The Annals of Probability
1975, Vol. 3, No. 1, 146–158

I-DIVERGENCE GEOMETRY OF PROBABILITY DISTRIBUTIONS AND MINIMIZATION PROBLEMS

BY I. CSISZÁR

Mathematical Institute of the Hungarian Academy of Sciences

Some geometric properties of PD's are established, Kullback's I-divergence playing the role of squared Euclidean distance. The minimum discrimination information problem is viewed as that of projecting a PD onto a convex set of PD's and useful existence theorems for and characterizations of the minimizing PD are arrived at. A natural generalization of known iterative algorithms converging to the minimizing PD in special situations is given; even for those special cases, our convergence proof is more generally valid than those previously published. As corollaries of independent interest, generalizations of known results on the existence of PD's or nonnegative matrices of a certain form are obtained. The Lagrange multiplier technique is not used.

1. Introduction. Capital P, Q, R will denote PD's (probability distributions) on a measurable space $(X, \mathscr{X})$ which will not be mentioned in the sequel. If $P \ll Q$ (or $Q \ll R$, etc.) the corresponding density (Radon–Nikodym derivative) will be denoted by $p_Q(x)$ (or $q_R(x)$, etc.); the argument x will be omitted if this does not cause ambiguity.

The I-divergence or Kullback–Leibler information number $I(P\,||\,Q)$—also called information for discrimination, information gain or entropy of P relative to Q—is defined as

$$I(P\,||\,Q) = \int \log p_Q\, dP = \int p_Q \log p_Q\, dQ \qquad \text{if} \quad P \ll Q \tag{1.1}$$
$$= +\infty \qquad \text{if} \quad P \not\ll Q\,.$$

If R is any PD with $P \ll R$, $Q \ll R$ (1.1) may be equivalently written as

$$I(P\,||\,Q) = \int p_R \log \frac{p_R}{q_R}\, dR\,. \tag{1.2}$$

Here and in the sequel we understand

$$\log 0 = -\infty\,, \qquad \log \frac{a}{0} = +\infty\,, \qquad 0 \cdot (\pm\infty) = 0\,. \tag{1.3}$$

$I(P\,||\,Q)$ is always nonnegative and vanishes only for $P = Q$.

We shall not be concerned with the information theoretic significance of I-divergence; rather, we look at it simply as a quantity measuring how much P differs from Q. Given a PD R, the set of PD's

$$S(R, \rho) = \{P : I(P\,||\,R) < \rho\} \qquad (0 < \rho \leqq \infty) \tag{1.4}$$

Received February 20, 1973; revised December 27, 1973.

AMS 1970 *subject classifications.* Primary 60-00-E05, 62-B10; Secondary 49-F22, 15-A48.

Key words and phrases. Geometry of probability distributions, minimum discrimination information, distributions with given marginals, contingency tables, iterative proportional fitting procedure.

Institute of Mathematical Statistics is collaborating with JSTOR to digitize, preserve, and extend access to *The Annals of Probability.* www.jstor.org

will be called an *I-sphere* with center R and radius ρ. If $\mathscr{E}$ is a convex set of PD's intersecting $S(R, \infty)$, a PD $Q \in \mathscr{E}$ satisfying

$$(1.5) \qquad I(Q \| R) = \min_{P \in \mathscr{E}} I(P \| R)$$

will be called the *I-projection* of R on $\mathscr{E}$. If such Q exists, the convexity of $\mathscr{E}$ guarantees its uniqueness since $I(P \| R)$ is strictly convex in P, as one immediately sees from (1.1).

As demonstrated by Kullback [14], minimization problems of type (1.5) play a basic role in the information-theoretic approach to statistics (see also [7], [9], [13], [17] etc.); they frequently occur also elsewhere, e.g., in the theory of large deviations, cf. Sanov [20], and in statistical physics, as maximization of entropy, cf. Jaynes [10]. In physics, the measure R is often not a PD; $R(X)$ may even be infinite. This does not make much difference in most respects, except that in the latter case the integral (1.1) may be negative, even $-\infty$ (which corresponds to infinite entropy), or undefined.

Let us emphasize that I-divergence is not a metric and in general the I-spheres $S(R, \rho)$ do not even define a topology (as a base of the neighborhood system of R). This negative statement remains true if $I(P \| Q)$ is replaced by the symmetric divergence $I(P \| Q) + I(Q \| P)$—used already by Jeffreys [11]—or by any reasonable function of $I(P \| Q)$ and $I(Q \| P)$, see Csiszár [3]. In spite of these discouraging facts, it will be shown that certain analogies exist between properties of PD's and Euclidean geometry, where I-divergence plays the role of squared Euclidean distance. In particular, a "geometric" approach will be helpful in the study of I-projections, i.e., of the extremum problem (1.5).

In Section 2, using an analogue of the parallelogram identity, we first prove that the I-projection always exists if the convex set $\mathscr{E}$ is closed in the topology of the variation distance

$$(1.6) \qquad |P - Q| = \int |p_R - q_R| \, dR$$

(where R is any PD with $P \ll R$, $Q \ll R$). Next we prove a lemma having the geometric interpretation that the PD's with $\int \log q_R \, dP = \rho$ form the "tangent hyperplane" of the I-sphere $S(R, \rho)$ at Q, where $\rho = I(Q \| R) < \infty$; for such P's

$$(1.7) \qquad I(P \| R) = I(P \| Q) + I(Q \| R) ,$$

which is an analogue of Pythagoras' theorem.

The resulting characterization of I-projection will be used in Section 3 to establish a necessary and a sufficient condition on the form of I-projection on a set $\mathscr{E}$ defined by linear constraints of a general type. In case of a finite number of integral constraints or marginal constraints, we obtain a necessary and sufficient characterization of I-projection. These results complete the known sufficient conditions following from the minimum discrimination information theorem of Kullback [14] and Kullback and Khairat [18]. As corollaries of independent interest, we arrive at generalizations of known results on existence of bivariate distributions or nonnegative matrices of a certain product form and with given marginals, see Hobby and Pyke [8] and, e.g., Sinkhorn [21].

