The Annals of Probability
1984, Vol. 12, No. 3, 768–793

SANOV PROPERTY, GENERALIZED *I*-PROJECTION AND A CONDITIONAL LIMIT THEOREM

BY IMRE CSISZÁR[1]

Mathematical Institute of the Hungarian Academy of Sciences, Budapest

Known results on the asymptotic behavior of the probability that the empirical distribution $\hat{P}_n$ of an i.i.d. sample $X_1, \cdots, X_n$ belongs to a given convex set Π of probability measures, and new results on that of the joint distribution of $X_1, \cdots, X_n$ under the condition $\hat{P}_n \in \Pi$ are obtained simultaneously, using an information-theoretic identity. The main theorem involves the concept of asymptotic quasi-independence introduced in the paper. In the particular case when $\hat{P}_n \in \Pi$ is the event that the sample mean of a V-valued statistic ψ is in a given convex subset of V, a locally convex topological vector space, the limiting conditional distribution of (either) X_i is characterized as a member of the exponential family determined by ψ through the unconditional distribution P_X, while $X_1, \cdots, X_n$ are conditionally asymptotically quasi-independent.

1. Introduction. Given a sequence $X_1, X_2, \cdots$ of independent random variables (rv's) distributed over an arbitrary measurable space $(S, \mathscr{B})$ with common distribution P_X, a set Π of probability measures (pm's) on $(S, \mathscr{B})$ is said to have the *Sanov property* if the empirical distribution $\hat{P}_n$ of $X^n = (X_1, \cdots, X_n)$ satisfies

$$\lim_{n\to\infty}(1/n)\log \Pr\{\hat{P}_n \in \Pi\} = -D(\Pi \,\|\, P_X). \tag{1.1}$$

Here the following notation is used:

$$D(\Pi \,\|\, Q) = \inf_{P\in\Pi} D(P \,\|\, Q), \tag{1.2}$$

where for any two pm's P and Q defined on the same measurable space

$$D(P \,\|\, Q) = \begin{cases} \int \log(dP/dQ)\, dP & \text{if } P \ll Q \\ +\infty & \text{else} \end{cases} \tag{1.3}$$

is the I-divergence or Kullback-Leibler information number, called briefly *divergence* in the sequel. The quantity (1.3) has also been called information for discrimination, cross-entropy, information gain, etc. The terminology adopted here is the one appearing most frequently in recent information theory literature, although it is at variance with the original terminology of Kullback and Leibler (1951) where "divergence" meant the Jeffreys divergence, i.e., the symmetrized

Received January 1983; revised December 1983.

[1] This work was done while the author was a visiting professor at Stanford University, partially supported by NSF grant ENG 79-08948.

AMS 1980 *subject classifications.* Primary 60F10, 60B10, 60B12; secondary 62B10, 94A17, 82A05.

Key words and phrases. Kullback-Leibler information, I-projection, large deviations in abstract space, exponential family, asymptotic quasi-independence, maximum entropy principle.

form of (1.3). A familiar alternative definition is

$$(1.4)\quad D(P \| Q) = \sup_{\mathscr{P}} D_{\mathscr{P}}(P \| Q); \quad D_{\mathscr{P}}(P \| Q) = \sum_{i=1}^{k} P(B_i)\log(P(B_i)/Q(B_i)),$$

where $\mathscr{P} = (B_1, \cdots, B_k)$ ranges over all finite (measurable) partitions of the underlying measurable space. Here one understands $0 \log 0 = 0 \log(0/0) = 0$, $a \log(a/0) = +\infty$ if $a > 0$; all logarithms in this paper are to the base e. For the equivalence of (1.3) and (1.4) and for the basic properties of divergence cf., e.g., Pinsker (1964).

The limit relation (1.1) is often referred to as Sanov's theorem. Papers establishing (1.1) under various conditions include Sanov (1957), Hoeffding (1965), Hoadley (1967), Borovkov (1967), Lanford (1972), Stone (1974), Donsker and Varadhan (1975–1976) (this highly sophisticated work deals with the Markov case), Bahadur and Zabell (1979) and Groeneboom, Oosterhoff and Ruymgaart (1979). For brevity, in the sequel we shall refer to the last two papers as BZ (1979) and GOR (1979). Familiarity with them or any of the other references is, however, not necessary for reading this paper.

There is a close relationship exhibited and exploited in our treatment, between the Sanov property and the limiting behavior of the conditional distribution of X_1 under the condition $\hat{P}_n \in \Pi$, as $n \to \infty$. The latter has been treated in the literature only when conditioning on empirical averages, i.e., on an event $(1/n) \sum_{i=1}^{n} f(X_i) \in I$ (where f is some real-valued function and I is an interval) or on the joint occurrence of a finite number of such events. This corresponds to letting Π be the set all pm's P satisfying $\int f\, dP \in I$, or the intersection of sets of this form. For such conditioning, Lanford (1973), Bártfai (1974), Vasicek (1980), Van Campenhout and Cover (1981) determined the limiting conditional distribution of X_1 under various assumptions, all of which implied the existence of a pm $P^* \in \Pi$ minimizing $D(P \| P_X)$ subject to $P \in \Pi$; the limiting conditional distribution of X_1 was equal to this P^*. Such results are relevant for statistical physics, since microcanonical distributions can be interpreted as conditional distributions of the mentioned kind. It has been argued, cf. Van Campenhout and Cover (1981), that such conditional limit theorems provide a justification of the "maximum entropy principle" in physics. Conditioning on exact values of empirical averages has also been considered, cf. Zabell (1980) and the literature cited there. Except for the discrete case, this problem requires stronger regularity conditions and may still not have direct implications for the previous one, cf. Van Campenhout and Cover (1981). Our generalization of these results will involve the concepts of generalized I-projection and convergence in information.

Let Q be a pm and Π be a convex set of pm's on $(S, \mathscr{B})$ such that $D(\Pi \| Q) < \infty$. As in Csiszár (1975), a pm P^* will be called the *I-projection* of Q on Π if $P^* \in \Pi$ and $D(P^* \| Q) = D(\Pi \| Q)$. This I-projection surely exists if Π is variation-closed, while in general there exists a pm P^* not necessarily in Π such that every sequence of pm's $P_n \in \Pi$ with $D(P_n \| Q) \to D(\Pi \| Q)$ converges to P^* in variation, cf. Csiszár (1975, Theorem 2.1 and Remark). The latter P^* will be called the *generalized I-projection* of Q on Π. Csiszár (1975, Theorem 2.2) proved that if P^*

is the I-projection of Q on Π then

$$D(P \| Q) \geq D(P \| P^*) + D(\Pi \| Q) \quad \text{for every} \quad P \in \Pi. \tag{1.5}$$

Topsoe (1979, Theorem 8) proved that (1.5) holds also if P^* is the generalized I-projection, called by him the "relative center of attraction"; moreover, (1.5) uniquely determines P^*. It follows from (1.5) that any sequence of pm's $P_n \in \Pi$ with $D(P_n \| Q) \to D(\Pi \| Q)$ *converges in information* to P^* in the sense of Csiszár (1962), i.e., $D(P_n \| P^*) \to 0$; this is a stronger property than convergence in variation, cf. Section 2. It is not possible, in general, to define a "closure" of Π such that every Q with $D(\Pi \| Q) < \infty$ had an I-projection on this closure equal to its generalized I-projection on Π, cf. Example 3.1. Further, while we always have

$$D(P^* \| Q) \leq D(\Pi \| Q) \tag{1.6}$$

by the lower semi-continuity of divergence, implied by (1.4), cf. Pinsker (1964), here the strict inequality is possible, cf. Example 3.2.

In this paper we will consider convex sets Π of pm's satisfying some slight additional conditions. These are fulfilled for the intersection of a finite number of sets of form $\{P: \int f\, dP \in I\}$ and also in other cases of interest. For such sets Π we will obtain a simple proof of the Sanov property, using an elementary information-theoretic identity. More importantly, our approach leads to the conclusion that the conditional distribution of (either) X_i under the condition $\hat{P}_n \in \Pi$ converges in information to the generalized I-projection P^* of P_X on Π. It will also turn out that $X_1, \cdots, X_n$ are *asymptotically quasi-independent* under the condition $\hat{P}_n \in \Pi$ in an information-theoretic sense, which has a very intuitive probabilistic meaning, cf. Definition 2.1 and the discussion thereafter. This appears highly relevant for the statistical physics problem hinted to above; results of this kind do not seem to have been published previously.

As a special case, we will consider the choice of Π which makes $\hat{P}_n \in \Pi$ equivalent to $(1/n) \sum_{i=1}^n \psi(X_i) \in C$ where ψ is a statistic with values in a topological vector space V and C is a convex subset of V. In this case we get more explicit results, based on representing the generalized I-projection of P_X on Π as a member of the exponential family determined by P_X and ψ.

2. Statement and discussion of the results. Let $(S, \mathscr{B})$ be an arbitrary measurable space, considered as fixed throughout this paper. We designate by Λ the set of all pm's on $(S, \mathscr{B})$ and by Λ_f its subset consisting of all atomic pm's with a finite number of atoms, i.e., $P \in \Lambda_f$ iff for every $B \in \mathscr{B}$

$$P(B) = \sum_{i=1}^k \alpha_i 1_B(s_i) \quad (s_i \in S,\ \alpha_i > 0,\ i = 1, \cdots, k;\ \sum_{i=1}^k \alpha_i = 1). \tag{2.1}$$

The *empirical distribution* of a sample $\mathbf{s} = (s_1, \cdots, s_n) \in S^n$ is the pm $\hat{P}_n(\mathbf{s}, \cdot) \in \Lambda_f$ defined by

$$\hat{P}_n(\mathbf{s}, B) = (1/n) \sum_{i=1}^k 1_B(s_i). \tag{2.2}$$

We consider as fixed also a sequence of independent S-valued rv's X_1, $X_2, \cdots$ with common distribution P_X. A formal definition of such rv's will not be needed, for we simply regard the nth Cartesian power of $(S, \mathscr{B}, P_X)$ as the

sample space of $X^n = (X_1, \cdots, X_n)$ and by probabilities of events determined in terms of X^n we formally mean P_X^n-measures of subsets of S^n where P_X^n is the nth Cartesian power of P_X. In particular, for any set of pm's $\Pi \subset \Lambda$ the probability that the empirical distribution $\hat{P}_n$ of $(X_1, \cdots, X_n)$ belongs to Π is, by definition,

$$\Pr\{\hat{P}_n \in \Pi\} = P_X^n(A_n); \quad A_n = \{\mathbf{s}: \hat{P}_n(\mathbf{s}, \cdot) \in \Pi\}. \tag{2.3}$$

The last probability is well defined if $A_n \in \mathscr{B}^n$. For conditions on $\Pi \subset \Lambda$ ensuring this for every n, cf. GOR (1979, Proposition 3.1). If $A_n \notin \mathscr{B}^n$, consider instead of (2.3) the upper and lower probabilities

$$\overline{\Pr}\{\hat{P}_n \in \Pi\} = P_X^n(\bar{A}_n), \quad \underline{\Pr}\{\hat{P}_n \in \Pi\} = P_X^n(\underline{A}_n) \tag{2.4}$$

where $\bar{A}_n \supset A_n$ and $\underline{A}_n \subset A_n$, respectively, are sets in $\mathscr{B}^n$ having minimum, respectively maximum, P_X^n-measure among all such sets. In this case the *Sanov property* is interpreted to mean that the limit relation (1.1) holds both for the upper and lower probabilities.

