Y} \xi(dx) \eta(dy) d^2(x, y) \\ &= \int_{X \times Y} \mu(dx\, dy) d(x, y) = E\{d\}. \end{aligned} \tag{10}$$

Note that, by concavity of the logarithm,

$$\log \operatorname{tr} P_x P_y = \log E\{d\} \geqq E\{\log d\} = I_{XY}, \tag{11}$$

where I_{XY} is the mutual information. If $\operatorname{tr} P_x P_y < \infty$, $P_x P_y$ is compact. This condition is satisfied in particular for $X = Y = R$, with μ a nondegenerate Gaussian distribution. A self-adjoint bounded operator is compact as soon as one of its powers is compact and a fortiori when the trace of a power is finite. One has

$$\begin{aligned} &\operatorname{tr} (P_x P_y)^n \\ &\quad = \int \mu(dx_1\, dy_1) \cdots \mu(dx_n\, dy_n) d(x_1, y_n) d(x_n, y_{n-1}) \cdots d(x_3, y_2) d(x_2, y_1). \end{aligned} \tag{12}$$

For a compact self-adjoint operator such as $P_x P_y$ the spectrum consists of real isolated eigenvalues of finite multiplicity with the exception of 0, with an associated orthonormal system of eigenvectors. As $\|P_x P_y\| = 1$ and $P_x P_y \geqq 0$ the

spectrum is contained in $[0, 1]$. The eigenvalues, repeated according to multiplicity, may be arranged in a nonincreasing sequence

$$1 \geqq \lambda_2 \geqq \lambda_3 \geqq \cdots,$$

where the first 1 corresponds to the constant functions and the other eigenvalues are those of $P_x^0 P_y^0$. Then

$$\cos^2 \theta = \|P_x^0 P_y^0\| = \lambda_2 \tag{13}$$

and the maximum correlation is attained [6].

LEMMA 2. *In the compact case, μ decomposes if and only if $\theta = 0$.*

Proof. If μ decomposes, then H_x^0 and H_y^0 have, by Lemma 1, a line in common, so that $\theta = 0$. Conversely, assume $\theta = 0$. Then by (13), $\lambda_2 = 1$. Let f be a normalized eigenvector of $P_x^0 P_y^0$ for this eigenvalue and let $g = P_y^0 f$. Then $P_x^0 g = P_x^0 P_y^0 f = f$. Hence $f - g$ is orthogonal to both f and g, implying $f = g$. Then H_x^0 and H_y^0 have f in common and by Lemma 1, μ decomposes.

In the absence of compactness θ may be zero as an infimum that is not attained, in which case H_x^0 and H_y^0 have no line in common and μ does not decompose. The following is an example of this situation [6]. Let $X = Y = [0, 1]$ with Borel sets. Let H be the region in $X \times Y$ defined by $\{x, y | (x - y)^2 \leqq c(x + y)^4\}$ for some $c > 0$. Let h be the Lebesgue measure $\lambda(H)$ of H and let $d\mu = h^{-1} \chi_H \, d\lambda$, i.e., uniform density on H with respect to λ. Then suitable functions $f(x)$, $g(y)$ supported on neighborhoods of the origin make arbitrarily small angles while μ does not decompose.

4. The finite case. When both X and Y have finite cardinalities, say m_1 and m_2, then μ is characterized by the matrix $M = [\mu_{ij}]$ of joint probabilities. Without loss of generality one may assume that the marginals

$$\xi_i = \sum_{j=1}^{m_2} \mu_{ij}, \quad \eta_j = \sum_{i=1}^{m_1} \mu_{ij}$$

are all strictly positive.

Then P_x can be represented by the $m_1 \times m_2$ stochastic matrix T^x with $T^x_{ij} = \mu_{ij}/\xi_i$ while P_y can be represented by the $m_2 \times m_1$ matrix Y^y with $T^y_{ji} = \mu_{ij}/\eta_j$. Then $P_x P_y$ is represented by $T^x T^y$, an $m_1 \times m_1$ stochastic matrix with

$$(T^x T^y)_{ij} = \sum_{k=1}^{m_2} \frac{\mu_{ik}\mu_{jk}}{\xi_i \eta_k}, \tag{14}$$

and similarly for $T^y T^x$ whose spectrum can differ from that of $T^x T^y$ only in the multiplicity of zero. That the self-adjoint operator $P^x P^y$ is represented by a not necessarily symmetric matrix is due to the fact that the natural basis in R^{m_1} is orthogonal but not orthonormal under the inner product $\langle x, y \rangle = \sum_{i=1}^{m_1} \xi_i x_i y_i$. If the basis is scaled by $\sqrt{\xi_i}$ to achieve orthonormality, the representation of $P^x P^y$ becomes symmetric, with entries

$$\sum_{k=1}^{m_2} \frac{\mu_{ik}\mu_{jk}}{\eta_k (\xi_i \xi_j)^{1/2}}, \tag{15}$$

and this matrix is nonnegative definite.