Another "geometric" result of Section 2 (asserting the transitivity of I-projection) is used in Section 3 to prove the convergence of an iterative algorithm for finding the I-projection, which generalizes the familiar iterative proportional fitting procedure (IPFP) for adjusting a contingency table to given marginal distributions. Though the proof works only for finite X, it is of more general validity than the known convergence proofs for the IPFP, even if attention is restricted to that case.

Our last result is an existence proof for a case not covered in Section 2.

After having submitted the first version of this paper, the author became aware of related work of Čencov [2]; he has developped a geometry of I-divergence, looking at it with the reversed order of P and Q. Apparently, there is no intersection between his results and ours, except for Theorem 3.3, see the discussion there.

2. General "geometric" results on I-projections.

THEOREM 2.1. *If the convex set $\mathscr{E}$ of* PD'*s is variation-closed then each R with $S(R, \infty) \cap \mathscr{E} \neq \emptyset$ has an I-projection on $\mathscr{E}$.*

PROOF. The idea is similar to the proof of existence of projection in Hilbert space. Pick a sequence $P_n \in \mathscr{E}$ with $I(P_n \| R) < \infty$ (in particular, $P_n \ll R$) such that

$$(2.1) \qquad I(P_n \| R) \to \inf_{P \in \mathscr{E}} I(P \| R) .$$

Since

$$(2.2) \qquad I(P_m \| R) + I(P_n \| R) = 2I\left(\frac{P_m + P_n}{2} \middle\| R\right) + I\left(P_m \middle\| \frac{P_m + P_n}{2}\right) + I\left(P_n \middle\| \frac{P_m + P_n}{2}\right)$$

(this analogue of the parallelogram identity is readily checked by writing all terms as integrals with respect to R, using (1.2)), where $(P_m + P_n)/2 \in \mathscr{E}$ by convexity, the last two terms of (2.2) must converge to 0 as $m, n \to \infty$.

Using the inequality

$$(2.3) \qquad |P - Q| \leqq (2I(P \| Q))^{\frac{1}{2}}$$

proved independently in [4], [12] and [15], one concludes that

$$|P_m - P_n| \leqq \left|P_m - \frac{P_m + P_n}{2}\right| + \left|P_n - \frac{P_m + P_n}{2}\right|$$

converges to 0 as $m, n \to \infty$ and, consequently, P_n converges in variation to some PD Q:

$$(2.4) \qquad |P_n - Q| = \int |p_{nR} - q_R|\, dR \to 0 \qquad (n \to \infty)$$

(the convergence in variation of the PD's $P_n \ll R$ to Q clearly implies $Q \ll R$).

In view of (1.1), from (2.4) follows by Fatou's lemma[1]

$$I(Q\,\|\,R) \leqq \liminf_{n\to\infty} I(P_n\,\|\,R)\,. \tag{2.5}$$

As $\mathscr{E}$ is variation-closed, we have $Q \in \mathscr{E}$. On account of (2.1) and (2.5), it follows that Q is the I-projection of R on $\mathscr{E}$.

REMARK. The only role of the hypothesis that $\mathscr{E}$ is variation-closed has been to ensure that the PD Q with the properties (2.4) and (2.5) belongs to $\mathscr{E}$. If this is ensured in some other way, the assertion still holds, see Theorem 3.3.

For any three PD's with $Q \ll R$ and either of $I(P\,\|\,Q) < \infty$ and $I(P\,\|\,R) < \infty$ (thus $P \ll R$, too), (1.1) and (1.2)—using (1.3) if necessary—give rise to the identity

$$\begin{aligned} I(P\,\|\,R) - I(P\,\|\,Q) &= \int \left(p_R \log p_R - p_R \log \frac{p_R}{q_R}\right) dR \\ &= \int p_R \log q_R \, dR = \int \log q_R \, dP\,. \end{aligned} \tag{2.6}$$

Our further results will be based on

LEMMA 2.1. *If $I(P\,\|\,Q)$ and $I(Q\,\|\,R)$ are finite, the "segment joining P and Q" does not intersect the I-sphere $S(R, \rho)$ with radius $\rho = I(Q\,\|\,R)$,* i.e., *$I(P_\alpha\,\|\,R) \geqq I(Q\,\|\,R)$ for each* PD

$$P_\alpha = \alpha P + (1-\alpha)Q\,, \qquad 0 \leqq \alpha \leqq 1\,, \tag{2.7}$$

iff

$$\int \log q_R \, dP \geqq I(Q\,\|\,R)\,. \tag{2.8}$$

If

$$Q = \alpha P + (1-\alpha)P'\,, \qquad 0 < \alpha < 1\,, \tag{2.9}$$

then $I(Q\,\|\,R) < \infty$ implies $I(P\,\|\,R) < \infty$, and the segment joining P and P' does not intersect $S(R, \rho)$ (with $\rho = I(Q\,\|\,R)$) iff

$$\int \log q_R \, dP = I(Q\,\|\,R)\,. \tag{2.10}$$

PROOF. The hypotheses imply $P \ll R$, $Q \ll R$. Let $p_\alpha = \alpha p_R + (1-\alpha)q_R$ denote the R-density of P_α defined by (2.7) (in particular, $p_0 = q_R$, $p_1 = p_R$). Since p_α is linear in α and $t \log t$ is convex, $p_\alpha \log p_\alpha$ is a convex function of α and its difference quotient

$$f_\alpha = \frac{1}{\alpha}(p_\alpha \log p_\alpha - q_R \log q_R) \tag{2.11}$$

converges non-increasingly (as $\alpha \downarrow 0$) to

$$\lim_{\alpha\downarrow 0} f_\alpha = \frac{\partial}{\partial \alpha} p_\alpha \log p_\alpha \bigg|_{\alpha=0} = (p_R - q_R)(\log q_R + 1)\,. \tag{2.12}$$

[1] (2.5) is a particular case of a more general lower semicontinuity property of I-divergence, see Pinsker [19], Section 2.4, Assertion 5.

