I-DIVERGENCE GEOMETRY OF PROBABILITY DISTRIBUTIONS
AND MINIMIZATION PROBLEMS

BY I. CSIZSÁR

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 $(X, \mathcal{F})$ 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 $q_q(x)$, 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

$$I(P || Q) = \int p_q \log p_q dP = \int p_q \log p_q dQ \quad \text{if} \quad P \ll Q$$

$$= +\infty \quad \text{if} \quad P \nsubseteq Q.$$ 

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

$$I(P || Q) = \int p_x \log \frac{p_x}{q_x} dR.$$ 

Here and in the sequel we understand

$$\log 0 = -\infty, \quad \frac{d}{dx} \log x = +\infty, \quad 0 \cdot (\pm \infty) = 0.$$ 

$I(P || Q)$ 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

$$S(R, \rho) = \{P : I(P || R) < \rho\} \quad (0 < \rho \leq \infty)$$

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

$$I(Q || R) = \min_{P \in \mathcal{E}} I(P || R)$$

will be called the I-projection of $R$ on $\mathcal{E}$. If such $Q$ exists, the convexity of $\mathcal{E}$ 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 $\mathcal{E}$ is closed in the topology of the variation distance

$$|P - Q| = \int |p_n - q_n| dR$$

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

$$I(P || R) = I(P || Q) + I(Q || R),$$

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 $\mathcal{E}$ 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 Cencov [2]; he has developed 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 \( \mathcal{E} \) of PD's is variation-closed then each R with \( S(R, \infty) \cap \mathcal{E} \neq \emptyset \) has an I-projection on \( \mathcal{E} \).

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

\[
I(P_n \| R) \to \inf_{P \in \mathcal{E}} I(P \| R).
\]

Since

\[
I(P_n \| R) + I(P_{n+1} \| R) = 2I(P_n + P_{n+1} \| R) + I(P_n \| P_{n+1} + \frac{P_n + P_{n+1}}{2}) + I(P_{n+1} \| P_{n+1} + \frac{P_n + P_{n+1}}{2})
\]

(this analogue of the parallelogram identity is readily checked by writing all terms as integrals with respect to R, using (1.2), where \( (P_n + P_{n+1})/2 \in \mathcal{E} \) by convexity, the last two terms of (2.2) must converge to 0 as \( n \to \infty \).

Using the inequality

\[
|P - Q| \leq 2I(P \| Q)
\]

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

\[
|P_n - P_{n+1}| \leq |P_n - P_{n+1} + \frac{P_n + P_{n+1}}{2}| + |P_n - P_{n+1} - \frac{P_n + P_{n+1}}{2}|
\]

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

\[
|P_n - Q| \to 0 \quad (n \to \infty)
\]

(2.4)

(2.11) \( f_\alpha = \frac{1}{\alpha} (p_{\alpha} \log p_{\alpha} - q_{\alpha} \log q_{\alpha}) \)

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

\[
\lim_{\alpha \downarrow 0} f_\alpha = \frac{\partial}{\partial \alpha} \Big|_{\alpha=0} p_{\alpha} \log p_{\alpha} = (p_{\alpha} - q_{\alpha}) \log q_{\alpha} + 1.
\]

(2.12)

\[
I(Q \| R) \leq \lim_{\alpha \downarrow 0} I(p_{\alpha} \| R).
\]

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

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

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

\[
I(P \| R) - I(Q \| R) = \int \left( p_{\alpha} \log p_{\alpha} - p_{\alpha} \log \frac{p_{\alpha}}{q_{\alpha}} \right) dR
\]

(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., \( I(P \| R) \geq I(Q \| R) \) for each PD

\[
P_{\alpha} = \alpha P + (1 - \alpha)Q, \quad 0 \leq \alpha \leq 1,
\]

(2.7)

\[
\int \log q_{\alpha} dP \leq I(Q \| R).
\]

(2.8)

\[
O = \alpha P + (1 - \alpha)Q', \quad 0 < \alpha < 1,
\]