By (13),

$$\cos\theta = [\lambda_2(T^xT^y)]^{1/2} = [\lambda_2(T^yT^x)]^{1/2}, \tag{16}$$

where $\lambda_2(\cdot)$ denotes the second largest eigenvalue, taking account of multiplicity. When $m_1 = m_2 = 2$, the second eigenvalue is the determinant as the largest eigenvalue is 1, so that $\cos\theta$ is the geometric mean of the magnitudes of the determinants of T^x and T^y. When $m_1 = m_2$ and μ is symmetric ($\mu_{ij} = \mu_{ji}$), then $T^x = T^y$ and in the orthonormal basis T^x and T^y are symmetric although not necessarily nonnegative definite. In that case $\cos\theta$ is the second largest *magnitude* eigenvalue of T^x.

In any case, using the natural basis of R^{m_1}, T^xT^y has the right eigenvector $v = \operatorname{col}(1, 1, \cdots, 1)$ and the left eigenvector $\xi = (\xi_1, \cdots, \xi_{m_1})$ for the eigenvalue 1, corresponding to the action of P_xP_y on the constant functions. Thus $P_x^0P_y^0$ is represented by the deflation $T^xT^y - v\xi$ which will have spectral radius $\cos^2\theta$.

5. Independent sequences of pairs of random variables. For $i = 1, \cdots, n$ let $(X_i \times Y_i, \mathscr{X}_i \times \mathscr{Y}_i, \mu_i)$ be separable probability spaces. Let

$$X = \prod_{i=1}^{n} X_i, \quad Y = \prod_{i=1}^{n} Y_i, \quad \mathscr{X} = \prod_{i=1}^{n} \mathscr{X}_i, \quad \mathscr{Y} = \prod_{i=1}^{n} \mathscr{Y}_i, \quad \mu = \prod_{i=1}^{n} \mu_i.$$

Then $(X \times Y, \mathscr{X} \times \mathscr{Y}, \mu)$ is the separable probability space corresponding to independent drawings of the pairs $(x_1, y_1), \cdots, (x_n, y_n)$ under their respective distributions. The maximum correlation $S(X, Y, \mu)$ between $(x_1, \cdots, x_n)$ and $(y_1, \cdots, y_n)$ is given by the following result.[3]

THEOREM 1. $S(X, Y, \mu) = \max_{1 \le i \le n} S(X_i, Y_i, \mu_i)$, *that is, the angle* θ *determined by* μ *is the minimum of the angles* θ_i *determined by the* μ_i.

Proof. By induction, it suffices to consider the case $n = 2$. The space $H_{12} = L_2(X \times Y, \mathscr{X} \times \mathscr{Y}, \mu)$ is the tensor product[4] of the spaces

$$H_1 = L_2(X_1 \times Y_1, \mathscr{X}_1 \times \mathscr{Y}_1, \mu_1) \quad \text{and} \quad H_2 = L_2(X_2 \times Y_2, \mathscr{X}_2 \times \mathscr{Y}_2, \mu_2).$$

Likewise,

$$\begin{aligned} H_{x_1x_2} = L_2(X, \mathscr{X}, \xi) &= L_2(X_1, \mathscr{X}_1, \xi_1) \otimes L_2(X_2, \mathscr{X}_2, \xi_2) \\ &= H_{x_1} \otimes H_{x_2} \end{aligned} \tag{17}$$

and $H_{y_1y_2} = H_{y_1} \otimes H_{y_2}$.

The operator $P_{x_1x_2}: H_{y_1y_2} \to H_{x_1x_2}$ of orthogonal projection, when applied to finite sums of products of functions of y_1 and y_2, satisfies

$$\begin{aligned} P_{x_1x_2} \sum_k a_k(y_1)b_k(y_2) &= E\Big\{\sum_k a_k(y_1)b_k(y_2) | x_1, x_2\Big\} \\ &= \sum_k E\{a_k(y_1)|x_1\}E\{b_k(y_2)|x_2\} \\ &= \sum_k (P_{x_1}a_k)(P_{x_2}b_k) \end{aligned} \tag{18}$$

[3] This result was stated as Theorem 6.2 in [9]. However the proof given in [9] is incorrect. It hinges on a chain of inequalities of which the last actually holds in the direction opposite to the one asserted.