$f_1 = p_R \log p_R - q_R \log q_R$ is R-integrable by assumption, thus (1.1), (2.11) and (2.12) give, by the monotone convergence theorem,

$$\text{(2.13)} \qquad \frac{d}{d\alpha} I(P_\alpha \| R)\Big|_{\alpha=0} = \lim\nolimits_{\alpha\downarrow 0} \textstyle\int f_\alpha \, dR = \int (p_R - q_R)(\log q_R + 1)\, dR$$

$$= \textstyle\int \log q_R \, dP - I(Q \| R) \, .$$

This proves that if (2.8) does not hold then

$$I(P_\alpha \| R) < I(P_0 \| R) = I(Q \| R) \qquad \text{for some } \alpha > 0 \, .$$

The converse is trivial: (2.8) implies $I(P \| R) \geqq I(Q \| R)$ by (2.6), and P_α also satisfies (2.8) if P does.

(2.9) with $I(Q \| R) < \infty$ implies $P \ll Q \ll R$, $p_R \leqq \alpha^{-1} q_R$, thus by (1.1) $I(P \| R) < \infty$, too, and similarly $I(P' \| Q) < \infty$. The last assertion of Lemma 2.1 follows from the first one, because (2.8) for both P and P' with strict inequality in either case would contradict to (2.9).

Lemma 2.1 means, intuitively, that the "tangent hyperplane" of $S(R, \rho)$ at Q consists of the PD's satisfying (2.10); according to (2.6), this is equivalent to (1.7), thus we have an analogue of Pythagoras' theorem. This geometric interpretation is limited, however, to $P \in S(R, \infty) \cup S(Q, \infty)$; if both $I(P \| R)$ and $I(Q \| R)$ are infinite, the integral $\int \log q_R \, dP$ may or may not be defined and if it is, its value may be arbitrary (the case of $P \in S(Q, \infty) \backslash S(R, \infty)$, i.e., $I(P \| Q) < I(P \| R) = \infty$ is not contained in Lemma 2.1, either; but then (2.6) applies and shows that (2.8) is trivially valid).

Lemma 2.1 and the identity (2.6) immediately give rise to

THEOREM 2.2. *A* PD $Q \in \mathscr{E} \cap S(R, \infty)$ *is the I-projection of R on the convex set* $\mathscr{E}$ *of* PD'*s* iff *every* $P \in \mathscr{E} \cap S(R, \infty)$ *satisfies* (2.8) *or, equivalently,* iff

$$\text{(2.14)} \qquad I(P \| R) \geqq I(P \| Q) + I(Q \| R) \qquad \textit{for every } P \in \mathscr{E} \, .$$

If the I-projection Q is an algebraic inner point of $\mathscr{E}$ *then* $\mathscr{E} \subset S(R, \infty)$ *and* (2.8) *and* (2.14) *hold with the equality.*

A $Q \in \mathscr{E}$ is called an algebraic inner point of $\mathscr{E}$ if for every $P \in \mathscr{E}$ there exist α and $P' \in \mathscr{E}$ satisfying (2.9).

REMARK. (2.14) shows, in particular, that if the I-projection Q of R on $\mathscr{E}$ exists then $P \ll Q$ for every $P \in \mathscr{E} \cap S(R, \infty)$. Thus, if some $P \in \mathscr{E}$ with $I(P \| Q) < \infty$ is measure-theoretically equivalent to R, then so is Q, as well.

Intuition suggests that if $\mathscr{E}$ is a *linear* set of PD's—i.e., if with P and P' also $\alpha P + (1 - \alpha) P'$ belongs to $\mathscr{E}$ for every real α for which it is a PD—then $\mathscr{E}$ always lies in the tangent hyperplane of $S(R, \rho)$ at Q, the I-projection of R on $\mathscr{E}$ (with $\rho = I(Q \| R)$), i.e., that the identity (1.7) is valid for every $P \in \mathscr{E}$. It will be shown in the next section that this conjecture is not generally true but in the most important cases—in particular, for finite X—it is. This additivity relation and its consequence, the following transitivity property of I-projection, proved for various particular cases by Kullback [14], [17], Ku and Kullback [13], etc., is very essential for informational statistical analysis.

THEOREM 2.3. *Let $\mathscr{E}$ and $\mathscr{E}_1 \subset \mathscr{E}$ be convex sets of* PD's, *let R have I-projection Q on $\mathscr{E}$ and I-projection Q_1 on $\mathscr{E}_1$, and suppose that the identity* (1.7) *holds for every $P \in \mathscr{E}$. Then Q_1 is the I-projection of Q on $\mathscr{E}_1$.*

PROOF. Applying (2.14) with Q_1 in the role of Q and (1.7) with Q_1 in the role of P, we have for $P \in \mathscr{E}_1$

$$I(P\,||\,R) \geqq I(P\,||\,Q_1) + I(Q_1\,||\,R) = I(P\,||\,Q_1) + I(Q_1\,||\,Q) + I(Q\,||\,R)\,. \tag{2.15}$$

Comparing (2.15) with (1.7), $I(Q\,||\,R)$ cancels out, yielding

$$I(P\,||\,Q) \geqq I(P\,||\,Q_1) + I(Q_1\,||\,Q) \qquad \text{for every } P \in \mathscr{E}_1\,. \tag{2.16}$$

Theorem 2.3 completes the geometric results on I-divergence needed for our purposes. Of course, intuition should be used with caution. For example, if R has I-projection Q on a convex set $\mathscr{E}$ of PD's, it does not follow that the elements of the "joining segment" of Q and R have the same I-projection on $\mathscr{E}$.