REMARK 2.1. If S is a completely regular topological space, $\mathscr{B}$ its Borel σ-algebra, and P_X satisfies the regularity condition

$$\lim P_X(G_\alpha) = P_X(\cup G_\alpha) \tag{2.5}$$

for every increasing net of open sets G_α then, by Csiszár (1970, Lemma 2), P_X^n can be uniquely extended to the Borel σ-algebra of S^n which, in general, is larger than $\mathscr{B}^n$, in such a way that both (2.5) and the Fubini theorem for evaluating integrals $\int f\, dP_X^n$ hold for the extended P_X^n and Borel-measurable functions f on S^n. This remark is relevant to our subject when S is a topological vector space, and one is interested in the event that the sample mean is in a given Borel subset of S, cf. BZ (1979). This event need not be in $\mathscr{B}^n$ but is obviously a Borel set in S^n. The regularity of a pm in the sense of BZ (1979, page 592) is a stronger condition than (2.5).

For any subset $A \in \mathscr{B}^n$ of S^n with $P_X^n(A) > 0$, we designate by $P_{X_i|A}$ and $P_{X^n|A}$ the conditional distribution of X_i and of $X^n = (X_1, \cdots, X_n)$, respectively, under the condition $X^n \in A$. Formally, $P_{X^n|A}$ is the pm defined on $(S, \mathscr{B})^n$ by

$$P_{X^n|A}(E) = P_X^n(E \cap A)/P_X^n(A) \quad (E \in \mathscr{B}^n) \tag{2.6}$$

and $P_{X_i|A}$ is its ith marginal. In the case when $P_{X_1|A} = \cdots = P_{X_n|A}$, this pm will be denoted simply by $P_{X|A}$. The definition of these conditional distributions is extended to the case $A \notin \mathscr{B}^n$ by setting $P_{X^n|A} = P_{X^n|\bar{A}}$ where $\bar{A} \in \mathscr{B}^n$ is such that $\bar{A} \supset A$ and $P_X^n(\bar{A})$ is minimum subject to these contraints. Clearly, this leads to an unambiguous definition whenever A has a positive outer P_X^n-measure. In particular, the conditional distributions

$$P_{X^n|\hat{P}_n \in \Pi} = P_{X^n|A_n}, \quad P_{X|\hat{P}_n \in \Pi} = P_{X|A_n} \tag{2.7}$$

—with A_n as in (2.3)—are well defined whenever $\overline{\Pr}\{\hat{P}_n \in \Pi\} > 0$.

DEFINITION 2.1. Let $X_{n1}, \cdots, X_{nn}$ be S-valued rv's with joint distribution $P^{(n)}$, $n = 1, 2, \cdots$, or let $X_1, X_2, \cdots$ be S-valued rv's and $A_n \subset S^n$ be sets with

$P_{X^n|A_n} = P^{(n)}$, $n = 1, 2, \cdots$. Then, if

$$\lim_{n\to\infty}(1/n)D(P^{(n)} \| Q^n) = 0 \tag{2.8}$$

for some pm $Q \in \Lambda$, we say that $X_{1n}, \cdots, X_{nn}$ are *asymptotically quasi-independent with limiting distribution* Q, or that $X_1, \cdots, X_n$ are *asymptotically quasi-independent under the condition* $X^n \in A_n$ *with limiting distribution* Q, respectively.

This terminology is motivated by the fact that whatever probabilistic statement holds, excepting an event of exponentially small probability, for i.i.d. rv's with common distribution Q, it holds with probability tending to 1 for $X_{1n}, \cdots, X_{nn}$ as in Definition 1 or, equivalently, with conditional probability tending to 1 for $X_1, \cdots, X_n$ given that $X^n \in A_n$. More precisely, it follows from (2.8) that for every $\alpha > 0$

$$\begin{gathered}\lim_{n\to\infty}P^{(n)}(B_n) = 0 \quad \text{if} \quad B_n \in \mathscr{B}^n, \\ Q^n(B_n) \le \exp(-\alpha n), \quad n = 1, 2, \cdots.\end{gathered} \tag{2.9}$$

In fact, the inequality $D_{\mathscr{P}}(P^{(n)} \| Q^n) \le D(P^{(n)} \| Q^n)$, cf. (1.4), applied to $\mathscr{P} = (B_n, S^n \backslash B_n)$ yields

$$P^{(n)}(B_n)\log\frac{P^{(n)}(B_n)}{Q^n(B_n)} + (1 - P^{(n)}(B_n))\log\frac{1 - P^{(n)}(B_n)}{1 - Q^n(B_n)} \le D(P^{(n)} \| Q^n);$$

hence (2.9) clearly follows if (2.8) holds. If in Definition 2.1 each X_{in} has the same distribution P_n or each X_i has the same conditional distribution P_n given that $X^n \in A_n$ $(i = 1, \cdots, n)$ then, by the inequality

$$D(P_n \| Q) \le (1/n)D(P^{(n)} \| Q^n), \tag{2.10}$$

Q is the limiting distribution for the individual rv's in the strong sense that $P_n \to Q$ in information, i.e., $D(P_n \| Q) \to 0$ as $n \to \infty$. (2.10) follows from the identity

$$D(P^{(n)} \| Q_1 \times \cdots \times Q_n) = D(P^{(n)} \| P_1 \times \cdots \times P_n) + \textstyle\sum_{i=1}^n D(P_i \| Q_i) \tag{2.11}$$

where $P_i \in \Lambda$, $Q_i \in \Lambda$, $i = 1, \cdots, n$ are arbitrary and $P^{(n)}$ is any pm on $(S, \mathscr{B})^n$ whose marginals are the P_i's. For a proof of (2.11), $P^{(n)} \ll P_1 \times \cdots \times P_n$ and $P_i \ll Q_i$, $i = 1, \cdots, n$ may be assumed, for else both sides are $+\infty$. Then

$$\frac{dP^{(n)}}{d(Q_1 \times \cdots \times Q_n)}(s_1, \cdots, s_n) = \frac{dP^{(n)}}{d(P_1 \times \cdots \times P_n)}(s_1, \cdots, s_n)\prod_{i=1}^n \frac{dP_i}{dQ_i}(s_i)$$

whence, taking logarithms and integrating with respect to $P^{(n)}$, (2.11) follows by (1.3). As another consequence of (2.11), we notice that if every k consecutive ones of the rv's in Definition 2.1 have the same joint distribution or conditional joint distribution given that $X^n \in A_n$, say $P_n^{(k)}$, then $P_n^{(k)} \to Q^k$ in information as $n \to \infty$.

Convergence in information is a stronger property than that in variation. In fact, the variation distance of any two pm's P and Q, i.e., the total variation

$|P - Q|$ of the signed measure $P - Q$, satisfies

$$|P - Q| \le \sqrt{2D(P \| Q)}, \tag{2.12}$$

cf. Csiszár (1967, Theorem 4.1). Further, if $P_n \to Q$ in information then

$$\int f\, dP_n \to \int f\, dQ \quad \text{if} \quad \int \exp(tf)\, dQ < \infty \quad \text{for} \quad |t| < \varepsilon \tag{2.13}$$

where $\varepsilon > 0$ is arbitrary, cf. Csiszár (1975, Lemma 3.1). Clearly, $P_n \to Q$ in variation is not sufficient for (2.13).

The results of GOR (1979) on the Sanov property are formulated in terms of what they call the τ-topology on Λ. This topology is defined by the basic neighborhoods of pm's $P \in \Lambda$ of form

$$U(P, \mathscr{P}, \varepsilon) = \{Q: |P(B_i) - Q(B_i)| < \varepsilon, i = 1, \cdots, k\} \tag{2.14}$$

where $\mathscr{P}$ ranges over all measurable partitions $\mathscr{P} = (B_1, \cdots, B_k)$ of $(S, \mathscr{B})$ and ε ranges over the positive numbers. We shall find it convenient to use a slightly weaker topology.

DEFINITION 2.2. The τ_0-topology on Λ is defined by the basic neighborhoods

$$\begin{aligned} &U_0(P, \mathscr{P}, \varepsilon) \\ &\quad = \{Q: |P(B_j) - Q(B_j)| < \varepsilon, Q(B_j) = 0 \text{ if } P(B_j) = 0, i = 1, \cdots, k\}. \end{aligned} \tag{2.15}$$

Notice that Λ_f is τ_0-open but not τ-open.

DEFINITION 2.3. A set of pm's $\Pi \subset \Lambda$ is *completely convex* if for every probability space $(\Omega, \mathscr{A}, \mu)$ and Markov kernel ν from $(\Omega, \mathscr{A})$ to $(S, \mathscr{B})$ such that $\nu(\omega, \cdot) \in \Pi$ for each $\omega \in \Omega$, the pm $\mu\nu$ defined by

$$\mu\nu(B) = \int \nu(\cdot, B)\, d\mu \quad (B \in \mathscr{B})$$

also belongs to Π. Further, a convex set of pm's $\Pi \subset \Lambda$ is *almost completely convex* if there exist completely convex subsets $\Pi_1 \subset \Pi_2 \subset \cdots$ of Π such that $\bigcup_{k=1}^{\infty} \Pi_k \supset \Pi \cap \Lambda_f$.