[4] This is shown in [4, pp. 51–52] under the assumption of separability. Note that the results concerning a single drawing are valid regardless of separability.

by the independence of (x_1, y_1) and (x_2, y_2). By continuity,

$$P_{x_1x_2} = P_{x_1} \otimes P_{x_2}. \tag{19}$$

However, applying (3) in the product space, one has the orthogonal decomposition

$$P_{x_1x_2} = P^0_{x_1x_2} \oplus I_{\gamma_{12}}; \tag{20}$$

while by (3) applied in the factor spaces and (19),

$$\begin{aligned} P_{x_1x_2} &= (P^0_{x_1} \oplus I_{\gamma_1}) \otimes (P^0_{x_2} \oplus I_{\gamma_2}) \\ &= (P^0_{x_1} \otimes P^0_{x_2}) \oplus (P^0_{x_1} \otimes I_{\gamma_2}) \oplus (I_{\gamma_1} \otimes P^0_{x_2}) \oplus (I_{\gamma_1} \otimes I_{\gamma_2}). \end{aligned} \tag{21}$$

This is an orthogonal decomposition of the operator, associated with the orthogonal decompositions

$$\begin{aligned} H_{x_1x_2} &= (H^0_{x_1} \oplus \gamma_1) \otimes (H^0_{x_2} \oplus \gamma_2) \\ &= (H^0_{x_1} \otimes H^0_{x_2}) \oplus (H^0_{y_1} \otimes \gamma_2) \oplus (\gamma_1 \otimes H^0_{x_2}) \oplus (\gamma_1 \otimes \gamma_2) \end{aligned}$$

and

$$H_{y_1y_2} = (H^0_{y_1} \otimes H^0_{y_2}) \oplus (H^0_{y_1} \otimes \gamma_2) \oplus (\gamma_1 \otimes H^0_{y_2}) \oplus (\gamma_1 \otimes \gamma_2) \tag{22}$$

of the domain and codomain.

In (22) the first term is the subspace of functions $g(y_1, y_2)$ satisfying $E\{g(y_1, y_2)|y_1\} = 0$ a.s. and $E\{g(y_1, y_2)|y_2\} = 0$ a.s., the second and third terms are the subspaces of functions of zero mean dependent only on y_1, respectively y_2, and the last term is the line of constant functions. Any function in $H_{y_1y_2}$ is uniquely decomposed according to

$$\begin{aligned} g(y_1, y_2) = (g - E\{g|y_1\} - E\{g|y_2\} + E\{g\}) + (E\{g|y_1\} - E\{g\}) \\ + (E\{g|y_2\} - E\{g\}) + E\{g\} \end{aligned} \tag{23}$$

and likewise for $H_{x_1x_2}$.

Now as $I_{\gamma_1} \otimes I_{\gamma_2} = I_{\gamma_{12}}$ comparison of (20) and (21) yields

$$P^0_{x_1x_2} = (P^0_{x_1} \otimes P^0_{x_2}) \oplus (P^0_{x_1} \otimes I_{\gamma_2}) \oplus (I_{\gamma_1} \otimes P^0_{x_2}). \tag{24}$$

By orthogonality,

$$\|P^0_{x_1x_2}\| = \max(\|P^0_{x_1} \otimes P^0_{x_2}\|, \|P^0_{x_1} \otimes I_{\gamma_2}\|, \|I_{\gamma_1} \otimes P^0_{x_2}\|) \tag{25}$$

and since $\|A \otimes B\| = \|A\| \, \|B\|$ one has

$$\begin{aligned} \cos\theta = \|P^0_{x_1x_2}\| &= \max(\cos\theta_1 \cos\theta_2, \cos\theta_1, \cos\theta_2) \\ &= \max(\cos\theta_1, \cos\theta_2) \\ &= \cos\min(\theta_1, \theta_2). \end{aligned} \tag{26}$$

This completes the proof.

By definition of θ, the infimum Δ^*_n of $E\{(f - g)^2\}$ over all real functions $f(x_1, \cdots, x_n)$, $g(y_1, \cdots, y_n)$ of zero mean and unit variance is $2(1 - \cos\theta) = 2(1 - \cos\min_{1 \leq i \leq n} \theta_i)$ and $\Delta^*_\infty = \inf_n \Delta^*_n = 2(1 - \cos\inf_n \theta_n)$. For identically distributed drawings, Δ^*_n is independent of n, a conclusion which remains valid if some of the pairs are reversed, e.g., for $f(x_1, y_2, x_3, y_4, \cdots)$, $g(y_1, x_2, y_3, x_4, \cdots)$.

6. Bounds on the probability of error. For the problem described in the Introduction, lower and upper bounds on $\varepsilon_\infty^*(\sigma)$ will now be derived in terms of the angle θ.