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

\[
\int \log q_{\alpha} dP \leq I(Q \| R).
\]

(2.9)

Proof. The hypotheses imply \( P \leq R, Q \leq R \). Let \( p_{\alpha} = \alpha p_{\alpha} + (1 - \alpha)q_{\alpha} \) denote the R-density of \( P_{\alpha} \) defined by (2.7) (in particular, \( p_{\alpha} = q_{\alpha} \) if \( p_{\alpha} = p_{\alpha} \)). Since \( p_{\alpha} \) is linear in \( \alpha \) and \( t \log t \) is convex, \( p_{\alpha} \log p_{\alpha} \) is a convex function of \( \alpha \) and its difference quotient

\[
(2.11)
\]
proven for various particular cases by Kullback [14], [17], Ku and Kullback [13], etc., is very important for informational statistical analysis.

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

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

\[
I(P || R) \geq I(P || Q_1) + I(Q_1 || R) = I(P || Q) + I(Q_1 || Q) + I(Q_1 || R).
\]

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

\[
I(P || R) \geq I(P || Q) + I(Q_1 || Q) \quad \text{for every } P \in \mathcal{E}_1.
\]

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 \( \mathcal{E} \) of PD's, it does not follow that the elements of the "joining segment" of \( Q \) and \( R \) have the same I-projection on \( \mathcal{E} \).

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 \( \mathcal{E} \) of PD's, if there exists a \( Q \in \mathcal{E} \) with R-density \( c \exp \{g(x)\} \) where \( \int g \, dP = \int g \, dP_0 \leq \infty \) for any \( P_0 \in \mathcal{E} \), then \( I(Q || R) = \min_{P \in \mathcal{E}} I(P || R) \); more exactly, in this case

\[
I(P || R) = I(P || Q) + I(Q || R) \quad \text{for all } P \in \mathcal{E}.
\]

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

Two particular cases deserve main attention:

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

\[
q_{\mathcal{E}}(x) = c \exp \sum_{i=1}^{k} t_i f_i(x)
\]

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

(B) \( (X, \mathcal{E}) = (X_0, \mathcal{E}_0) \times (X_1, \mathcal{E}_1) \) and \( \mathcal{E} \) consists of the PD's \( P \) with given marginals \( P_1 \) on \( (X_1, \mathcal{E}_1) \), \( i = 1, 2 \). Then, if a \( Q \in \mathcal{E} \) with

\[
q_i(x_1, x_2) = \delta(x_1) \delta(x_2), \quad \log a \in L(P_0), \quad \log b \in L(P_1)
\]

exists, it is the I-projection of \( R \) on \( \mathcal{E} \) 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 \( (X, \mathcal{E}) = X_{i=1}^n (X_i, \mathcal{E}_i) \) and \( \mathcal{E} \) consists of the PD's with given marginals (types shown by the indices) \( P_1, \ldots, P_n \), say, then the extension of Corollary
Let the PD $R$ be determined by the condition \( q_0(x) = c \exp g(x) \) where
\[
(3.6) \quad g(x) = -\lim_{x \to 1} f \left( x, n \right) = -1 \quad \text{if} \quad 0 < x < \frac{1}{4} = 0 \quad \text{if} \quad \frac{1}{4} \leq x < 1.
\]

Then $Q \in \mathcal{S}$, and on account of Fatou's lemma
\[
(3.7) \quad \int q_0 \, dP = \int c - \int \lim_{n \to 1} f \, dP \\
\geq \log c - \frac{1}{2} = \log \left( \|Q\| R \right)
\]
for all $P \in \mathcal{S}$. This means, by Theorem 2.2, that $Q$ is the $I$-projection of $R$ on $\mathcal{S}$. It is easy to find $P \in \mathcal{S}$ for which in (3.7) the strict inequality holds, e.g., the PD with $Q$-density
\[
p_0(x) = \begin{cases} \frac{1}{5x^4} & \text{if } 0 < x < \frac{1}{4} \\ 0 & \text{if } \frac{1}{4} \leq x < \frac{1}{2} \\ 2 & \text{if } \frac{1}{2} \leq x < 1. \end{cases}
\]

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 \( \int q_0 \, dP < \int \|Q\| R \) for the $P$ defined by (3.8); this means that $Q$ cannot be the $I$-projection of $R$ on $\mathcal{S}$, 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 $\mathcal{S}$ then for $N = \{ x : q_0(x) = 0 \}$ necessarily $P(N) = 0$ for each $P \in \mathcal{S} \cap S(R, \infty)$; see the remark to Theorem 2.2.