3. Minimizing I-divergence under linear constraints. A general formulation of a useful result known as minimum discrimination information theorem (Kullback [14], Kullback and Khairat [18]) is the following: For any (not necessarily convex) set $\mathscr{E}$ of PD's, if there exists a $Q \in \mathscr{E}$ with R-density $c \exp g(x)$ where $\int g\, dP_1 = \int g\, dP_2 < \infty$ for any $P_1, P_2 \in \mathscr{E}$, then $I(Q\,||\,R) = \min_{P \in \mathscr{E}} I(P\,||\,R)$; more exactly, in this case

$$I(P\,||\,R) = I(P\,||\,Q) + I(Q\,||\,R) \qquad \text{for all } P \in \mathscr{E}\,. \tag{3.1}$$

Observe that this immediately follows from the identity (2.6).

Two particular cases deserve main attention:

(A) $\mathscr{E}$ is defined by constraints of form $\int f_i\, dP = a_i$, $i = 1, \cdots, k$. Then, if a $Q \in \mathscr{E}$ with

$$q_R(x) = c \exp \textstyle\sum_{i=1}^{k} t_i f_i(x) \tag{3.2}$$

exists, it is the I-projection of R on $\mathscr{E}$ and (3.1) holds.

(B) $(X, \mathscr{X}) = (X_1, \mathscr{X}_1) \times (X_2, \mathscr{X}_2)$ and $\mathscr{E}$ consists of the PD's P with given marginals P_i on $(X_i, \mathscr{X}_i)$, $i = 1, 2$. Then, if a $Q \in \mathscr{E}$ with

$$q_R(x_1, x_2) = a(x_1)b(x_2)\,, \qquad \log a \in L_1(P_1)\,, \qquad \log b \in L_1(P_2) \tag{3.3}$$

exists, it is the I-projection of R on $\mathscr{E}$ and, again, (3.1) holds.

Our next aim is to complete the mentioned results for cases (A) and (B). We shall not explicitly consider the equally important case of PD's on a multiple product space with given marginals of certain (arbitrary) types, since the extension of our results from case (B) to that case is trivial. For example, if $(X, \mathscr{X}) = \times_{i=1}^{4} (X_i, \mathscr{X}_i)$ and $\mathscr{E}$ consists of the PD's with given marginals (of types shown by the indices) P_{123}, P_{124} and P_{34}, say, then the extension of Corollary

3.1 below is that a $Q \in \mathscr{E}$ is the I-projection of R on $\mathscr{E}$ iff

$$q_R(x_1, x_2, x_3, x_4) = a(x_1, x_2, x_3)b(x_1, x_2, x_4)c(x_3, x_4)$$

with $\log a \in L_1(P_{123})$, $\log b \in L_1(P_{124})$, $\log c \in L_1(P_{34})$ except, possibly, for a set N where q_R vanishes and $P(N) = 0$ whenever $P \in \mathscr{E}$, $I(P \| R) < \infty$; then (3.1) holds, too.

The following theorem concerns sets of PD's defined by linear constraints of a general type. Since no existence assertions will be made, we need not formally exclude even $\mathscr{E} = \emptyset$, i.e., contradicting constraints.

THEOREM 3.1. *Let $\{f_\gamma\}_{\gamma \in \Gamma}$ be an arbitrary set of real-valued $\mathscr{X}$-measurable functions on X and $\{a_\gamma\}_{\gamma \in \Gamma}$ be real constants. Let $\mathscr{E}$ be the set of all those* PD*'s P on $(X, \mathscr{X})$ for which the integrals $\int f_\gamma \, dP$ exist and equal a_γ $(\gamma \in \Gamma)$. Then, if a* PD *R has I-projection Q on $\mathscr{E}$, its R-density is of form*

$$(3.4) \qquad \begin{aligned} q_R(x) &= c \exp g(x) && \textit{if} \quad x \notin N \\ &= 0 && \textit{if} \quad x \in N \end{aligned}$$

where N has $P(N) = 0$ for every $P \in \mathscr{E} \cap S(R, \infty)$ and g belongs to the closed subspace of $L_1(Q)$ spanned by the f_γ's. On the other hand, if a $Q \in \mathscr{E}$ has R-density of form (3.4) *where g belongs to the linear space spanned by the f_γ's (without closure) then Q is the I-projection of R on $\mathscr{E}$ and* (3.1) *holds.*

COROLLARY 3.1. *In case* (A) *or* (B) *above, a $Q \in \mathscr{E}$ is the I-projection of R on $\mathscr{E}$* iff *q_R is of form* (3.2) *or* (3.3), *respectively, except possibly for a set N where q_R vanishes and $P(N) = 0$ foı every $P \in \mathscr{E} \cap S(R, \infty)$; in both cases, the identity* (3.1) *holds. If, in particular, some $P \in \mathscr{E}$ with $I(P \| R) < \infty$ is measure-theoretically equivalent to R then* (3.2) *or* (3.3) *is necessary and sufficient for Q to be the I-projection of R on $\mathscr{E}$.*

Before giving the proof, let us show by an example that for the I-projection on a set $\mathscr{E}$ defined by linear constraints the identity (3.1) is not generally true (contrary to geometric intuition) and neither the necessary nor the sufficient condition of Theorem 3.1 is both necessary and sufficient, in general.

EXAMPLE. Let X be the unit interval, $\mathscr{X}$ the Borel σ-algebra and Q the Lebesgue measure. Let $\mathscr{E}$ be the set of PD's satisfying $\int f_n \, dP = \frac{1}{4}$, $n = 1, 2, \cdots$, where

$$(3.5) \qquad \begin{aligned} f_n(x) &= 1 + \frac{n^{\frac{1}{2}}}{4} && \text{if} \quad 0 < x < \frac{1}{4n} \\ &= 1 && \text{if} \quad \frac{1}{4n} \leqq x < \tfrac{1}{4} \\ &= -\frac{1}{4n^{\frac{1}{2}}} && \text{it} \quad \tfrac{1}{4} \leqq x < \tfrac{1}{2} \\ &= 0 && \text{if} \quad \tfrac{1}{2} \leqq x < 1\,. \end{aligned}$$