For a single drawing from $(X \times Y, \mathscr{X} \times \mathscr{Y}, \mu)$, with $S(X, Y, \mu) = \cos\theta$, let $\varphi: X \to \{0, 1\}$, $\psi: Y \to \{0, 1\}$ be measurable functions and let $A \subset X$, $B \subset Y$ be the sets of which they are the characteristic functions. Set $q = \xi(A)$, $r = \eta(B)$, $a = \mu(A \times B)$, $b = \mu((X - A) \times B)$, $c = \mu(A \times (Y - B))$ and $d = \mu((X - A) \times (Y - B))$. Then one has

$$a + c = q, \quad b + d = 1 - q, \quad a + b = r, \quad c + d = 1 - r, \quad b + c = \varepsilon, \tag{27}$$

where ε is the probability of error.

Note that (27) implies

$$\varepsilon \geqq |q - r|. \tag{28}$$

THEOREM 2. *The probability of error is bounded from below by*

$$\varepsilon \geqq 2(1 - \cos\theta)[q(1 - q)r(1 - r)]^{1/2}.$$

Proof. Let $f(x) = 1 - q$ for $x \in A$, $-q$ for $x \notin A$ and $g(y) = 1 - r$ for $y \in B$, $-r$ for $y \notin B$. Then f and g have zero mean, so that

$$E\{f(x)g(y)\} \leqq \cos\theta E\{f^2(x)\}^{1/2} E\{g^2(y)\}^{1/2}$$

or

$$a(1 - q)(1 - r) - bq(1 - r) - c(1 - q)r + dqr \leqq \cos\theta[q(1 - q)r(1 - r)]^{1/2}.$$

Eliminating a, b, c, d in favor of q, r, ε by (27) yields

$$\varepsilon \geqq q(1 - r) + r(1 - q) - 2\cos\theta[q(1 - q)r(1 - r)]^{1/2},$$

which is sharp in case of independence ($\theta = \pi/2$), and may be written

$$\varepsilon \geqq (\sqrt{q(1 - r)} - \sqrt{r(1 - q)})^2 + 2(1 - \cos\theta)[q(1 - q)r(1 - r)]^{1/2},$$

which completes the proof.

Now consider a sequence $(X_i \times Y_i, \mathscr{X}_i \times \mathscr{Y}_i, \mu_i)$ and measurable Boolean functions

$$\varphi_n: \prod_{i=1}^{n} X_i \to \{0, 1\}, \quad \psi_n: \prod_{i=1}^{n} Y_i \to \{0, 1\}$$

with

$$q_n = \Pr\{\varphi_n = 1\}, \quad r_n = \Pr\{\psi_n = 1\}, \quad \varepsilon_n = \Pr\{\varphi_n \neq \psi_n\}$$

and $\cos\theta_i = S(X_i, Y_i, \mu_i)$.

Combining Theorems 1 and 2 one has the corollaries

$$\varepsilon_n \geqq 2(1 - \cos \min_{1 \leqq i \leqq n} \theta_i)[q_n(1 - q_n)r_n(1 - r_n)]^{1/2}. \tag{29}$$

For $q_n, r_n \in [\sigma, 1 - \sigma]$,

$$\varepsilon_n^*(\sigma) \geqq 2(1 - \cos \min_{1 \leqq i \leqq n} \theta_i)\sigma(1 - \sigma) \tag{30}$$

as well as

$$\varepsilon_\infty^*(\sigma) \geqq 2(1 - \cos \inf_{i \geqq 1} \theta_i)\sigma(1 - \sigma). \tag{31}$$

For the case of identically distributed drawings,

$$\varepsilon_\infty^*(\sigma) \geqq 2\sigma(1 - \sigma)(1 - \cos \theta), \tag{32}$$

where θ is the angle for a single drawing. Note that (32) holds for arbitrary μ_i provided their angles are all the same, and in particular if some of the pairs are reversed.

Note that while the bound (32) is independent of n, $\varepsilon_n^*(\sigma)$ will in general not be constant. For finite X, Y only a finite number of functions φ_n, ψ_n satisfy the σ constraint for each n. This number increases with n and so does the multiplicity of eigenvalue λ_2. As n increases it may become possible to find two-valued functions which are closer to the eigenspace for λ_2. If this eigenspace contains a two-valued function, it achieves the bound (32). Such is the case for $n = 1$ and $\sigma = 1/2$ and μ the 2×2 matrix

$$\begin{pmatrix} (1 - \delta)/2 & \delta/2 \\ \delta/2 & (1 - \delta)/2 \end{pmatrix}$$

for which $\varepsilon_n^*(1/2) = \delta$ for all n, a result obtained previously by A. Wyner using a lemma of Wyner and Ziv [7]. Examples of decreases in $\varepsilon_n^*(\sigma)$ with n were first obtained by R. L. Graham.