Let $\mathcal{S}' \subset \mathcal{S}$ be the set of PD's $P \in \mathcal{S}$ with $p_0(x) \leq 2$. If $P \in \mathcal{S}'$, there is a $P' \in \mathcal{S}'$ with $p_0'(x) = 2 - p_0(x)$, and with it $Q = (P + P')/2$; thus $Q$ is an algebraic inner point of $\mathcal{S}'$. Applying Theorem 2.2 to $\mathcal{S}'$ instead of $\mathcal{S}$ we obtain
\[
(3.9) \quad \int q_0(p_0 - 1) \, dQ = 0 \quad \text{for all } P \in \mathcal{S}'.
\]

But for any $\mathcal{S}$-measurable function $h$ with $|h(x)| \leq 1$ such that
\[
(3.10) \quad \int h \, dQ = 0 \quad \text{and} \quad \int f \, h \, dQ = 0 \quad \text{for each } r \in \Gamma,
\]
there exists a $P \in \mathcal{S}'$ with $p_0 = 1 + h$. Thus (3.9) gives
\[
(3.11) \quad \int q_0 h \, dQ = 0
\]
for all such $h$ and therefore also for all $h \in L_0(Q)$ satisfying (3.10).

Hence follows that $q_0$ belongs to the $(\text{closed})$ subspace of $L_0(Q)$ spanned by $1$ and the $f_\gamma$. 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 $L_0(Q)$ vanishing on the mentioned subspace but not at $q_0$; since the dual of $L_0(Q)$ is $L_0(Q)$ ([22] page 115), this is a contradiction. This proves the first assertion of Theorem 3.1.

**Example.** Let $X$ be the unit interval, $\mathcal{B}$ the Borel $\sigma$-algebra and $Q$ the Lebesgue measure. Let $\mathcal{S}$ be the set of PD's satisfying $\int f \, dP = \frac{1}{n}$, $n = 1, 2, \ldots,$ where
\[
(3.5) \quad f_\gamma(x) = \begin{cases} 1 + \frac{n}{4} & \text{if } 0 < x < \frac{1}{4n} \\ 1 & \text{if } \frac{1}{4n} \leq x < \frac{1}{4} \\ -\frac{1}{4n^2} & \text{if } \frac{1}{4} \leq x < \frac{1}{2} \\ 0 & \text{if } \frac{1}{2} \leq x < 1. \end{cases}
\]
The second part is much easier. Suppose that \( q_a \) is of the stated form. Since \( g \) is a finite linear combination of \( f_i \)'s, \( \int q_a dP \) is constant for \( P \in \mathcal{C} \) and
\[
(3.12) \quad \int \log q_a dP = \log c + \int g dP = \text{const} = I(\mathcal{Q}||P)
\]
for \( P \in \mathcal{C}, P \neq Q \). But for \( P \in \mathcal{C} \) both \( I(P||R) < \infty \) (by hypothesis) and \( I(\mathcal{Q}||Q) < \infty \) (by definition) imply \( P < Q \); 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 \( f_i \)'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 \( q_a \) and guarantees the validity of (3.1) for the \( L \)-projection \( Q \) if the linear space spanned by the \( f_i \)'s is closed in \( L(P) \) for each \( P \in \mathcal{C} \). 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 \( L \)-projection open. If \( \mathcal{C} \) is variation-closed, as in the case of bounded \( f_i \)'s or in case (B), Theorem 2.1 guarantees the existence provided that \( \mathcal{C} \neq \emptyset \) and \( I(P||R) < \infty \) for some \( P \in \mathcal{C} \). 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 \( (X_i, \mathcal{X}_i') \), \( i = 1, 2 \) and \( R \) on \( (X_1 \times X_2, \mathcal{X}_1 \times \mathcal{X}_2) \), there exists a \( PD \) \( Q \) on the product space with marginals \( P_1 \) and \( P_2 \) and with \( R \)-density of form \( q(x_1)q(x_2) \) if \( q \in L(P_1) \), \( q \in L(P_2) \) iff there exists any \( P \)-measure-theoretically equivalent to \( R \) which has the prescribed marginals and satisfies \( I(P||R) < \infty \).