Let the PD R be determined by the condition $q_R(x) = c \exp g(x)$ where

$$\begin{aligned} g(x) = -\lim_{n\to\infty} f_n(x) &= -1 \quad \text{if} \quad 0 < x < \tfrac{1}{4} \\ &= 0 \quad \text{if} \quad \tfrac{1}{4} \leqq x < 1\,. \end{aligned} \tag{3.6}$$

Then $Q \in \mathscr{E}$, and on account of Fatou's lemma

$$\begin{aligned} \int \log q_R \, dP &= \log c - \int \lim_{n\to\infty} f_n \, dP \\ &\geqq \log c - \tfrac{1}{4} = I(Q \,||\, R) \end{aligned} \tag{3.7}$$

for all $P \in \mathscr{E}$. This means, by Theorem 2.2, that Q is the I-projection of R on $\mathscr{E}$. It is easy to find $P \in \mathscr{E}$ for which in (3.7) the strict inequality holds, e.g. the PD with Q-density

$$\begin{aligned} p_Q(x) &= \frac{1}{5x^{\frac{1}{2}}} \quad \text{if} \quad 0 < x < \tfrac{1}{4} \\ &= 0 \quad \text{if} \quad \tfrac{1}{4} \leqq x < \tfrac{3}{5} \\ &= 2 \quad \text{if} \quad \tfrac{3}{5} \leqq x < 1\,. \end{aligned} \tag{3.8}$$

Thus (3.1) is false in this case; in particular, Q cannot meet the sufficient condition of Theorem 3.1. If $g(x)$ is given the opposite sign and R is defined accordingly, we obtain $\int \log q_R \, dP < I(Q \,||\, R)$ for the P defined by (3.8); this means that Q cannot be the I-projection of R on $\mathscr{E}$, showing that the necessary condition of Theorem 3.1 is not sufficient.

Proof of Theorem 3.1. If Q is the I-projection of R on $\mathscr{E}$ then for $N = \{x : q_R(x) = 0\}$ necessarily $P(N) = 0$ for each $P \in \mathscr{E} \cap S(R, \infty)$; see the remark to Theorem 2.2.

Let $\mathscr{E}' \subset \mathscr{E}$ be the set of PD's $P \in \mathscr{E}$ with $p_Q(x) \leqq 2$. If $P \in \mathscr{E}'$, there is a $P' \in \mathscr{E}'$ with $p_Q'(x) = 2 - p_Q(x)$, and with it $Q = (P + P')/2$; thus Q is an algebraic inner point of $\mathscr{E}'$. Applying Theorem 2.2 to $\mathscr{E}'$ instead of $\mathscr{E}$ we obtain $\int \log q_R \, dP = I(Q \,||\, R)$, i.e.,

$$\int \log q_R (p_Q - 1) \, dQ = 0 \qquad \text{for all } P \in \mathscr{E}'\,. \tag{3.9}$$

But for any $\mathscr{X}$-measurable function h with $|h(x)| \leqq 1$ such that

$$\int h \, dQ = 0 \quad \text{and} \quad \int f_\gamma h \, dQ = 0 \qquad \text{for each } \gamma \in \Gamma\,, \tag{3.10}$$

there exists a $P \in \mathscr{E}'$ with $p_Q = 1 + h$. Thus (3.9) gives

$$\int \log q_R h \, dQ = 0 \tag{3.11}$$

for all such h and therefore also for all $h \in L_\infty(Q)$ satisfying (3.10).

Hence follows that $\log q_R$ belongs to the (closed) subspace of $L_1(Q)$ spanned by 1 and the f_γ's. In fact, were this not the case, in view of the Hahn–Banach theorem ([22] page 106) there would exist a bounded linear functional on $L_1(Q)$ vanishing on the mentioned subspace but not at $\log q_R$; since the dual of $L_1(Q)$ is $L_\infty(Q)$ ([22] page 115), this is a contradiction. This proves the first assertion of Theorem 3.1.

The second part is much easier. Suppose that q_R is of the stated form. Since g is a finite linear combination of f_γ's, $\int g\,dP$ is constant for $P \in \mathscr{E}$ and

$$(3.12) \qquad \int \log q_R\,dP = \log c + \int g\,dP = \text{const} = I(Q\,||\,P)$$

for $P \in \mathscr{E}$, $P \ll Q$. But for $P \in \mathscr{E}$ both $I(P\,||\,R) < \infty$ (by hypothesis) and $I(P\,||\,Q) < \infty$ (by definition) imply $P \ll Q$; thus (3.1) follows from (2.6) and (3.12).

To prove the corollary, observe that case (B) does fit into the considered model, taking for f_γ's the P_i-integrable functions $f(x_i)$, $i = 1, 2$ (looking at them as functions of (x_1, x_2)). Theorem 3.1 clearly gives a necessary and sufficient condition on q_R and guarantees the validity of (3.1) for the I-projection Q if the linear space spanned by the f_γ's is closed in $L_1(P)$ for each $P \in \mathscr{E}$. But the latter hypothesis is fulfilled in both cases (A) and (B), completing the proof.

Theorem 3.1 and its corollary leaves the question of existence of I-projection open. If $\mathscr{E}$ is variation-closed, as in the case of bounded f_γ's or in case (B), Theorem 2.1 guarantees the existence provided that $\mathscr{E} \neq \emptyset$ and $I(P\,||\,R) < \infty$ for some $P \in \mathscr{E}$. For case (A) with not bounded f_i's, see Theorem 3.3 below.

As a consequence of Corollary 3.1 and Theorem 2.1 we obtain

COROLLARY 3.2. *To given* PD'*s* P_i *on* $(X_i, \mathscr{X}_i)$, $i = 1, 2$ *and* R *on* $(X_1 \times X_2, \mathscr{X}_1 \times \mathscr{X}_2)$, *there exists a* PD Q *on the product space with marginals* P_1 *and* P_2 *and with R-density of form* $a(x_1)b(x_2)$, $\log a \in L_1(P_1)$, $\log b \in L_1(P_2)$ iff *there exists any* P *measure-theoretically equivalent to* R *which has the prescribed marginals and satisfies* $I(P\,||\,R) < \infty$.