Notice that Λ_f is convex but, typically, not even almost completely convex. For an important class of almost completely but, in general, not completely convex sets of pm's cf. Lemma 4.3.

Now we can formulate our basic result.

THEOREM 1. *Let $\Pi \subset \Lambda$ be an almost completely convex set of pm's and, if $D(\Pi \| P_X) < \infty$, let P^* denote the generalized I-projection of P_X on Π. Then*

$$(1/n)\log \overline{\Pr}\{\hat{P}_n \in \Pi\} \le -D(\Pi \| P_X) \quad \textit{for every} \quad n \tag{2.16}$$

and in case $D(\Pi \| P_X) < \infty$ we have for each $\Pi' \subset \Pi$ with $\overline{\Pr}\{\hat{P}_n \in \Pi'\} > 0$

$$(1/n)\log \overline{\Pr}\{\hat{P}_n \in \Pi'\} \le -D(\Pi \| P_X) - (1/n)D(P_{X^n|\hat{P}_n \in \Pi'} \| P^{*n}). \tag{2.17}$$

If $\Pi' \subset \Pi$ *is such that*

$$D(\text{int}_{\tau_0}\Pi' \,\|\, P_X) = D(\Pi \,\|\, P_X) < \infty \tag{2.18}$$

then (Π *and*) Π' *has the Sanov property, and* $X_1, \cdots, X_n$ *are asymptotically quasi-independent under the condition* $\hat{P}_n \in \Pi'$ *with limiting distribution* P^*.

Theorem 1 will be proved in Section 4. The proof is remarkably simple, the main tool being the identity (2.11). Of course, (2.17) implies that $X_1, \cdots, X_n$ are asymptotically quasi-independent under the condition $\hat{P}_n \in \Pi'$ with limiting distribution P^* whenever $D(\Pi' \,\|\, P_X) = D(\Pi \,\|\, P_X) < \infty$ and Π' has the Sanov property; (2.18) is just a sufficient condition for the latter. The intuitive probabilistic implications of asymptotic quasi-independence have been discussed after Definition 2.1. We notice that (2.10) and (2.17) yield

$$D(P_{X|\hat{P}_n \in \Pi'} \,\|\, P^*) \leq -(1/n)\log \overline{\Pr}\{\hat{P}_n \in \Pi'\} - D(\Pi \,\|\, P_X), \tag{2.19}$$

relating the speed of the convergence in information $P_{X|\hat{P}_n \in \Pi'} \to P^*$ to the speed of convergence in the Sanov property.

The next two theorems characterize generalized I-projections on certain important kinds of sets of pm's as members of exponential families. They will be proved in Section 3 as consequences of Lemma 3.4, a generalization of Csiszár (1975, Theorem 3.1). Our final result, Theorem 4, proved in Section 4, is an application of Theorems 1 and 3 to sample means of a statistic taking values in a topological vector space. It comprises a large derivation theorem and a conditional limit theorem involving such statistics.

The relation of Theorems 1–4 to previous results will be discussed at the end of this section.

THEOREM 2. *If* $\Pi \subset \Lambda$ *is defined by*

$$\Pi = \left\{P: \int f_i \, dP \geq 0, \, i = 1, \cdots, k\right\} \tag{2.20}$$

where $f_1, \cdots, f_k$ *are given measurable functions on* $(S, \mathscr{B})$, *for a pm* $Q \in \Lambda$ *we have* $D = D(\Pi \,\|\, Q) < \infty$ *iff there exists a* $P \in \Pi$ *with* $P \ll Q$. *Then the generalized I-projection* P^* *of* Q *on* Π *has* Q*-density of form*

$$\frac{dP^*}{dQ} = \begin{cases} \exp\{D + \sum_{i=1}^k \vartheta_i^* f_i\} & \text{on } \{s: (f_1(s), \cdots, f_k(s)) \in M\} \\ 0 & \text{elsewhere} \end{cases} \tag{2.21}$$

where M *is a linear subspace of* R^k *and* $\vartheta^* = (\vartheta_1^*, \cdots, \vartheta_k^*) \in R_+^k$. (2.21) *holds with* $M = R^k$, *i.e.*, P^* *belongs to the exponential family* $\{P_\vartheta: \vartheta \in \Theta\}$ *defined by*

$$\begin{aligned} &\frac{dP_\vartheta}{dQ} = \frac{\exp \sum_{i=1}^k \vartheta_i f_i}{\int \exp(\sum_{i=1}^k \vartheta_i f_i) \, dQ}, \\ &\Theta = \left\{\vartheta = (\vartheta_1, \cdots, \vartheta_k): \int \exp(\textstyle\sum_{i=1}^k \vartheta_i f_i) \, dQ < \infty\right\} \end{aligned} \tag{2.22}$$

iff there exists a $P \in \Pi$ *with* $P \equiv Q$, *where* $\equiv$ *designates mutual absolute continuity.*

Under the last condition

$$D = D(\Pi \,\|\, Q) = \max_{\vartheta \in R_+^k} \left[-\log \int \exp\left(\sum_{i=1}^k \vartheta_i f_i\right) dQ \right] \tag{2.23}$$

where the maximum is attained iff $P_\vartheta = P^*$.

COROLLARY. *The statements of Theorem* 2 *hold also for* $\Pi = \{P: \int f_i \, dP = 0, i = 1, \cdots, k\}$, *with the only modification that* R_+^k *is to be replaced everywhere by* R^k.

Now let V be a locally convex topological vector space and consider pm's on the Borel σ-algebra of V. The *expectation* or resultant of such a pm P is defined by

$$E(P) = v_0 \quad \text{if} \quad \int \vartheta(\cdot) \, dP = \vartheta(v_0) \quad \text{for each} \quad \vartheta \in V' \tag{2.24}$$

if such a $v_0 \in V$ exists, while else $E(P)$ is undefined. Here V' is the set of all continuous linear functionals on V. A useful fact is

$$A \subset V \text{ compact, convex}, \quad P(A) = 1 \Rightarrow E(P) \text{ exists}, \quad E(P) \in A, \tag{2.25}$$

cf. Choquet (1969, page 115). The *support* of P is the set of those $v \in V$ which do not have a neighborhood of P-measure 0. Let Q be a *convex-tight* pm on V, i.e., such that there exist subsets K_n of V with the properties

$$K_1 \subset K_2 \subset \cdots, \quad \text{each compact and convex}, \quad Q(K_n) \to 1 \quad \text{as} \quad n \to \infty. \tag{2.26}$$

If V is a separable Fréchet or, in particular, Banach space then every pm on V is convex-tight.

THEOREM 3. *Given* V *and* Q *as above, let* $C \subset V$ *be a convex set whose interior has a nonvoid intersection with the convex hull of the support of* Q. *Then, with the notation*

$$\Pi(C) = \{P: E(P) \in C\}, \quad \Pi_0(C) = \{P: P(K_n) = 1 \text{ for some } n\} \cap \Pi(C) \tag{2.27}$$

for sets K_n *satisfying* (2.26), *we have*

$$D(\Pi(C) \,\|\, Q) = D(\Pi_0(\text{int } C) \,\|\, Q) = D < \infty \tag{2.28}$$

and the common generalized I-projection P^* *of* Q *on* $\Pi(C)$ *and* $\Pi_0(\text{int } C)$ *belongs to the exponential family* $\{P_\vartheta: \vartheta \in \Theta\}$ *defined by*

$$\frac{dP_\vartheta}{dQ} = \frac{\exp \vartheta(\cdot)}{\int \exp \vartheta(\cdot) \, dQ}, \quad \Theta = \left\{ \vartheta: \vartheta \in V', \int \exp \vartheta(\cdot) \, dQ < \infty \right\}. \tag{2.29}$$

Further,

$$D = \max_{\vartheta \in V'} \left[\inf_{v \in C} \vartheta(v) - \log \int \exp \vartheta(\cdot) \, dQ \right] \tag{2.30}$$

where the maximum is attained iff $P_\vartheta = P^*$.

REMARK 2.2. To compare Theorems 2 and 3, consider the former in the case $S = R^k$, $f_i(x_1, \cdots, x_k) = x_i - r_i$, $i = 1, \cdots, k$, where $r_1, \cdots, r_k$ are real constants. This is equivalent to the general case, even when setting $r_1 = \cdots = r_k = 0$, cf. the proof of Theorem 2. We get, in particular, that the generalized I-projection P^* of Q on Π belongs to the exponential family of pm's with Q-density

$$\frac{dP_\vartheta}{dQ}(x) = \frac{\exp \sum_{i=1}^k \vartheta_i x_i}{\int \exp(\sum_{i=1}^k \vartheta_i x_i)\, dQ} \tag{2.31}$$

iff a $P \in \Pi$ with $P \equiv Q$ exists; further, then

$$D(\Pi \,\|\, Q) = \max_{\vartheta \in R_+^k} \left[\sum_{i=1}^k \vartheta_i r_i - \log \int \exp(\sum_{i=1}^k \vartheta_i x_i)\, dQ \right] \tag{2.32}$$

where the maximum is attained iff $P_\vartheta = P^*$. Theorem 3, when specialized to the case $V = R^k$, $C = \{x\colon x_i \geq r_i,\ i = 1, \cdots, k\}$, gives the same results under the condition that the convex hull of the support of Q intersects the interior of C, a condition equivalent to

$$\int x_i\, dP > r_i, \quad i = 1, \cdots, k \quad \text{for some} \quad P \ll Q. \tag{2.33}$$

Since (2.33) implies the existence of a $P \in \Pi$ with $P \equiv Q$ but not conversely, we see that Theorem 2 is somewhat stronger than the corresponding special case of the more general Theorem 3. This is of advantage, e.g., in obtaining the corollary.

THEOREM 4. *Let ψ be a measurable mapping of $(S, \mathscr{B})$ into a locally convex topological vector space V, let Q be the image of P_X under ψ, and suppose that Q is convex-tight. Then, for any convex set $C \subset V$ whose interior has a nonvoid intersection with the support of Q, we have*

$$\lim_{n\to\infty} (1/n) \log \Pr\{(1/n) \textstyle\sum_{i=1}^k \psi(X_i) \in C\} = -D \tag{2.34}$$

where $D < \infty$ is given by (2.30). *Further, $X_1, \cdots, X_n$ are asymptotically quasi-independent under the condition*

$$(1/n) \textstyle\sum_{i=1}^n \psi(X_i) \in C \tag{2.35}$$

with limiting distribution P^*, where

$$\frac{dP^*}{dP_X} = c \exp \vartheta^*(\psi(\cdot)), \tag{2.36}$$

with $\vartheta^* \in V'$ attaining the maximum in (2.30).