Considering \( R \leq P_1 \times P_2 \) with density \( f(x_1, x_2) \) and using \( P = P_1 \times P_2 \), Corollary 3.2, we obtain for the existence of a \( PD \) with marginals \( P_1 \) and \( P_2 \) and having \( P \)-density of form \( q(x_1)q(x_2) \) the sufficient condition \( \log f \in L(P_1 \times P_2) \). 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 \( 0 < a \leq f(x_1, x_2) \leq K \).

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 \( PD \) 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 \( P \leq R \) 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.

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 \( Q_0 \) on the product space starting from \( Q_0 = R \) to obtain \( Q_k \), the probabilities of \( Q_0 \) are multiplied by the ratios of the corresponding marginal probabilities of the \( n \)th constraint and of \( Q_{k-1} \). Here the constraints are looked at cyclically repeated. As shown by Ireland and Kullback [9], \( Q_k \) is just the \( L \)-projection of \( Q_{k-1} \) on \( \mathcal{C}_k \), where \( \mathcal{C}_k \) is the set of \( PD \)'s satisfying the \( k \)th constraint, and \( Q = \lim_\rightarrow Q \) (if it exists) is the \( L \)-projection of \( R \) on \( \mathcal{C} = \bigcap_\rightarrow \mathcal{C}_k \), 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 \( L \)-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 \( \mathcal{C}_k, \ldots, \mathcal{C}_0 \) be arbitrary linear sets of \( PD \)'s on a finite set \( X \) with \( \mathcal{C}_0 = \bigcap_\rightarrow \mathcal{C}_k \neq \emptyset \), let \( R \) be any \( PD \) of which there exists \( P \in \mathcal{C} \) with \( P \leq R \), and define \( Q_0, Q_1, \ldots \) recursively by letting \( Q_0 \) be the \( L \)-projection of \( Q_{k-1} \) on \( \mathcal{C}_k \), \( n = 1, 2, \ldots \) where \( Q_k = R \) and
\[
(3.13) \quad Q_n = Q_{n-1} \quad \text{if} \quad n = mk + i, \quad 1 \leq i \leq k.
\]
Then \( Q_k \) converges (pointwise or, equivalently, in variation) to the \( L \)-projection \( R \) on \( \mathcal{C} \).

**Proof.** Any linear set \( \mathcal{C} \) 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 \( \mathcal{C} \) is closed and can be defined by a finite number of linear constraints. In view of Theorem 2.1 and Corollary 3.1, the \( L \)-projection of \( R \) on such an \( \mathcal{C} \) always exists if \( I(P||R) < \infty \)—now equivalent to \( P \leq R \)—for some \( P \in \mathcal{C} \), and then (3.1) holds, as well.

Under the hypotheses of Theorem 3.2, it follows that the \( L \)-projections \( Q_0, Q_1, \ldots \) and \( Q \) exist and \( I(P||Q_n) = I(P||Q_{n-1}) + I(Q_{n-1}||Q_{n-1}) \) for any \( P \in \mathcal{C} \).