Considering $R \ll P_1 \times P_2$ with density $f(x_1, x_2)$ and using $P = P_1 \times P_2$ in Corollary 3.2, we obtain for the existence of a PD with marginals P_1 and P_2 and having $P_1 \times P_2$-density of form $a(x_1)b(x_2)f(x_1, x_2)$ the sufficient condition $\log f \in L_1(P_1 \times P_2)$. It is interesting to compare this with a result of Hobby and Pyke [8]; their theorem, when specialized to our problem, gives the sufficient condition $0 < a \leqq f(x_1, x_2) \leqq K$.

Specializing Corollary 3.2 to finite X_1 and X_2, we obtain

COROLLARY 3.3. *Let* A *be an* $m \times n$ *matrix with nonnegative elements. For the existence of positive diagonal matrices* D_1 *and* D_2 *such that the row and column sums of* D_1AD_2 *be given positive numbers, it is necessary and sufficient that some* B *with nonnegative elements and with the given row and column sums have the same zero entries as* A (*if any*).

PROOF. Without any loss of generality, the elements of A and both the given row and column sums may be assumed to add up to one. Then A defines a PD R on $X_1 \times X_2$ and D_1AD_2 defines a PD having R-density of form (3.3). Since for PD's on finite sets $P \ll R$ implies $I(P\,||\,R) < \infty$, the assertion follows from Corollary 3.2.

Corollary 3.3 solves a matrix-theoretic problem, partial solutions of which have been given by many authors. Sinkhorn [21] has shown that the positivity of A is a sufficient condition by proving the convergence of the iterative proportional fitting procedure (IPFP) dating back to Deming and Stephan [5]. This IPFP and its extensions are widely used in the analysis of contingency tables. Of the extensive literature of the subject we mention here only Ireland and Kullback [9], Ku and Kullback [13] and Fienberg [6]; further references may be found there.[2]

The IPFP for adjusting a PD R given on a finite product space to k marginal constraints, i.e., to given marginal distributions of arbitrary types, consists in the successive calculation of PD's Q_n on the product space starting from $Q_0 = R$: to obtain Q_n, the probabilities of Q_{n-1} are multiplied by the ratios of the corresponding marginal probabilities of the nth constraint and of Q_{n-1}. Here the constraints are looked at cyclically repeated. As shown by Ireland and Kullback [9], Q_n is just the I-projection of Q_{n-1} on $\mathscr{E}_n$, where $\mathscr{E}_i$ is the set of PD's satisfying the ith constraint, and $Q = \lim_{n\to\infty} Q_n$ (if it exists) is the I-projection of R on $\mathscr{E} = \bigcap_{i=1}^k \mathscr{E}_i$, the set of PD's satisfying all k marginal constraints. Kullback [16] generalized the method for the non-discrete case, too.

Motivated by the approach of Ireland and Kullback [9], we are going to formulate the procedure in a general setup and prove convergence to the required I-projection, provided that X is a finite set. Unlike previous convergence proofs for the IPFP (see Fienberg [6] and the references there), we shall not need any assumption on the positivity of the probabilities of R. It should be noted that the convergence proof in [9] is incomplete since formula (4.38) does not imply (4.39); in [16] there is a similar flaw.

THEOREM 3.2. *Let $\mathscr{E}_1, \cdots, \mathscr{E}_k$ be arbitrary linear sets of* PD'*s on a finite set X with $\mathscr{E} = \bigcap_{i=1}^k \mathscr{E}_i \neq \emptyset$, let R be any* PD *to which there exists $P \in \mathscr{E}$ with $P \ll R$, and define $Q_1, Q_2, \cdots$ recursively by letting Q_n be the I-projection of Q_{n-1} on $\mathscr{E}_n$, $n = 1, 2, \cdots$ where $Q_0 = R$ and*

$$\mathscr{E}_n = \mathscr{E}_i \quad \text{if} \quad n = mk + i\,, \qquad 1 \leqq i \leqq k\,. \tag{3.13}$$

Then Q_n converges (*pointwise or, equivalently, in variation*) *to the I-projection Q of R on $\mathscr{E}$.*

PROOF. Any linear set $\mathscr{E}$ of PD's on a finite set X of size r, say, can be looked at as the intersection of a linear subset of E^r with the simplex representing the PD's on X. Hence $\mathscr{E}$ is closed and can be defined by a finite number of linear constraints. In view of Theorem 2.1 and Corollary 3.1, the I-projection of R on such an $\mathscr{E}$ always exists if $I(P\,||\,R) < \infty$—now equivalent to $P \ll R$—for some $P \in \mathscr{E}$, and then (3.1) holds, as well.

Under the hypotheses of Theorem 3.2, it follows that the I-projections Q_1, $Q_2, \cdots$ and Q exist and $I(P\,||\,Q_{n-1}) = I(P\,||\,Q_n) + I(Q_n\,||\,Q_{n-1})$ for any $P \in \mathscr{E}_n$,

[2] The iterative algorithms suggested in [17] apparently do not belong to the class of generalizations of the IPFP considered below. But also for the problems considered there, it is straightforward to give convergent iterations within the framework of Theorem 3.2.

$n = 1, 2, \cdots$. Setting $P = Q$, in particular, we obtain by induction

$$(3.14) \qquad I(Q \| R) = I(Q \| Q_n) + \textstyle\sum_{i=1}^{n} I(Q_i \| Q_{i-1}), \qquad n = 1, 2, \cdots .$$

Since X is finite, each subsequence of Q_n contains a convergent subsequence; it suffices to show that $Q_{n_l} \to Q'$ implies $Q' = Q$. First verify $Q' \in \mathscr{E}$. We have from (3.14)

$$\textstyle\sum_{i=1}^{\infty} I(Q_i \| Q_{i-1}) \leqq I(Q \| R) < \infty$$

thus $I(Q_n \| Q_{n-1}) \to 0$, implying $|Q_n - Q_{n-1}| \to 0$ by (2.3). Thus $Q_{n_l+1}, \cdots, Q_{n_l+k}$ also converge to Q' as $l \to \infty$. Since these PD's belong to (a cyclic permutation of) the closed sets $\mathscr{E}_1, \cdots, \mathscr{E}_k$ respectively, see (3.13), we conclude $Q' \in \bigcap_{i=1}^{k} \mathscr{E}_i = \mathscr{E}$.