In the "nonmeasurable case", i.e., when the set

$$A_n = \{(s_1, \cdots, s_n)\colon (1/n) \textstyle\sum_{i=1}^n \psi(s_i) \in C\}$$

is not in $\mathscr{B}^n$, (2.34) is interpreted to mean that the limit relation holds for both the upper and lower probabilities, cf. (2.4). The asymptotic quasi-independence assertion means, of course, that (2.8) holds for $P_{X^n|A_n}$ and P^* in the role of $P^{(n)}$

and Q. We notice that as Theorem 4 is a consequence of Theorem 1, the bound (2.19) holds for the present case. It now gives

$$D(P_{X|(1/n)\sum_{i=1}^n \psi(X_i)\in C} \,\|\, P^*) \le -(1/n)\log \overline{\Pr}\{(1/n)\sum_{i=1}^n \psi(X_i) \in C\} - D. \tag{2.37}$$

We conclude this section by discussing the relation of Theorems 1–4 to previous results.

The more interesting part of the basic Theorem 1 is the asymptotic quasi-independence assertion. Related results available in the literature, referred to in the Introduction, concern the case when the condition $\hat{P}_n \in \Pi$ represents a finite number of constraints on sample means; then, under various regularity hypotheses, the convergence of $P_{X|\hat{P}_n\in\Pi}$ to the I-projection of P_X on Π has been established. These results are generalized here in four directions:

(i) more general sets Π of pm's are considered
(ii) the I-projection of P_X on Π need not exist
(iii) a stronger kind of convergence $P_{X|\hat{P}_n\in\Pi} \to P^*$ is established, namely convergence in information, and its speed is related to the speed of convergence in the Sanov property, cf. (2.19)
(iv) perhaps most importantly, the rv's $X_1, \cdots, X_n$ under the condition $\hat{P}_n \in \Pi$ are shown to jointly behave, in a sense, like i.i.d. rv's with common distribution P^*.

Theorem 1 also contains the most general sufficient conditions known to us for a convex set of pm's to have the Sanov property. An interesting point is that the upper bound needed for the Sanov property holds for every n rather than asymptotically as $n \to \infty$. This bound (2.16) may be viewed as a general Chernoff bound. For the discrete case, (2.16) appears in Csiszár and Körner (1981, page 43). A result equivalent to a special case of (2.16) is the "multidimensional Berstein-Chernoff inequality" of Bártfai (1977, Theorem 2). The asymptotic upper bound

$$\lim \sup_{n\to\infty}(1/n)\log \overline{\Pr}\{\hat{P}_n \in \Pi\} \le -D(\Pi \,\|\, P_X)$$

has been established by GOR (1979, Lemmas 2.4 and 3.1A) for sets $\Pi \subset \Lambda$ closed in the τ-topology, cf. (2.14), but otherwise arbitrary. Its lower counterpart

$$\lim \inf_{n\to\infty}(1/n)\log \overline{\Pr}\{\hat{P}_n \in \Pi\} \ge -D(\Pi \,\|\, P_X) \tag{2.38}$$

was proved by GOR (1979, Lemma 2.1B) for $\Pi \subset \Lambda$ open in the τ-topology or, equivalently, for $\Pi \subset \Lambda$ satisfying $D(\Pi \,\|\, P_X) = D(\text{int}_\tau \Pi \,\|\, P_X)$. It follows similarly, cf. Lemma 4.1, that $D(\Pi \,\|\, P_X) = D(\text{int}_{\tau_0}\Pi \,\|\, P_X)$ is already sufficient for (2.38). Actually, hypothesis (2.18) of Theorem 1 has been imposed but for ensuring (2.38) with Π' in the role of Π. For an alternative approach yielding (2.38) in great generality cf. Bahadur, Zabell and Gupta (1980).

It has been known for a long time that if

$$\Pi = \left\{P: \int f_i \, dP = 0,\ i = 1, \cdots, k\right\} \tag{2.39}$$

as in the Corollary of Theorem 2 then, providing the exponential family (2.22)

contains a pm $P_{\vartheta^*} \in \Pi$, this P_{ϑ^*} is the I-projection of Q on Π. This is obvious from the identity

$$D(P \| Q) = D(P \| P_{\vartheta^*}) + D(P_{\vartheta^*} \| Q) \quad (P \in \Pi) \tag{2.40}$$

which can be considered as an analogue of Pythagoras' theorem, the divergence playing the role of squared Euclidean distance. Various special cases of (2.40) were used in statistical inference by Kullback (1959), and a systematic approach to geometric properties of pm's based on (2.40) was developed by Čencov (1972). The well-known exponential family representation of I-projections on Π as in (2.39) was extended to generalized I-projections by Jupp and Mardia (1983), generalizing a previous result of Topsoe (1979). Their result is, in our terminology, that if

$$\max_{\vartheta \in R^k} \left[-\log \int \exp\left(\sum_{i=1}^k \vartheta_i f_i\right) dQ \right] \tag{2.41}$$

is attained then this maximum equals $D(\Pi \| Q)$ and the generalized I-projection of Q on Π equals P_{ϑ^*} if $\vartheta = \vartheta^*$ attains the maximum in (2.41). The corollary of Theorem 2 shows that on the other hand, (2.41) is necessary for the generalized I-projection to belong to the exponential family (2.22), and it provides another necessary and sufficient condition for the latter, namely the existence of a $P \in \Pi$ with $P \equiv Q$. The corollary also provides a characterization of generalized I-projection when this condition is not fulfilled.

For the purpose of this paper, sets of pm's of form (2.39) are of minor interest for they typically do not meet the hypothesis (2.18) of Theorem 1. Characterizations of generalized I-projections on sets of pm's as in Theorems 2 and 3 have not been considered in the literature. Nevertheless, formula (2.32)—equivalent to (2.23)—has been proved by GOR (1979, Theorem 5.1) under a condition equivalent to (2.33). Further, BZ (1979, Theorem 3.2 and Theorem 3.3d) proved a minimax counterpart of formula (2.30) for open convex sets $C \subset V$, namely, with our notation, that

$$D(\Pi(C) \| Q) = \inf_{v \in C} \sup_{\vartheta \in V'} \left[\vartheta(v) - \log \int \exp \vartheta(\cdot)\, dQ \right], \tag{2.42}$$

under a slightly stronger hypothesis on Q than ours. They did not assume that C intersected the convex hull of the support of Q; notice, however, that under the hypothesis of BZ (loc. cit) formula (2.30) trivially holds if C is disjoint from the convex hull of the support of Q, since then C can be separated from the latter by a hyperplane. We believe it is intuitively suggestive to regard formulas for $D(\Pi \| Q)$ as corollaries of the exponential family representation of generalized I-projection. This has also lead to a simple proof of the identity (2.28) which was, in a sense, a missing link between the approaches of GOR (1979) and BZ (1979) to large deviation theorems, cf. the next paragraph.

A general large deviation theorem for empirical means such as (2.34) has been established by BZ (1979, Theorems 2.3 and 3.2) and GOR (1979, Corollary 4.2). The former authors proved (2.34) for open convex sets $C \subset V$ with the alternative

formula (2.42) for D. In the finite-dimensional case the same was proved also by Bártfai (1977, Theorem 1) who considered nonconvex open sets C, too. GOR (loc. cit.) used effectively the hypotheses of Theorem 4 and identified the limit in (2.34) as $D = D(\Pi_0(C) \parallel Q)$ (with our notation, cf. (2.27)); they did not show that this result was equivalent to that of BZ (1979). In the special case

$$V = R^k, \quad \psi = (f_1, \cdots, f_k), \quad C = \{(x_1, \cdots, x_k): x_i \geq r_i, i = 1, \cdots, k\}$$

the hypothesis of Theorem 4 reduces to (2.33), with $Q = P_X$. Then (2.36) gives

$$\frac{dP^*}{dP_X} = c \exp \sum_{i=1}^k \vartheta_i^* f_i \quad \text{with} \quad \vartheta_i^* \geq 0, \quad i = 1, \cdots, k,$$

and Theorem 4 gives the "d-dimensional Chernoff theorem" of GOR (1979, Theorem 5.1). The main new contribution of Theorem 4 is the asymptotic quasi-independence assertion. It considerably extends previous limit theorems on the conditional distribution of X_1, referred to in the Introduction, while retaining the exponential family representation of the limiting distribution P^*. The bound (2.37) may also be of interest. Results of this kind have apparently not been published previously.

3. Generalized I-projection. Given a pm $Q \in \Lambda$ write

$$\Lambda_Q = \{P: P \in \Lambda, D(P \parallel Q) < \infty\}. \tag{3.1}$$

LEMMA 3.1. *Let P^* be the generalized I-projection of Q on a convex set of pm's $\Pi \subset \Lambda$ with $D(\Pi \parallel Q) < \infty$, and let $P' \ll Q$, $P' \neq P^*$ be arbitrary. Then*

$$D(\Pi \parallel Q) = \inf_{P \in \Pi \cap \Lambda_Q} \int \log \frac{dP^*}{dQ}\, dP > \inf_{P \in \Pi \cap \Lambda_Q} \int \log \frac{dP'}{dQ}\, dP. \tag{3.2}$$

In particular, for any measurable function f on $(S, \mathscr{B})$ such that $\int f\, dP \geq 0$ for every $P \in \Pi \cap \Lambda_Q$, we have

$$D(\Pi \parallel Q) \geq -\log \int \exp f\, dQ. \tag{3.3}$$

If here the equality holds then

$$\frac{dP^*}{dQ} = \left[\int \exp f\, dQ\right]^{-1} \exp f. \tag{3.4}$$

PROOF. On account of (1.5), we have for every $P \in \Pi \cap \Lambda_Q$

$$D(\Pi \parallel Q) \leq D(P \parallel Q) - D(P \parallel P^*) = \int \log \frac{dP^*}{dQ}\, dP \leq D(P \parallel Q)$$

which proves the equality in (3.2). Further, as (1.5) uniquely determines P^*, in case $P' \neq P^*$ there exists $P \in \Pi \cap \Lambda_Q$ such that $D(P \parallel Q) < D(P \parallel P') + D(\Pi \parallel Q)$.