*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.*
\[ n = 1, 2, \ldots \] Setting \( P = Q \), in particular, we obtain by induction
\[ I(Q || R) = I(Q || Q_n) + \sum_{i=1}^{n} I(Q_i || Q_{i-1}) \]
\( n = 1, 2, \ldots \)

Since \( X \) is finite, each subsequence of \( Q_{n} \) contains a convergent subsequence; it suffices to show that \( Q_{n} \to Q' \) implies \( Q' = Q \). First verify \( Q' \in \mathcal{F} \). We have from (3.14)
\[ \sum_{i=1}^{\infty} I(Q_i || Q_{i-1}) \leq I(Q || R) < \infty \]
thus \( I(Q_n || Q_{n-1}) \to 0 \), implying \( |Q_n - Q_{n-1}| \to 0 \) by (2.3). Thus \( Q_{n+1}, \ldots, Q_{n+m} \) also converge to \( Q' \) as \( t \to \infty \). Since these \( \mathcal{D} \)'s belong to (a cyclic permutation of) the closed sets \( \mathcal{F}_0, \ldots, \mathcal{F}_n \) respectively, see (3.13), we conclude \( Q' \in \bigcap_{i=0}^{\infty} \mathcal{F}_i \).

Repeated application of Theorem 2.3 shows that \( Q \) is the \( \mathcal{D} \)-projection on \( \mathcal{F} \) of \( Q_1, \ldots, Q_n \), as well, thus
\[ I(P || Q_n) = I(P || Q) + I(Q || Q_n) \quad \text{for all} \quad P \in \mathcal{F}, \]
\( n = 1, 2, \ldots \). Applying this to \( P = Q' \) and \( Q = Q_n \), we obtain \( I(Q' || Q) = 0 \), i.e., \( Q' = Q_n \), because for finite \( X \), \( Q_n \to Q' \) implies \( I(Q' || Q_n) \to 0 \). The proof is complete.

Finally, let us return to the problem of existence of \( \mathcal{D} \)-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 \( Q \in \mathcal{F} \) with \( R \)-density (3.2). This is not easy but has been done under fairly general conditions by Cencov [2], Theorem 23.1. Here we show how the method of Theorem 2.1 applies to this case. Our hypothesis will be that
\[ T_n = \{(t_1, \ldots, t_k) : \exp \sum_{i=1}^{k} t_i f_i(x) \} \quad \text{is} \quad R \text{-integrable} \]
is open in \( E^n \); this clearly implies that \( f_i(x) \exp \sum_{j=1}^{k} t_j f_j(x) \) is \( R \)-integrable for every \( (t_1, \ldots, t_k) \in \mathcal{T}_n \), \( i = 1, \ldots, k \). (In the first version of this paper, \( T_n = E^n \) was assumed. The strengthening has been inspired by Cencov's result, loc. cit., which implies the existence of \( \mathcal{D} \)-projection even under a slightly weaker hypothesis.)

**Theorem 3.3.** Let \( a_1, \ldots, a_n \) be the set of points satisfying \( \int f_i dP = a_i \), \( i = 1, \ldots, k \) and let \( A_n \) be the set of points \( a_1, \ldots, a_n \in E^n \) for which \( \mathcal{D}(a_1, \ldots, a_n) \) contains some point with \( I(P || R) < \infty \). Then, supposing that \( T_n \) is open, the \( \mathcal{D} \)-projection of \( R \) on \( \mathcal{D}(a_1, \ldots, a_n) \) exists for each inner point \( a_1, \ldots, a_n \) of \( A_n \), and its \( R \)-density is of form (3.2).