Repeated application of Theorem 2.3 shows that Q is the I-projection on $\mathscr{E}$ of $Q_1, \cdots, Q_n, \cdots$, as well, thus

$$(3.15) \qquad I(P \| Q_n) = I(P \| Q) + I(Q \| Q_n) \qquad \text{for all } P \in \mathscr{E},$$

$n = 1, 2, \cdots$. Applying this to $P = Q'$, we obtain $I(Q' \| Q) = 0$, i.e., $Q' = Q$, because for finite $XQ_{n_l} \to Q'$ implies $I(Q' \| Q_{n_l}) \to 0$. The proof is complete.

Finally, let us return to the problem of existence of I-projection in case (A), if the f_i's are not necessarily bounded. One possible approach is to prove in a direct way that there exists a $Q \in \mathscr{E}$ with R-density (3.2). This is not easy but has been done under fairly general conditions by Čencov [2], Theorem 23.1. Here we show how the method of Theorem 2.1 applies to this case. Our hypothesis will be that

$$(3.16) \qquad T_R = \{(t_1, \cdots, t_n) : \exp \textstyle\sum_{i=1}^{k} t_i f_i(x) \text{ is } R\text{-integrable}\}$$

is an open set in E^k; this clearly implies that $f_i(x) \exp \sum_{i=1}^{k} t_i f_i(x)$ is R-integrable for every $(t_1, \cdots, t_k) \in T_R$, $i = 1, \cdots, k$. (In the first version of this paper, $T_R = E^k$ was assumed. The strengthening has been inspired by Čencov's result, *loc. cit.*, which implies the existence of I-projection even under a slightly weaker hypothesis.)

THEOREM 3.3. *Let $\mathscr{E}(a_1, \cdots, a_k)$ be the set of* PD'*s satisfying $\int f_i \, dP = a_i$, $i = 1, \cdots, k$ and let A_R be the set of points $(a_1, \cdots, a_k) \in E^k$ for which $\mathscr{E}(a_1, \cdots, a_k)$ contains some P with $I(P \| R) < \infty$. Then, supposing that T_R is open, the I-projection of R on $\mathscr{E}(a_1, \cdots, a_k)$ exists for each inner point $(a_1, \cdots, a_k)$ of A_R, and its R-density is of form* (3.2).

REMARKS. It can be shown that the interior of A_R coincides with that of the convex hull of the support of R', the image of R in E^k at the mapping $x \to (f_1(x), \cdots, f_k(x))$. Thus, assuming that the f_i's are linearly independent mod R, the interior of A_R is nonvoid. If $(a_1, \cdots, a_k) \in A_R$ is on the boundary of A_R, typically there still exists the I-projection of R on $\mathscr{E}(a_1, \cdots, a_k)$ but its R-density vanishes on a set N of $R(N) > 0$. These problems will not be entered here.

We shall need the following lemma of some independent interest.

LEMMA 3.1. *For any (measurable) function $f(x)$ for which $e^{tf(x)}$ is Q-integrable if $|t|$ is sufficiently small, $I(P_n \| Q) \to 0$ implies $\int f\, dP_n \to \int f\, dQ$.*

PROOF. Let p_n denote the Q-density of P_n; it surely exists if $I(P_n \| Q) < \infty$. In view of (2.3), $I(P_n \| Q) \to 0$ implies $|P_n - Q| = \int |p_n - 1|\, dQ \to 0$, hence on $A_k = \{x : |f(x)| \leqq K\}$ we have $\int_{A_k} f\, dP_n \to \int_{A_k} f\, dQ$. Thus it suffices to show that to any $\varepsilon > 0$ there exists K such that

$$\text{(3.17)} \qquad \limsup_{n\to\infty} \int_{X\setminus A_k} |f|\, dP_n = \limsup_{n\to\infty} \int_{X\setminus A_k} |f| p_n\, dQ < \varepsilon\,.$$

But $I(P_n \| Q) = \int p_n \log p_n\, dQ \to 0$ implies $\lim_{n\to\infty} \int_A p_n \log p_n\, dQ = 0$ for every $A \in \mathscr{X}$ (apply Fatou's lemma to both A and $X\setminus A$). Choosing $t > 0$ and K to satisfy $\int_{X\setminus A_k} e^{t|f|}\, dQ < \varepsilon t$, (3.17) follows from the inequality $ab < a \log a + e^b$ (see [1] Section 15), substituting $a = p_n(x)$, $b = t|f(x)|$.

PROOF OF THEOREM 3.3. On account of the convexity of $I(P \| R)$ in P, A_R is a convex set and

$$\text{(3.18)} \qquad F(a_1, \cdots, a_k) = \inf_{P \in \mathscr{E}(a_1,\cdots,a_k)} I(P \| Q)$$

is a finite valued convex function on A_R. Hence, if $(a_1, \cdots, a_k)$ is an inner point of A_R, there exists $(t_1, \cdots, t_k)$ such that

$$\text{(3.19)} \qquad F(b_1, \cdots, b_k) \geqq F(a_1, \cdots, a_k) + \sum_{i=1}^k t_i(b_i - a_i) \qquad \text{for all } (b_1, \cdots, b_k) \in A_R\,.$$

First we show that $(t_1, \cdots, t_k) \in T_R$, see (3.16).