If here $P' \ll Q$, it follows that

$$D(\Pi \,\|\, Q) > D(P \,\|\, Q) - D(P \,\|\, P') = \int \log \frac{dP'}{dQ}\, dP,$$

completing the proof of (3.2). To prove (3.3), we may suppose that $\int \exp f\, dQ < \infty$. Then, applying (3.2) to the pm P' with Q-density $[\int \exp f\, dQ]^{-1} \exp f$, we obtain

$$D(\Pi \,\|\, Q) \geq \inf_{P \in \Pi \cap \Lambda_Q} \int \log \frac{dP'}{dQ}\, dP = -\log \int \exp f\, dQ + \inf_{P \in \Pi \cap \Lambda_Q} \int f\, dP$$

where the inequality is strict unless $P' = P^*$. As $\int f\, dP \geq 0$ for every $P \in \Pi \cap \Lambda_Q$ by assumption, the proof is complete. □

LEMMA 3.2. *For convex subsets $\Pi' \subset \Pi$ of Λ, $D(\Pi \,\|\, Q) = D(\Pi' \,\|\, Q) < \infty$ implies that the generalized I-projections of Q on Π and Π' are the same.*

PROOF. Obvious from the definition of generalized I-projection. □

We notice that the converse implication is false, cf. Example 3.2.

LEMMA 3.3. *Let ψ be a measurable mapping of $(S, \mathscr{B})$ into another measurable space $(V, \mathscr{C})$ and let Π be the set of all pm's $P \in \Lambda$ whose image under ψ belongs to a given convex set $\tilde{\Pi}$ of pm's on $(V, \mathscr{C})$. Then for arbitrary $Q \in \Lambda$ and its ψ-image $\tilde{Q}$ we have*

$$D(\Pi \,\|\, Q) = D(\tilde{\Pi} \,\|\, \tilde{Q}). \tag{3.5}$$

If $D(\Pi \,\|\, Q) < \infty$, the generalized I-projections P^* of Q on Π and $\tilde{P}^*$ of $\tilde{Q}$ on $\tilde{\Pi}$ are related by

$$\frac{dP^*}{dQ}(s) = \frac{d\tilde{P}^*}{d\tilde{Q}}(\psi(s)) \quad [Q]. \tag{3.6}$$

PROOF. It follows from (1.4) that $D(P \,\|\, Q) \geq D(\tilde{P} \,\|\, \tilde{Q})$ for each $P \in \Lambda$ and its ψ-image $\tilde{P}$. Further, for any $\tilde{P} \in \tilde{\Pi}$ with $D(\tilde{P} \,\|\, \tilde{Q}) < \infty$ the pm $P \in \Lambda$ determined by

$$\frac{dP}{dQ}(s) = \frac{d\tilde{P}}{d\tilde{Q}}(\psi(s))$$

has ψ-image $\tilde{P}$ and it satisfies $D(P \,\|\, Q) = D(\tilde{P} \,\|\, \tilde{Q})$ by (1.3). This proves (3.5), and implies that if $\tilde{P}_n \in \tilde{\Pi}$, $D(\tilde{P}_n \,\|\, \tilde{Q}) \to D(\tilde{\Pi} \,\|\, \tilde{Q})$ then the pm's $P_n \in \Pi$ determined by $(dP_n/dQ)(s) = (d\tilde{P}_n/d\tilde{Q})(\psi(s))$ satisfy $D(P_n \,\|\, Q) \to D(\tilde{\Pi} \,\|\, \tilde{Q}) = D(\Pi \,\|\, Q)$. Hence (3.6) follows by the definition of generalized I-projection. □

EXAMPLE 3.1. As in Csiszár (1962, page 154), consider pm's P, Q_m and

$R_{m,n}$ $(1 < m < n)$ on $\mathbb{N}$, the positive integers, with probability mass functions

$$P(i) = 2^{-i}, \quad q_m(i) = \begin{cases} p_i & \text{if } i < m \\ c_m i^{-3} & \text{if } i \geq m, \end{cases} \quad r_{m,n}(i) = \begin{cases} q_m(i) & \text{if } i < n \\ c_{m,n} i^{-2} & \text{if } i \geq n. \end{cases}$$

Further, let Π be the set of pm's on $\mathbb{N}$ with expectation ≥ 3, possibly infinite. Then $R_{m,n} \in \Pi$ and $D(R_{m,n} \| Q_m) \to 0$ as $n \to \infty$. Hence, if generalized I-projections on Π could be represented as I-projections on some "closure" Π^* of Π, we certainly had $Q_m \in \Pi^*$, $m = 1, 2, \cdots$. On the other hand, P has I-projection P^* on Π with $p^*(i) = \frac{1}{3}(\frac{2}{3})^i$, and $D(Q_m \| P) \to 0$ as $m \to \infty$ rules out the equality of this P^* to the I-projection of P on the hypothetical Π^*.

EXAMPLE 3.2. Let S be the positive half-line, let Q be the pm with distribution function $1 - (s + 1)^{-3}e^{-s}$, i.e., with density function

$$q(s) = \frac{s + 4}{(s + 1)^4} e^{-s}.$$

Let Π_a be the set of all pm's P on S with expectation $\int s \, dP \geq a$. Consider the pm's $P_n \in \Pi_a$ with density functions

$$p_n(s) = \begin{cases} c_n \exp(t_n s) q(s), & 0 \leq s \leq n \\ 0, & s > n \end{cases}$$

where c_n and t_n are determined from the conditions $\int p_m \, ds = 1$, $\int s p_m \, ds = a$. Let $a \geq \frac{2}{3}$. Then $t_n \geq 1$, $t_n \to 1$ as $n \to \infty$, thus

$$D(P_n \| Q) = \log c_n + \int t_n s p_n \, ds = -\log \int_0^n \exp(t_n s) q \, ds + t_n a$$

$$\leq -\log \int_0^n e^s q \, ds + a t_n \to -\log \int_0^\infty e^s q \, ds + a = a - \log \tfrac{3}{2}.$$

Hence $D(\Pi_a \| Q) \leq a - \log \frac{3}{2}$. Further, let P^* be the element with parameter $t = 1$ of the exponential family of pm's with densities of from $c \exp(ts)q(s)$, i.e., the pm with density

$$p^*(s) = \frac{2}{3} \frac{s + 4}{(s + 1)^4}.$$

Then

$$\int \log \frac{dP^*}{dQ} \, dP = \log \frac{2}{3} + \int s \, dP \geq a - \log \frac{3}{2} \quad \text{for every} \quad P \in \Pi_a.$$

This proves by Lemma 3.1 that P^* is the generalized I-projection of Q on Π_a, and $D(\Pi_a \| Q) = a - \log \frac{3}{2}$ $(a \geq \frac{2}{3})$. This example shows that for convex sets $\Pi' \subset \Pi$ it may happen that Q has the same generalized I-projection on Π and Π' and still $D(\Pi' \| Q) > D(\Pi \| Q)$, and also that in (1.6) the strict inequality is possible.

Let $\mathscr{F}$ be a family of real-valued measurable functions on $(S, \mathscr{B})$. We designate by $\Pi(\mathscr{F})$ and $\Pi'(\mathscr{F})$ the set of all pm's $P \in \Lambda$ for which the integrals $\int f\,dP$ exist and are nonnegative, respectively positive, for each $f \in \mathscr{F}$:

$$\Pi(\mathscr{F}) = \left\{P: \int f\,dP \geq 0, \quad f \in \mathscr{F}\right\}$$
$$\Pi'(\mathscr{F}) = \left\{P: \int f\,dP > 0, \quad f \in \mathscr{F}\right\}. \tag{3.7}$$

For a subset $K \in \mathscr{B}$ of S, $\Pi(\mathscr{F} \mid K)$ and $\Pi'(\mathscr{F} \mid K)$ will designate the subsets of $\Pi(\mathscr{F})$ and $\Pi'(\mathscr{F})$ consisting of the pm's with $P(K) = 1$. Further, if $\mathscr{K} = \{K_i\}_{i=1}^{\infty}$ is a sequence of sets such that

(3.8) $K_i \in \mathscr{B}$, $K_i \subset K_{i+1}$, each $f \in \mathscr{F}$ is bounded on K_i, $i = 1, 2, \cdots$

we write

$$\Pi(\mathscr{F} \mid \mathscr{K}) = \cup_{i=1}^{\infty} \Pi(\mathscr{F} \mid K_i), \quad \Pi'(\mathscr{F} \mid \mathscr{K}) = \cup_{i=1}^{\infty} \Pi'(\mathscr{F} \mid K_i). \tag{3.9}$$

If to a given $Q \in \Lambda$ there exists a $P_0 \in \Pi'(\mathscr{F})$ or $P_0 \in \Pi'(\mathscr{F} \mid \mathscr{K})$, respectively, such that $D(P_0 \| Q) < \infty$ then

$$D(\Pi'(\mathscr{F}) \| Q) = D(\Pi(\mathscr{F}) \| Q),$$
$$D(\Pi'(\mathscr{F} \mid \mathscr{K}) \| Q) = D(\Pi(\mathscr{F} \mid \mathscr{K}) \| Q). \tag{3.10}$$

This follows from the facts that for any $P_1 \in \Pi(\mathscr{F})$ $(\Pi(\mathscr{F} \mid \mathscr{K}))$ we have $P_\alpha = (1 - \alpha)P_0 + \alpha P_1 \in \Pi'(\mathscr{F})$ $(\Pi'(\mathscr{F} \mid \mathscr{K}))$ if $0 \leq \alpha \leq 1$ and, by convexity,

$$\begin{aligned} \limsup_{\alpha \to 1} D(P_\alpha \| Q) &\leq \lim_{\alpha \to 1}[(1 - \alpha)D(P_0 \| Q) + \alpha D(P_1 \| Q)] \\ &= D(P_1 \| Q). \end{aligned} \tag{3.11}$$

The main tool to the proof of Theorems 2 and 3 is the following generalization of Csiszár (1975, Theorem 3.1).