Remarks. It can be shown that the interior of \( A_n \) coincides with that of the convex hull of the support of \( R' \), the image of \( R \) in \( E^n \) at the mapping \( x \to (f_1(x), \ldots, f_k(x)) \). Thus, assuming that the \( f_i \)'s are linearly independent mod \( R \), the interior of \( A_n \) is nonvoid. If \( a_1, \ldots, a_n \in A_n \) is on the boundary of \( A_n \), typically there still exists the \( \mathcal{D} \)-projection of \( R \) on \( \mathcal{D}(a_1, \ldots, a_n) \) 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 \( e^{f(x)} \) 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 sufficiently 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 \( A_n = \{ x : f(x) \leq K \} \) we have \( \int_a f dP_n \to \int_a f dQ \). Thus it suffices to show that to any \( \varepsilon > 0 \) there exists \( K \) such that
\[ \limsup_{n \to \infty} \int_{A_n} |f| dP_n = \limsup_{n \to \infty} \int_{A_n} |f| dP_n < \varepsilon. \]
But \( I(P_n || Q) = \int f \log P_n - dQ \to 0 \) implies \( \limsup_{n \to \infty} \int f \log P_n - dQ = 0 \) for every \( A \) in \( \mathcal{F} \) (apply Fatou's lemma to both \( A \) and \( A \setminus A \)). Choosing \( \varepsilon > 0 \) and \( K \) to satisfy \( \int_{A_n} e^{f(x)} dQ < \varepsilon \), (3.17) follows from the inequality \( ab < a \log a + e^{a} \) (see [1] Section 15), substituting \( a = p_n(x) \), \( b = f(x) \).

**Proof of Theorem 3.3.** On account of the convexity of \( I(P || R) \) in \( P, A_n \) is a convex set and
\[ F(a_1, \ldots, a_n) = \inf_{P \in \mathcal{P}_n(a_1, \ldots, a_n)} I(P || Q) \]
is a finite valued convex function on \( A_n \). Hence, if \( a_1, \ldots, a_n \) is an inner point of \( A_n \), there exists \( (t_1, \ldots, t_k) \) such that
\[ F(b_1, \ldots, b_k) \leq F(a_1, \ldots, a_n) + \sum_{i=1}^{k} t_i(b_i - a_i) \]
for all \( b_1, b_2, \ldots, b_k \in A_n \). First we show that \( (t_1, \ldots, t_k) \in T_n \), see (3.16). Let \( P_n \in \mathcal{P}_n(a_1, \ldots, a_n) \), \( I(P_n || R) \to F(a_1, \ldots, a_n) \); then \( P_n \) converges in variation to some \( Q \) by the proof of Theorem 2.1. Let \( f(x) = f_n(x) \) if \( t_n f_n < K_n \) and \( f(x) \equiv 0 \) else, where \( K_n \uparrow \infty \), and let \( Q_n \) be the PD with \( R \)-density (3.20)
\[ q_n(x) = c_n \exp \sum_{i=1}^{k} t_i f_i(x) \].
From (3.20) and (1.1) follows
\[ I(Q_n || R) = \lim \log q_n dP_n + \sum_{i=1}^{k} t_i f_i dQ_n - \int f_i dP_n \]
Since \( (0, \ldots, 0) \in T \), the \( P_n \)'s are \( R \)-integrable and thus \( Q_n \)-integrable, too; it follows that for large \( n \) \( q_n dP_n \) is arbitrarily close to \( \int f_i dQ_n = b_i \), say (note that the sequence \( c_n \) is non-increasing). Choosing the \( K \) properly, also \( b_i \) close to \( \int f_i dP_n = a_i \) if \( n \) is large, and then the identical (3.21) and (2.6) compared with the inequality (3.19) (with \( b_i \) in the role of \( b_0 \)) 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, \( (t_1, \ldots, t_k) \in T_n \).

Setting \( b_i = \frac{1}{2} f_i dQ \), similarly to (3.21) we have
\[ I(Q || R) = \int \log q_n dP_n + \sum_{i=1}^{k} t_i(b_i - a_i) \]
whence—again by (2.6) and (3.19)—we obtain \( I(P_n || Q) \to 0 \). Using the assumption that \( T_n \) is an open set, Lemma 3.1 gives \( \int f dQ = \lim_{n \to \infty} \int f dP_n = a_i \), \( i = 1, \ldots, k \). The proof is complete.
REMARK. It follows that $F(a_1, \ldots, a_i)$—see (3.18)—is differentiable at every inner point of $A_i$ and grad $F(a_1, \ldots, a_i) = (t_1, \ldots, t_i)$ is just the parameter vector in (3.2) for the 1-projection $Q$ of $R$ on $Z(a_1, \ldots, a_i)$.

REFERENCES


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