Let $P_n \in \mathscr{E}(a_1, \cdots, a_k)$, $I(P_n \| R) \to F(a_1, \cdots, a_k)$; then P_n converges in variation to some Q by the proof of Theorem 2.1. Let $f_i^{(n)}(x) = f_i(x)$ if $t_i f_i(x) \leqq K_n$ and $f_i^n(x) = 0$ else, where $K_n \uparrow \infty$, and let Q_n be the PD with R-density

$$\text{(3.20)} \qquad q_{nR}(x) = c_n \exp \sum_{i=1}^k t_i f_i^n(x)\,.$$

From (3.20) and (1.1) follows

$$\text{(3.21)} \qquad I(Q_n \| R) = \int \log q_{nR}\, dP_n + \sum_{i=1}^k t_i\left(\int f_i^n\, dQ_n - \int f_i^n\, dP_n\right).$$

Since $(0_1, \cdots, 0) \in T_R$ and T_R is open, the f_i's are R-integrable and thus Q_n-integrable, too; it follows that for large n $\int f_i^n\, dQ_n$ is arbitrarily close to $\int f_i\, dQ_n = b_i^n$, say (note, that the sequence c_n is non-increasing). Choosing the K_n's properly, also $\int f_i^n\, dP_n$ will be close to $\int f_i\, dP_n = a_i$ if n is large, and then the identities (3.21) and (2.6) compared with the inequality (3.19) (with b_i^n in the role of b_i) give rise to $I(P_n \| Q_n) \to 0$.

On account of (2.3), it follows that the Q_n's with R-density (3.20) also converge in variation to Q, hence the latter has R-density (3.2); in particular, $(t_1, \cdots, t_k) \in T_R$.

Setting $b_i = \int f_i\, dQ$, similarly to (3.21) we have

$$\text{(3.22)} \qquad I(Q \| R) = \int \log q_R\, dP_n + \sum_{i=1}^k t_i(b_i - a_i)\,,$$

whence—again by (2.6) and (3.19)—we obtain that $I(P_n \| Q) \to 0$. Using the assumption that T_R is an open set, Lemma 3.1 gives $\int f_i\, dQ = \lim_{n\to\infty} \int f_i\, dP_n = a_i$, $i = 1, \cdots, k$. The proof is complete.

REMARK. It follows that $F(a_1, \cdots, a_k)$—see (3.18)—is differentiable at every inner point of A_R and grad $F(a_1, \cdots, a_k) = (t_1, \cdots, t_k)$ is just the parameter vector in (3.2) for the I-projection Q of R on $\mathscr{E}(a_1, \cdots, a_k)$.

REFERENCES

[1] BECKENBACH, E. F. and BELLMAN, R. (1961). *Inequalities*. Springer-Verlag, Berlin.

[2] ČENCOV, N. N. (1972). *Statistical Decision Rules and Optimal Decisions* (in Russian). Nauka, Moskow.

[3] CSISZÁR, I. (1964). Über topologische und metrische Eigenschaften der relativen Information der Ordnung α. *Trans. Third Prague Conference of Information Theory etc.* Publishing House of the Czehoslovak Academy of Sciences, Prague, 63–73.

[4] CSISZÁR, I. (1967). Information-type measures of difference of probability distributions and indirect observations. *Studia Sci. Math. Hungar.* **2** 299–318.

[5] DEMING, W. E. and STEPHAN, F. F. (1940). On a least square adjustment of a sampled frequency table when the expected marginal totals are known. *Ann. Math. Statist.* **11** 427–444.

[6] FIENBERG, S. (1970). An iterative procedure for estimation in contingency tables. *Ann. Math. Statist.* **41** 907–917.

[7] GOOD, I. J. (1963). Maximum entropy for hypothesis formulation, especially for multidimensional contingency tables. *Ann. Math. Statist.* **34** 911–934.

[8] HOBBY, C. and PYKE, R. (1965). Doubly stochastic operators obtained from positive operators. *Pacific J. Math.* **15** 153–157.

[9] IRELAND, C. T. and KULLBACK, S. (1968). Contingency tables with given marginals. *Biometrika* **55** 179–188.

[10] JAYNES, E. T. (1957). Information theory and statistical mechanics. *Phys. Rev.* **106** 620–630.

[11] JEFFREYS, H. (1948). *Theory of Probability*. 2nd ed. Clarendon Press, Oxford.

[12] KEMPERMAN, J. H. B. (1967). On the optimum rate of transmitting information. *Probability and Information Theory*. Lecture Notes in Mathematics, Springer-Verlag, 126–169.

[13] KU, H. H. and KULLBACK, S. (1968). Interaction in multidimensional contingency tables: an information theoretic approach. *J. Res. Nat. Bur. Standards* **72** 159–199.

[14] KULLBACK, S. (1959). *Information Theory and Statistics*. Wiley, New York.

[15] KULLBACK, S. (1967). A lower bound for discrimination information in terms of variation. *IEEE Trans. Information Theory*. **IT-13** 126–127.

[16] KULLBACK, S. (1968). Probability densities with given marginals. *Ann. Math. Statist.* **39** 1236–1243.

[17] KULLBACK, S. (1971). Marginal homogeneity of multidimensional contingency tables. *Ann. Math. Statist.* **42** 594–606.

[18] KULLBACK, S. and KHAIRAT, M. A. (1966). A note on minimum discrimination information. *Ann. Math. Statist.* **37** 279–280.

[19] PINSKER, M. S. (1964). *Information and Information Stability of Random Variables and Processes*. Holden-Day, San Francisco.

[20] SANOV, I. N. (1957). On the probability of large deviations of random variables (in Russian). *Mat. Sbornik* **42** 11–44.

[21] SINKHORN, R. (1967). Diagonal equivalence to matrices with prescribed row and column sums. *Amer. Math. Monthly* **74** 402–405.

[22] YOSIDA, K. (1965). *Functional Analysis*. Springer-Verlag, Berlin.

MATHEMATICAL INSTITUTE
HUNGARIAN ACADEMY OF SCIENCES
1053 BUDAPEST, REÁLTANODA-U. 13–15
HUNGARY