LEMMA 3.4. *Suppose that $\mathscr{F}$ is a convex cone, i.e., $f_i \in \mathscr{F}$, $\alpha_i \geq 0$, $i = 1, \cdots, n$ implies*

$$\textstyle\sum_{i=1}^{n} \alpha_i f_i \in \mathscr{F}.$$

a. *If $Q \in \Lambda$ is a pm with $D(\Pi(\mathscr{F}) \| Q) = D < \infty$ whose I-projection P^* on $\Pi(\mathscr{F})$ exists then $\log(dP^*/dQ) - D$ belongs to the $L_1(P^*)$-closure of $\mathscr{F}$.*

b. *If $Q \in \Lambda$ is a pm with $D(\Pi(\mathscr{F} \mid \mathscr{K}) \| Q) = D < \infty$ and P^* is its generalized I-projection on $\Pi(\mathscr{F} \mid \mathscr{K})$ then there exists a sequence of functions $f_n \in \mathscr{F}$ such that*

$$\log \frac{dP^*}{dQ} = D + \lim_{n \to \infty} f_n \qquad [P^*]. \tag{3.12}$$

PROOF. a. Let $\mathscr{F}_1$ be the set of functions of form $f + g$, $f \in \mathscr{F}$, $g \geq 0$, g

bounded. We first show that if

$$P^* \in \Pi(\mathscr{F}), \quad D(P^* \| Q) = D, \quad \varphi = \log \frac{dP^*}{dQ} - D \tag{3.13}$$

then φ belongs to the $L_1(P^*)$-closure of $\mathscr{F}_1$. In fact, else φ could be separated from the convex cone $\mathscr{F}_1$ by a hyperplane in $L_1(P^*)$, i.e.,

$$\int \varphi h \, dP^* < \inf_{f \in \mathscr{F}_1} \int fh \, dP^* = 0 \tag{3.14}$$

would hold for some $h \in L_\infty(P^*)$. As $\mathscr{F}_1$ contains the nonnegative bounded functions, here necessarily $h \geq 0$ P^*-a.e., and h can be chosen to satisfy $\int h \, dP^* = 1$. Since $\mathscr{F} \subset \mathscr{F}_1$, (3.14) implies that the pm P_0 with $dP_0/dP^* = h$ belongs to $\Pi(\mathscr{F})$. We also have $P_0 \in \Lambda_Q$ since $P^* \in \Lambda_Q$ and h is bounded. Thus (3.2) gives

$$\int \varphi h \, dP^* = \int \left(\log \frac{dP^*}{dQ} - D \right) dP_0 \geq 0,$$

a contradiction to (3.14).

By what we have proved, there exist functions $f_n \in \mathscr{F}$ and $g_n \geq 0$, g_n bounded, such that

$$f_n + g_n \to \varphi \quad \text{in} \quad L_1(P^*). \tag{3.15}$$

By (3.13) we have $\int f_n \, dP^* \geq 0$, $\int \varphi \, dP^* = 0$. As (3.15) implies

$$\int f_n \, dP^* + \int g_n \, dP^* \to \int \varphi \, dP^*,$$

it follows that $\| g_n \|_{L_1(P^*)} = \int g_n \, dP^* \to 0$. This and (3.15) complete the proof of part a.

b. For n so large that $D_n = D(\Pi(\mathscr{F} \mid K_n) \| Q) < \infty$, let P_n^* denote the I-projection of Q on $\Pi(\mathscr{F} \mid K_n)$. This I-projection exists, since (3.8) implies that $\Pi(\mathscr{F} \mid K_n)$ is variation-closed. By part a, $\varphi_n = \log(dP_n^*/dQ) - D_n$ belongs to the $L_1(P_n^*)$-closure of $\mathscr{F}$, in particular, there exists $f_n \in \mathscr{F}$ such that

$$|\varphi_n - f_n| < 1/n \quad \text{except for a set} \quad A_n \quad \text{with} \quad P_n^*(A_n) < 1/n. \tag{3.16}$$

Further, as by (3.9)

$$P_n^* \in \Pi(\mathscr{F} \mid \mathscr{K}), \quad D(P_n^* \| Q) = D_n \to D = D(\Pi(\mathscr{F} \mid \mathscr{K}) \| Q),$$

we have $| P_n^* - P^* | \to 0$ by the definition of generalized I-projection. In particular, $P_n^*(A_n) \to 0$ implies $P^*(A_n) \to 0$, hence from (3.16)

$$\varphi_n - f_n \to 0 \quad \text{in } P^*\text{-measure.} \tag{3.17}$$

On the other hand, from $| P_n^* - P^* | \to 0$ and $D_n \to D$ it follows that

$$\varphi_n \to \varphi = \log \frac{dP^*}{dQ} - D \quad \text{in } P^*\text{-measure.} \tag{3.18}$$

(3.17) and (3.18) show that $f_n \to \varphi$ in P^*-measure. Since any sequence of

functions converging in measure has an a.e. convergent subsequence, this proves (3.12). □

PROOF OF THEOREM 2. It suffices to prove the theorem in the special case $S = R^k$, $f_i(x_1, \cdots, x_k) = x_i$, $i = 1, \cdots, k$, when

$$\Pi = \left\{P: \int x_i \, dP \geq 0, \, i = 1, \cdots, k\right\}. \tag{3.19}$$

In fact, in the general case consider the mapping $\psi: S \to R^k$ defined by $\psi(s) = (f_1(s), \cdots, f_k(s))$. Then (2.20) is the set of those pm's $P \in \Lambda$ whose ψ-image belongs to (3.19), and Lemma 3.3 gives that if Theorem 2 is true in the mentioned special case then it is true in general.

We henceforth consider Π as in (3.19) and a pm Q on R^k such that $P \ll Q$ for some $P \in \Pi$. Let Π_0 be the subset of Π consisting of pm's with bounded support, i.e.,

$$\Pi_0 = \left\{P: P(K_n) = 1 \text{ for some } n; \int x_i \, dP \geq 0, \, i = 1, \cdots, k\right\} \tag{3.20}$$

$$\text{where} \quad K_n = \{x: |x_i| \leq n, \, i = 1, \cdots, k\}.$$

We will prove that

$$D_0 = D(\Pi_0 \| Q) < \infty \tag{3.21}$$

and characterize the generalized I-projection P_0^* of Q on Π_0. Then the proof will be completed by checking that $D(\Pi \| Q) = D(\Pi_0 \| Q)$, $P^* = P_0^*$.

The set (3.20) of pm's on R^k is of form $\Pi(\mathscr{F} \mid \mathscr{K})$, cf. (3.9), with the convex cone $\mathscr{F}$ of linear functions $f(x) = \sum_{i=1}^k \vartheta_i x_i$, $\vartheta = (\vartheta_1, \cdots, \vartheta_k) \in R_+^k$, and with the sequence of "cubes" K_n defined in (3.20). As the a.e. limit of a sequence of functions in this $\mathscr{F}$ is itself a.e. equal to a function in $\mathscr{F}$, Lemma 3.4 gives that

$$\log \frac{dP_0^*}{dQ}(x) = D_0 + \sum_{i=1}^k \vartheta_i^* x_i \quad [P^*] \quad \text{with} \quad \vartheta^* \in R_+^k, \tag{3.22}$$

supposing that (3.21) holds.

Let M designate the smallest linear subspace of R^k with the property

$$P(M) = 1 \quad \text{for every} \quad P \in \Pi \quad \text{with} \quad P \ll Q. \tag{3.23}$$

Then, of course, $Q(M) > 0$. We are going to show that, for every n, the set of pm's on R^k with the properties

$$P \ll Q, \quad \frac{dP}{dQ} \quad \text{bounded},$$

$$M \cap K_n \subset \left\{x: \frac{dP}{dQ}(x) > 0\right\} \subset M \cap K_m \quad \text{for some} \quad m \geq n \tag{3.24}$$

contains some $P_n \in \Pi$. Since then $P_n \in \Pi_0$ and $D(P_n \| Q) < \infty$, this will prove (3.21). Further, as (1.5) then implies $P_n \ll P_0^*$, $n = 1, 2, \cdots$, hence and from

(3.23) it will also follow that

$$\frac{dP_0^*}{dQ} \text{ is positive exactly on } M \quad [Q]. \tag{3.25}$$

Let us denote by E_n the set of expectation vectors of all pm's with the properties (3.24). We have to prove that $E_n \cap R_+^k \neq \emptyset$. Let $F \subset M$ be the closed convex hull of the support of the restriction of Q to M, and let F_0 be the interior of F relative to its affine hull. Clearly, $E_n \subset F$ and E_n is dense in F. Since E_n is convex, it follows that $E_n \supset F_0$. Hence it suffices to show that $F_0 \cap R_+^k \neq \emptyset$. Supposing the contrary, F_0 and $M \cap R_+^k$ could be separated by a (dim $M - 1$)-dimensional linear subspace M_1 of M, thus a pm $P \ll Q$ with $P(M) = 1$ could not belong to Π unless its support were contained in M_1. This contradicts to the minimality of M subject to (3.23). According to the previous paragraph, thereby we have established (3.21) and (3.25).

Now we show that $D = D(\Pi \,\|\, Q)$ equals $D_0 = D(\Pi_0 \,\|\, Q)$ and the generalized I-projection P^* of Q on Π equals P_0^*. By (3.23) and (3.25), every $P \in \Pi$ with $P \ll Q$ satisfies $P \ll P_0^*$. Hence from (3.22) and (3.19) we obtain for every such P

$$\int \log \frac{dP_0^*}{dQ}\, dP = D_0 + \sum_{i=1}^k \vartheta_i^* \int x_i\, dP \geq D_0. \tag{3.26}$$

By Lemma 3.1, the infimum for all $P \in \Pi$ with $D(P \,\|\, Q) < \infty$ of the first integral in (3.26) cannot exceed D, and would be strictly less than D were P_0^* not equal to P^*. Since $D \leq D_0$ by definition, this proves the asserted equalities. Thus we have proved, cf. (3.22), (3.25), that

$$\frac{dP^*}{dQ}(x) = \begin{cases} \exp\{D + \sum_{i=1}^k \vartheta_i^* x_i\} & \text{if } x \in M \\ 0 & \text{else} \end{cases} \quad \text{where } \vartheta^* \in R_+^k. \tag{3.27}$$

If there exists a $P \in \Pi$ with $P \equiv Q$ then (3.23) implies $Q(M) = 1$, hence (3.27) holds with $M = R^k$. In this case P^* belongs to the exponential family $\{P_\vartheta\colon \vartheta \in \Theta\}$ where

$$\frac{dP_\vartheta}{dQ}(x) = \frac{\exp \sum_{i=1}^k \vartheta_i x_i}{\int \exp(\sum_{i=1}^k \vartheta_i x_i)\, dQ},$$

$$\Theta = \left\{\vartheta\colon \int \exp(\textstyle\sum_{i=1}^k \vartheta_i x_i)\, dQ < \infty\right\}. \tag{3.28}$$

We notice that (3.27) gives

$$D = -\log \int \exp(\textstyle\sum_{i=1}^k \vartheta_i^* x_i)\, dQ \quad \text{if } P_{\vartheta^*} = P^*. \tag{3.29}$$

Finally, applying (3.3) to $f(x) = \sum_{i=1}^k \vartheta_i x_i$ with $\vartheta \in R_+^k$, we see that

$$D \geq -\log \int \exp(\textstyle\sum_{i=1}^k \vartheta_i x_i)\, dQ \quad \text{for each } \vartheta \in R_+^k,$$

and this inequality is strict unless $P_\vartheta = P^*$. This and (3.29) prove that if $P \equiv Q$

for some $P \in \Pi$ then

$$D = \max_{\vartheta \in R_+^k} \left[-\log \int \exp\left(\sum_{i=1}^k \vartheta_i x_i\right) dQ \right], \tag{3.30}$$

where the maximum is attained iff $P_\vartheta = P^*$.