THEOREM 3. *For identically distributed pairs with angle θ one has, for all $\sigma \in (0, 1/2)$, $\varepsilon_\infty^*(\sigma) \leqq \theta/\pi$.*

Proof. By the definition of θ, there exist, for arbitrary $\delta > 0$, real functions $f(x_1)$, $g(y_1)$ of zero mean and unit variance for which $E\{f(x_1)g(y_1)\} = \cos \alpha$ with $\theta \leqq \alpha \leqq \min(\theta + \delta, \pi/2)$.

Let

$$f_n(x_1, \cdots, x_n) = n^{-1/2} \sum_{i=1}^{n} f(x_i)$$

and

$$g_n(y_1, \cdots, y_n) = n^{-1/2} \sum_{i=1}^{n} g(y_i).$$

Then f_n and g_n have zero mean and unit variance and $E\{f_n g_n\} = \cos \alpha$. Let $\varphi_n(x_1, \cdots, x_n) = 1$ for $f_n > 0$, 0 for $f_n \leqq 0$ and let $\psi_n(y_1, \cdots, y_n) = 1$ for $g_n > 0$, 0 for $g_n \leqq 0$.

By the central limit theorem the pair of random variables (f_n, g_n) converges in law to the pair (u, v), where u and v are real, jointly Gaussian random variables with mean zero and covariance matrix

$$\begin{pmatrix} 1 & \cos \alpha \\ \cos \alpha & 1 \end{pmatrix}.$$

Then $\Pr\{\varphi_n = 1\} \to \Pr\{u > 0\} = \frac{1}{2}$ and $\Pr\{\psi_n = 1\} \to \frac{1}{2}$. Hence for any $\sigma \in (0, 1/2)$ both probabilities fall inside $[\sigma, 1 - \sigma]$ for all sufficiently large n. On

the other hand, $\varepsilon_n = \Pr\{\varphi_n \neq \psi_n\}$ tends to $\Pr\{u > 0, v \leqq 0\} + \Pr\{u \leqq 0, v > 0\} = \alpha/\pi$. Hence $\varepsilon^*_\infty(\sigma) \leqq \alpha/\pi \leqq (\theta + \delta)/\pi$ and as δ is arbitrary the theorem is proved.

For sequences of pairs with nonidentical distributions μ_i with angles θ_i, the argument of Theorem 3 can be extended to give, for $0 < \sigma < 1/2$,

$$\varepsilon^*_\infty(\sigma) \leqq \pi^{-1} \liminf_{n\to\infty} \theta_n \geqq \pi^{-1} \inf_{n \geqq 1} \theta_n$$

and when $\inf_n \theta_n$ is not attained equality holds on the right.

Combining Theorems 2 and 3 gives the following result which, in particular, contains the Slepian conjecture.

THEOREM 4. *For identically distributed pairs with angle θ, the following three statements are equivalent*: (i) $\theta = 0$, (ii) $\varepsilon^*_\infty(\sigma) = 0$ *for some* $\sigma \in (0, 1/2)$, (iii) $\varepsilon^*_\infty(\sigma) = 0$ *for all* $\sigma \in (0, 1/2)$. *If the maximum correlation is attained (as in the compact case), these statements are in addition equivalent to* (iv) μ *decomposes.*

Allowing independent randomization of the decisions does not change the bounds obtained above. Indeed, such randomization amounts to the use of functions $f_n(u, x_1, \cdots, x_n)$, $g_n(v, y_1, \cdots, y_n)$, where u, v are independent random variables characterizing the randomizing devices. Then the pair (u, v) has angle $\pi/2$ and by Theorem 1 the angle between $(u, x_1, \cdots, x_n)$ and $(v, y_1, \cdots, y_n)$ is just the angle between $(x_1, \cdots, x_n)$ and $(y_1, \cdots, y_n)$.