This completes the proof of Theorem 2 for the case $S = R^k$, $f_i(x_1, \cdots, x_k) = x_i$, $i = 1, \cdots, k$. According to the first paragraph of the proof, thereby Theorem 2 is proved in general. The corollary follows from the theorem simply by applying it to the $2k$ functions $f_1, -f_1, \cdots, f_k, -f_k$. □

PROOF OF THEOREM 3. We first prove that

$$D(\Pi_0(\text{int } C) \,\|\, Q) = D < \infty, \quad P^* \equiv Q, \tag{3.31}$$

where P^* denotes the generalized I-projection of Q on $\Pi_0(\text{int } C)$. By assumption, some $v_0 \in \text{int } C$ belongs to the convex hull of the support of Q, i.e.,

$$v_0 = \sum_{i=1}^k \alpha_i v_i, \quad \alpha_i > 0, \quad i = 1, \cdots, k, \quad \sum_{i=1}^k \alpha_i = 1, \tag{3.32}$$

where $Q(v_i + N) > 0$, $i = 1, \cdots, k$ for each 0-neighborhood N in V. Pick a convex and closed 0-neighborhood N such that $v_0 + N \subset \text{int } C$, and let n_0 be so large that the compact, convex sets $A_i = (v_i + N) \cap K_{n_0}$, $i = 1, \cdots, k$ have positive Q-measure, where the compact convex sets $K_1 \subset K_2 \subset \cdots$ with $Q(K_n) \to 1$ are those appearing in (2.27). Then for the pm

$$P_0 = \sum_{i=1}^k \alpha_i R_i \quad \text{with} \quad R_i(\cdot) = Q(\cdot \cap A_i)/Q(A_i)$$

we have for $n \geq n_0$

$$E(P_0) \in \text{int } C, \quad D(P_0 \,\|\, Q) < \infty, \quad P_0(K_n) = 1, \tag{3.33}$$

where the first property follows from (2.25), (3.32) and the condition $v_0 + N \subset \text{int } C$. This already proves the first part of (3.31). Further, as the pm's P_n defined by

$$P_n = (1 - \alpha_n)P_0 + \alpha_n Q_n, \quad Q_n(\cdot) = Q(\cdot \cap K_n)/Q(K_n), \quad m \geq n_0$$

with sufficiently small $\alpha_n > 0$ also satisfy (3.33), we have

$$Q_n \ll P_n \ll P^* \ll Q \quad \text{for every} \quad n \geq n_0.$$

As $Q(K_n) \to 1$, this proves the second part of (3.31).

Since the assertions of Theorem 3 are invariant under translations, we henceforth assume that $0 \in \text{int } C$ and that in (3.33) we have $E(P_0) = 0$. Then by the bipolar theorem (Köthe, 1960, page 248)

$$\text{cl } C = C^{00} = \{v\colon \vartheta(v) \leq 1 \text{ for all } \vartheta \in C^0\} \tag{3.34}$$

where

$$C^0 = \{\vartheta\colon \vartheta \in V',\ \vartheta(v) \leq 1 \text{ for all } v \in C\} \tag{3.35}$$

is the polar of C with respect to the duality $\langle V, V' \rangle$. By the Alaoglu-Bourbaki

theorem (Köthe, 1960, page 250), C^0 is compact in the weak topology of V', i.e., in the topology of pointwise convergence of funtionals.

On account of (2.24), (2.25) and (3.34), $\Pi_0(\text{cl } C)$ and $\Pi_0(\text{int } C)$ defined by (2.27) can be represented in the form (3.9):

$$\Pi_0(\text{cl } C) = \Pi(\mathscr{F} \mid \mathscr{K}), \quad \Pi_0(\text{int } C) = \Pi'(\mathscr{F} \mid \mathscr{K}) \tag{3.36}$$

where $\mathscr{K}$ is the sequence $K_1 \subset K_2 \subset \cdots$ appearing in (2.27), and $\mathscr{F}$ is the convex cone of functions $f = a(1 - \vartheta)$ $(a \geq 0, \vartheta \in C^0)$ on V. In particular, by (3.10) and Lemma 3.2

$$D(\Pi_0(\text{cl } C) \parallel Q) = D(\Pi_0(\text{int } C) \parallel Q) = D$$

and the generalized I-projection of Q on $\Pi_0(\text{cl } C)$ and $\Pi_0(\text{int } C)$ is the same P^*. As we already know that $P^* \equiv Q$, Lemma 3.4 gives

$$\log \frac{dP^*}{dQ} = D + \lim_{n\to\infty} a_n(1 - \vartheta_n) \qquad [Q] \tag{3.37}$$

$$(a_n \geq 0, \quad \vartheta_n \in C^0, \quad n = 1, 2, \cdots).$$

Using the compactness of C^0, hence it is easy to conclue that P^* belongs to the exponential family (2.29). In fact, if $a_n \to 0$ then the right side of (3.37) equals the constant D; then necessarily $D = 0$ and $P^* = Q$. Suppose therefore that $a_{n_k} \to a_0 > 0$ for some sequence $\{n_k\}$, where $a_0 = \infty$ is not yet excluded. Then the a.e. convergence of $a_n(1 - \vartheta_n)$ to a finite limit implies the same for ϑ_{n_k}; moreover, $a_0 = \infty$ is possible only if $\vartheta_{n_k} \to 1$ $[Q]$. Let $\vartheta_0 \in C^0$ be a "cluster point" of the sequence ϑ_{n_k} in the topology of pointwise convergence. Then $\lim \vartheta_{n_k}(v) = \vartheta_0(v)$ for each $v \in V$ such that the limit exists, i.e., for Q—a.e. $v \in V$. The possibility of $\vartheta_0(\cdot) = 1$ $[Q]$ and thus that of $a_0 = \infty$ is ruled out by (3.33) with $E(P_0) = 0$, cf. (2.24). Thus (3.37) yields

$$\log \frac{dP^*}{dQ} = D + a_0(1 - \vartheta_0) \qquad [Q] \qquad (a_0 \geq 0, \vartheta_0 \in C^0), \tag{3.38}$$

and P^* belongs to the exponential family (2.29), with $\vartheta = -a_0\vartheta_0$.

Now, writing $\vartheta^* = -a_0\vartheta_0$, we have $\vartheta^*(v) \geq -a_0$ for all $v \in C$ by (3.35), thus (3.38) gives

$$\begin{aligned} D &= -\log \int \exp[a_0(1 - \vartheta_0(\cdot))]\, dQ \\ &= -a_0 - \log \int \exp \vartheta^*(\cdot)\, dQ \leq \inf_{v\in C}\vartheta^*(v) - \log \int \exp \vartheta^*(\cdot)\, dQ. \end{aligned} \tag{3.39}$$

On the other hand, applying (3.3) with the choice $f = f_\vartheta = \vartheta(\cdot) - \inf_{v\in C}\vartheta(v)$, where $\vartheta \in V'$ with $\inf_{v\in C}\vartheta(v) > -\infty$ is arbitrary, we get

$$D(\Pi(C) \parallel Q) \geq \inf_{v\in C}\vartheta(v) - \log \int \exp \vartheta(\cdot)\, dQ, \quad \vartheta \in V', \tag{3.40}$$

with strict inequality unless $[\int \exp f_\vartheta\, dQ]^{-1} \exp f_\vartheta = dP_\vartheta/dQ$, cf. (2.29), equals the

Q-density of the generalized I-projection of Q on $\Pi(C)$. Comparing (3.39) and (3.40) with the obvious inequality $D = D(\Pi_0(\text{int } C) \,\|\, Q) \geq D(\Pi(C) \,\|\, Q)$, it follows that in the latter the equality holds and, in particular, the generalized I-projections of Q on $\Pi(C)$ and $\Pi_0(\text{int } C)$ are the same. Moreover, we have also obtained that the necessary and sufficient condition of the equality in (3.40) is $P_\vartheta = P^*$. This completes the proof of Theorem 3.

4. Proof of the limit theorems. We send forward two simple lemmas.

LEMMA 4.1. *If $\Pi \subset \Lambda$ is a relatively τ_0-open subset of $\Pi \cup \Lambda_f$, i.e., if every $P \in \Pi$ has a τ_0-neighborhood, cf. Definition 2.2, such that $U_0(P, \mathscr{P}, \varepsilon) \cap \Lambda_f \subset \Pi$, then*

$$\liminf_{n\to\infty}(1/n)\log \underline{\Pr}\{\hat{P}_n \in \Pi\} \geq -D(\Pi \,\|\, P_X). \tag{4.1}$$

Further, (4.1) holds for every $\Pi \subset \Lambda$ such that $D(\Pi \,\|\, P_X) = D(\text{int}_{\tau_0}\Pi \,\|\, P_X)$.

PROOF. It suffices to prove the first assertion, for the second one follows by applying it to the τ_0-open set $\text{int}_{\tau_0}\Pi$ instead of Π.