7. Multivalued decisions. The condition $\sigma \leqq \Pr\{\varphi_n = 1\} \leqq 1 - \sigma$ is equivalent to a lower bound on the entropy of φ_n. When φ_n, ψ_n are allowed to take their values in a common finite set of cardinality k_n it is tempting to use the requirement that $H(\varphi_n)$, the entropy of φ_n, satisfy $H(\varphi_n) \geqq h_0 > 0$, to rule out the trivial constant functions. While this is adequate when k_n is independent of n, it will not be if k_n can grow with n. To see this consider two-valued functions $\varphi(x)$, $\psi(y)$ for which $\Pr\{\psi = \varphi = 0\} = 1 - 2\delta$, $\Pr\{\varphi = 0, \psi = 1\} = \Pr\{\varphi = 1, \psi = 0\} = \delta$ and $\Pr\{\psi = \varphi = 1\} = 0$. The probability of error is $\varepsilon = 2\delta$ and the entropy of φ is $h(\delta) = -\delta \log \delta - (1 - \delta) \log (1 - \delta)$. Now suppose that the sets $\{x|\varphi(x) = 1\}$ and $\{y|\psi(y) = 1\}$ can be divided into k parts of equal probability.[5] Let $\varphi_k(x)$ take the value 0 for $\varphi(x) = 0$ and the values $1, 2, \cdots, k$ on the k parts of the remainder of X, and likewise for ψ_k. Then the probability of error $\varepsilon = \Pr\{\varphi_k \neq \psi_k\} = 2\delta$ as before, while the entropy H of φ_k has become $h(\delta) + \delta \log k$. This can be made as large as desired by the choice of k. In particular, one can obtain $\varepsilon \to 0$, $H \to \infty$ in this entirely trivial way. It is simply not true that more entropy for the same probability of error means that more is being achieved, in fact, in this example error is committed with probability one in all the additional levels which furnish the extra entropy.

However, a lower bound on $H(\varphi)/\log k_n$ is one possibility that is adequate for the case of decisions with increasingly many values. For a sequence of independent pairs drawn from $(X \times Y, \mathscr{X} \times \mathscr{Y}, \mu)$, let $\varphi_n: X^n \to K_n = \{1, \cdots, k_n\}$, $\psi_n: Y^n \to K_n$ be measurable functions, and let $H(\varphi_n)$, $H(\psi_n)$ denote the entropies of the distributions generated by φ_n, ψ_n on K_n. Let $\varepsilon_n = \Pr\{\varphi_n \neq \psi_n\}$. Then one has the following theorem.

[5] This could also be achieved by (independent) randomization of the decisions.

THEOREM 5. *If* $\cos\theta = S(X, Y, \mu) < 1$ *and* $\lim_{n\to\infty} \varepsilon_n = 0$, *then*

$$\lim_{n\to\infty} H(\varphi_n)/\log k_n = \lim_{n\to\infty} H(\psi_n)/\log k_n = 0.$$

Proof. For each n let $A_n \subset K_n$ be a set for which the minimum of $|\Pr\{\varphi_n \in A_n\} - \Pr\{\varphi_n \notin A_n\}|$ subject to $\Pr\{\varphi_n \in A_n\} \leqq \Pr\{\varphi_n \notin A_n\}$ is attained. Let

$$q_n = \Pr\{\varphi_n \in A_n\}$$

noting $q_n \leqq 1/2$, and let $p_i = \Pr\{\varphi_n = i\}$, $i = 1, \cdots, k_n$. If $i \notin A_n$ and $p_i \neq 0$, then $p_i \geqq 1 - 2q_n$; for if $0 < p_i < 1 - 2q_n$, then either $A_n \cup \{i\}$ or its complement would furnish a lower minimum. Then

$$\begin{aligned} H(\varphi_n) &= -\sum_{i\notin A_n} p_i \log p_i - \sum_{i\in A_n} p_i \log p_i \\ &\leqq -\sum_{i\notin A_n} p_i \log(1 - 2q_n) - q_n \sum_{i\in A_n} \frac{p_i}{q_n} \log \frac{p_i}{q_n} - q_n \sum_{i\in A_n} \frac{p_i}{q_n} \log q_n \qquad (33) \\ &\leqq -(1 - q_n)\log(1 - 2q_n) + q_n \log(\operatorname{card} A_n) - q_n \log q_n \\ &\leqq h(q_n) + (1 - q_n)\log\frac{1 - q_n}{1 - 2q_n} + q_n \log k_n \end{aligned}$$

and likewise for $H(\psi_n)$ with set B_n of probability r_n.

Now let φ'_n, ψ'_n be the two-valued functions taking the value 1 on A_n (resp. B_n) and 0 otherwise. Let $\varepsilon'_n = \Pr\{\varphi'_n \neq \psi'_n\}$. Since $\varphi_n = \psi_n$ implies $\varphi'_n = \psi'_n$ one has $\varepsilon'_n \leqq \varepsilon_n$. Hence $\varepsilon'_n \to 0$ which by (28) implies $q_n - r_n \to 0$ and by (29), with $\theta > 0$, $q_n(1 - q_n)r_n(1 - r_n) \to 0$. By construction $q_n, r_n \in [0, 1/2]$. Hence $q_n \to 0$ and $r_n \to 0$. Then by (33), $H(\varphi_n)/\log k_n \to 0$ and likewise for $H(\psi_n)/\log k_n$, completing the proof.

In particular, one has the following corollary, obtained in a different way by Gács and Körner [1] for finite X, Y.

COROLLARY. *In the compact case, if* μ *does not decompose, if* $\varepsilon_n \to 0$ *and if* $\log k_n \leqq cn$, *then* $\lim_{n\to\infty} H(\varphi_n)/n = \lim_{n\to\infty} H(\psi_n)/n = 0$.

Note that the assumption $\log k_n \leqq cn$ is automatically fulfilled when X and Y are finite sets.

8. Functions of the full sequences. Consider functions f, g defined on the infinite sequences, $f: \prod_{i=1}^{\infty} X_i \to R$, $g = \prod_{i=1}^{\infty} Y_i \to R$.

THEOREM 6. *If* $f(x_1, \cdots) = g(y_1, \cdots)$ *a.s. and none of the* μ_i *decomposes, then* f *and* g *are degenerate* (*i.e., equal to a constant a.s.*).

Proof. Replacing f by $\tanh f$, g by $\tanh g$ if necessary, one may assume that f and g are bounded. Let

$$X^k = \prod_k^{\infty} X_i, \quad Y^k = \prod_k^{\infty} Y_i, \quad \mathscr{X}^k = \prod_k^{\infty} \mathscr{X}_i, \quad \mathscr{Y}^k = \prod_k^{\infty} \mathscr{Y}_i, \quad \mu^k = \prod_k^{\infty} \mu_i.$$

Then

$$\prod_k^{\infty} (X_i \times Y_i, \mathscr{X}_i \times \mathscr{Y}_i, \mu_i) = (X^k \times Y^k, \mathscr{X}^k \times \mathscr{Y}^k, \mu^k).$$

With $x^k = (x_k, x_{k+1}, \cdots)$, $y^k = (y_k, y_{k+1}, \cdots)$, let

$$f^k(x^k) = E\{f(x^1)|x^k\}, \qquad g^k(y^k) = E\{g(y^1)|y^k\} \qquad \text{a.s. } (\mu^k)$$

and note that

$$f^{k+1}(x^{k+1}) = E\{f^k(x^k)|x^{k+1}\}, \qquad g^{k+1}(y^{k+1}) = E\{g^k(y^k)|y^{k+1}\} \qquad \text{a.s. } (\mu^{k+1}).$$

The three relations

$$f(x^1) = f^k(x^k) \quad \text{a.s. } (\mu^1),$$
$$g(y^1) = g^k(y^k) \quad \text{a.s. } (\mu^1),$$
$$f^k(x^k) = g^k(y^k) \quad \text{a.s. } (\mu^k)$$

hold trivially for $k = 1$. Suppose they hold for $k \leqq n$. Then to show that they hold for $k = n + 1$ one need only establish the three relations $f^n(x^n) = f^{n+1}(x^{n+1})$ a.s. (μ^n), $g^n(y^n) = g^{n+1}(y^{n+1})$ a.s. (μ^n) and $f^{n+1}(x^{n+1}) = g^{n+1}(y^{n+1})$ a.s. (μ^{n+1}).

For $E \in \mathcal{X}^{n+1} \times \mathcal{Y}^{n+1}$, $x_n \in X_n$, $y_n \in Y_n$ let

$$a(x_n, E) = \int_E f^n(x_n, x^{n+1})\mu^{n+1}(dx^{n+1}\, dy^{n+1}),$$
$$b(y_n, E) = \int_E g^n(y_n, y^{n+1})\mu^{n+1}(dx^{n+1}\, dy^{n+1}).$$

Then $a(\cdot, E), b(\cdot, E)$ are bounded and measurable on $\mathcal{X}_n, \mathcal{Y}_n$. Let F be the set in $\mathcal{X}_n \times \mathcal{Y}_n$ on which $a(x_n, E) > b(y_n, E)$. One has

$$\int_F \mu_n(dx_n\, dy_n)(a(x_n, E) - b(y_n, E)) = \int_{F \times E} \mu^n(dx^n\, dy^n)(f^n(x^n) - g^n(y^n))$$

which vanishes by the induction hypothesis. Hence $\mu_n(F) = 0$ and likewise for the reverse inequality, so that, for all E,

$$a(x_n, E) = b(y_n, E) \qquad \text{a.s. } (\mu_n).$$

But since μ_n is indecomposable, one has by Lemma 1,

$$a(x_n, E) = b(y_n, E) = \nu(E) \qquad \text{a.s. } (\mu_n), \tag{34}$$

where $\nu(E)$ is a constant depending only on E.

Let

$$\psi(x^{n+1}) = \int_{X_n \times Y_n} \mu_n(dx_n\, dy_n) f^n(x_n, x^{n+1})$$
$$= E\{f^n(x^n)|x^{n+1}, y^{n+1}\}.$$

By (34),

$$\nu(E) = \int_{X_n \times Y_n} \mu_n(dx_n\, dy_n) a(x_n, E)$$
$$= \int_E \mu^{n+1}(dx^{n+1}\, dy^{n+1})\psi(x^{n+1}),$$

that is, v is a measure on $\mathcal{X}^{n+1} \times \mathcal{Y}^{n+1}$ with Radon–Nikodym derivative $dv/d\mu^{n+1} = \psi$.

One has, a.s. (μ^n),

$$\begin{aligned} f^{n+1}(x^{n+1}) &= E\{f^n(x^n)|x^{n+1}\} = E\{E\{f^n(x^n)|x^{n+1}, y^{n+1}\}|x^{n+1}\} \\ &= E\{\psi(x^{n+1})|x^{n+1}\} = \psi(x^{n+1}). \end{aligned} \tag{35}$$

For any set $F \in \mathcal{X}_n \times \mathcal{Y}_n$, by (34),

$$\int_{F \times E} \mu_n(dx_n\, dy_n)(v(E) - a(x_n, E)) = 0$$

or

$$\int_{F \times E} \mu^n(dx^n\, dy^n)(\psi(x^{n+1}) - f^n(x^n)) = 0.$$

As $F \times E$ ranges over all measurable rectangles, one has, using (35),

$$f^n(x^n) = \psi(x^{n+1}) = f^{n+1}(x^{n+1}) \qquad \text{a.s. } (\mu^n).$$

Likewise,

$$g^n(y^n) = g^{n+1}(y^{n+1}) \qquad \text{a.s. } (\mu^n).$$

Finally, since the set on which $f^{n+1} \neq g^{n+1}$ is a μ^n null set belonging to $\mathcal{X}^{n+1} \times \mathcal{Y}^{n+1}$, it is the cylindrical extension of a μ^{n+1} null set, i.e., $f^{n+1}(x^{n+1}) = g^{n+1}(y^{n+1})$ a.s. (μ^{n+1}), completing the induction.

Thus $f = f^n$ a.s. (μ^1) for all n, so that $f = \limsup f^n$ a.s. (μ^1). However, the function $\limsup f^n$ is measurable on the tail field of an independent sequence. By the Kolmogorov zero-one law, f, and likewise g, is degenerate as claimed.

COROLLARY. *A finite or countable product of indecomposable joint distributions μ_i is indecomposable.*

In particular, in the noncompact case, if each μ_i is indecomposable but has angle $\theta_i = 0$ (as an infimum which is not attained), the products have the same property. Furthermore, $E\{(f - g)^2\}$ which can already be made arbitrarily small by functions f, g of zero mean and unit variance dependent only on x_1, y_1 cannot be made to vanish even with f, g depending on the infinite sequences.

For the case of X, Y finite, the corollary contains the fact that Kronecker products of indecomposable matrices are indecomposable which is equivalent to the fact that anded products of connected bipartite graphs[6] are connected.

Acknowledgments. The author is indebted to R. L. Graham, J. Körner, S. P. Lloyd, J. Mazo, L. A. Shepp, D. Slepian and A. D. Wyner for fruitful discussions of this topic.

[6] That is, (a, a') is joined to (b, b') in the product graph if and only if a is joined to b *and* a' is joined to b' in the factor graphs.

REFERENCES

[1] P. GÁCS AND J. KÖRNER, *Common information is far less than mutual information*, Problems of Contr. and Inform. Th., 2 (1973), pp. 149–162.

[2] H. GEBELEIN, *Das statistische Problem der Korrelation als Variations—und Eigenwertproblem und sein Zusammenhang mit der Ausgleichungsrechnung*, Z. Angew. Math. Mech., 21 (1941), pp. 364–379.

[3] A. O. HIRSCHFELD, *A connection between correlation and contingency*, Proc. Cambridge Philos. Soc., 31 (1935), pp. 520–524.

[4] M. REED AND B. SIMON, *Methods of Modern Mathematical Physics, I: Functional Analysis*, Academic Press, New York, 1972.

[5] A. RÉNYI, *New version of the probabilistic generalization of the large sieve*, Acta Math. Hungar., 10 (1959), pp. 217–226.

[6] ———, *On measures of dependence*, Ibid., 10 (1959), pp. 441–451.

[7] A. D. WYNER, *A theorem on the entropy of certain binary sequences and applications (part II)*, IEEE Trans. Information Theory, IT-19 (1973), pp. 772–777.

[8] ———, *The common information of two dependent random variables*, Ibid., to appear.

[9] P. CSÁKI AND J. FISCHER, *On the general notion of maximal correlation*, Magyar Tud. Akad. Mat. Kutató Int. Közl., 8 (1963), pp. 27–51.