We may suppose that $D(\Pi \,\|\, P_X) < \infty$. Given any $\delta > 0$, pick $P \in \Pi$ with

$$D(P \,\|\, P_X) < D(\Pi \,\|\, P_X) + \delta \tag{4.2}$$

and find $\mathscr{P} = (B_1, \cdots, B_k)$ and $\varepsilon > 0$ such that $U_0(P, \mathscr{P}, \varepsilon) \cap \Lambda_f \subset \Pi$. Choose $0 < \varepsilon' < \varepsilon$ so small that for nonnegative $r_1, \cdots, r_k$ with

$$|r_i - P(B_i)| < \varepsilon', \quad r_i = 0 \quad \text{if} \quad P(B_i) = 0, \quad i = 1, \cdots, k \tag{4.3}$$

we have

$$\left| r_i \log \frac{r_i}{P_X(B_i)} - P(B_i)\log \frac{P(B_i)}{P_X(B_i)} \right| < \frac{\delta}{k}, \quad i = 1, \cdots, k. \tag{4.4}$$

This is possible since $P(B_i) > P_X(B_i) = 0$ is ruled out by (4.2). For sufficiently large n there exist nonnegative integers $\ell_1, \cdots, \ell_k$ such that $\sum_{i=1}^k \ell_i = n$ and $r_i = \ell_i/n$ satisfy (4.3). Then

$$\begin{aligned} \underline{\Pr}\{\hat{P}_n \in \Pi\} &\geq \Pr\{\hat{P}_n \in U_0(P, \mathscr{P}, \varepsilon) \cap \Lambda_f\} \\ &\geq \Pr\{\hat{P}_n(B_i) = \ell_i, i = 1, \cdots, k\} = \frac{\ell_1! \cdots \ell_k!}{n!} \prod_{i=1}^k [P_X(B_i)]^{\ell_i} \\ &\geq (n+1)^{-k}\exp\left\{-n \sum_{i=1}^k r_i \log \frac{r_i}{P_X(B_i)}\right\}. \end{aligned} \tag{4.5}$$

Here the last inequality follows because

$$\frac{\ell_1! \cdots \ell_k!}{n!} \geq (n+1)^{-k}\exp\{-n \textstyle\sum_{i=1}^k r_i \log r_i\};$$

this can be checked by Stirling's approximation or for a simple elementary proof cf. Csiszár and Körner (1981, page 30). Since $r_1, \cdots, r_k$ satisfy (4.3) and therefore

(4.4), we have by (1.4)

$$\sum_{i=1}^{k} r_i \log \frac{r_i}{P_X(B_i)} \leq D_{\mathscr{P}}(P \,\|\, P_X) + \delta \leq D(P \,\|\, P_X) + \delta.$$

Using this and (4.2), we obtain from (4.5)

$$\liminf_{n\to\infty} (1/n) \log \underline{\Pr}\{\hat{P} \in \Pi\} \geq -D(\Pi \,\|\, P_X) - 2\delta.$$

Since $\delta > 0$ was arbitrary, the proof is complete. □

LEMMA 4.2. *If* $\Pi \subset \Lambda$ *is completely convex and* $\Pi' \subset \Pi$, $\overline{\Pr}\{\hat{P}_n \in \Pi'\} > 0$ *then* $P_{X|\hat{P}_n \in \Pi'} \in \Pi$.

PROOF. Write

$$A' = \{\mathbf{s}\colon \hat{P}_n(\mathbf{s}, \cdot) \in \Pi'\}. \tag{4.6}$$

Then, cf. (2.7), $P_{X|\hat{P}_n \in \Pi'} = P_{X_i|A'}$ is the one-dimensional marginal, not depending on i, of $P_{X^n|\hat{P}_n \in \Pi'} = P_{X^n|A'}$; the latter is defined as in (2.6) if $A' \in \mathscr{B}^n$ and by $P_{X^n|A'} = P_{X^n|A}$ else, where $A \in \mathscr{B}^n$, $A \supset A'$ and $P_X^n(A)$ is minimum subject to these conditions, i.e.,

$$P_X^n(A) = \overline{\Pr}\{\hat{P}_n \in \Pi'\}, \quad \text{cf. (2.4)}.$$

Integrating the identity (2.2) with respect to $P_{X^n|A'}$ yields for every $B \in \mathscr{B}$

$$\int \hat{P}_n(\cdot, B)\, dP_{X^n|A'} = \frac{1}{n} \sum_{i=1}^{n} P_{X_i|A'}(B) = P_{X|\hat{P}_n \in \Pi'}(B). \tag{4.7}$$

If $A' \in \mathscr{B}^n$, the integral in (4.7) may be restricted to A', and $P_{X|\hat{P}_n \in \Pi'} \in \Pi$ follows from (4.6) and $\Pi' \subset \Pi$ by the definition of complete convexity. Turning to the case $A' \notin \mathscr{B}^n$, we notice that since the outer $P_{X^n|A'}$-measure of A' is equal to 1, there exists a unique pm P' on the σ-algebra of subsets $F = A' \cap E$ $(E \in \mathscr{B}^n)$ of A' such that $P'(F) = P_{X^n|A'}(E)$. Thus the integral in (4.7) can be written as an integral on A' with respect to P', i.e.,

$$P_{X|\hat{P}_n \in \Pi'}(B) = \int_{A'} \hat{P}_n(\cdot, B)\, dP' \quad (B \in \mathscr{B}). \tag{4.8}$$

Hence $P_{X|\hat{P}_n \in \Pi'} \in \Pi$ follows as before. □

PROOF OF THEOREM 1. Let Π' be an arbitrary subset of the almost completely convex set of pm's $\Pi \subset \Lambda$ such that $\overline{\Pr}\{\hat{P}_n \in \Pi'\} > 0$. By Definition 2.3, there exist completely convex subsets $\Pi_1 \subset \Pi_2 \subset \cdots$ of Π such that, with the notation $\Pi_k' = \Pi_k \cap \Pi'$,

$$\{\mathbf{s}\colon \hat{P}_n(\mathbf{s}, \cdot) \in \Pi'\} = \cup_{k=1}^{\infty} \{\mathbf{s}\colon \hat{P}_n(\mathbf{s}, \cdot) \in \Pi_k'\}. \tag{4.9}$$

Fixing n, consider sets $A \in \mathscr{B}^n$, $A_k \in \mathscr{B}^n$, $k = 1, 2, \cdots$ with $A \supset \{\mathbf{s}\colon \hat{P}_n(\mathbf{s}, \cdot) \in \Pi'\}$, $A_k \supset \{\mathbf{s}\colon \hat{P}_n(\mathbf{s}, \cdot) \in \Pi_k'\}$, whose P_X^n-measure is minimum subject to these

conditions. On account of (4.9), these sets may be chosen such that

$$A_1 \subset A_2 \subset \cdots; \quad A = \cup_{k=1}^{\infty} A_k, \tag{4.10}$$

thus

$$P_X^n(A_k) \to P_X^n(A) = \overline{\Pr}\{\hat{P}_n \in \Pi'\} \quad \text{as} \quad k \to \infty. \tag{4.11}$$

We assume, without any loss of generality, that $P_X^n(A_k) > 0$ for each k, and consider the pm's

$$P_{X^n|\hat{P}_n \in \Pi'} = P_{X^n|A}, \quad P_{X^n|\hat{P}_n \in \Pi_k'} = P_{X^n|A_k} \tag{4.12}$$

defined as in (2.6).

Applying the identity (2.11) with the choice $P^{(n)} = P_{X^n|A_k}$, $Q_1 = \cdots = Q_k = P_X$, it follows that

$$\begin{aligned} -\log P_X^n(A_k) &= D(P_{X^n|A_k} \| P_X^n) \\ &= D(P_{X^n|A_k} \| P_{X|A_k}^n) + nD(P_{X|A_k} \| P_X) \end{aligned} \tag{4.13}$$

where $P_{X|A_k}$, the (unique) one-dimensional marginal of $P_{X^n|A_k}$, satisfies

$$P_{X|A_k} = P_{X|\hat{P}_n \in \Pi_k'} \in \Pi_k \subset \Pi \tag{4.14}$$

by Lemma 4.2. Since (4.14) implies $D(P_{X|A_k} \| P_X) \geq D(\Pi \| P_X)$, (4.11) and (4.13) give

$$(1/n)\log \overline{\Pr}\{\hat{P}_n \in \Pi'\} = \lim_{k\to\infty}(1/n)\log P_X^n(A_k) \leq -D(\Pi \| P_X).$$

This, with the choice $\Pi' = \Pi$, already proves (2.16). If $D(\Pi \| P_X) < \infty$ (4.14) implies also

$$D(P_{X|A_k} \| P_X) \geq D(P_{X|A_k} \| P^*) + D(\Pi \| P_X) \tag{4.15}$$

by (1.5). From (4.13) and (4.15) we obtain, using (2.11) once more (now with $Q_1 = \cdots = Q_n = P^*$), that

$$-\log P_X^n(A_k) \geq D(P_{X^n|A_k} \| P^{*n}) + nD(\Pi \| P_X). \tag{4.16}$$

Here, on account of (4.10) and (2.6),

$$D(P_{X^n|A_k} \| P^{*n}) \to D(P_{X^n|A} \| P^{*n}) \quad \text{as} \quad k \to \infty.$$

Thus, recalling (4.11) and (4.12), the assertion (2.17) follows from (4.16). Under hypothesis (2.19) we have by Lemma 4.1

$$\liminf_{n\to\infty}(1/n)\log \underline{\Pr}\{\hat{P}_n \in \Pi'\} \geq -D(\Pi \| P_X).$$

This and (2.17) complete the proof of Theorem 1. □

LEMMA 4.3. *For any family $\mathcal{F}$ of real-valued measurable functions on $(S, \mathcal{B})$ and any sequence $\mathcal{K} = \{K_i\}_{i=1}^{\infty}$ of subsets of S with the properties* (3.8), *the sets of pm's $\Pi(\mathcal{F} \mid \mathcal{K})$ and $\Pi'(\mathcal{F} \mid \mathcal{K})$ defined by* (3.9) *are almost completely convex.*

PROOF. For $\mu\nu$ as in Definition 3.3

$$\int f\,d\mu\nu = \int \left(\int f\,d\nu(\omega, \cdot) \right) d\mu \tag{4.17}$$

for every measurable and bounded f; hence, if $\nu(\omega, \cdot) \in \Pi(\mathscr{F} \mid K_i)$ for each $\omega \in \Omega$ then also $\mu\nu \in \Pi(\mathscr{F} \mid K_i)$. Since $\Pi(\mathscr{F} \mid \mathscr{K}) = \cup_{i=1}^{\infty} \Pi(\mathscr{F} \mid K_i)$ by definition, this proves that $\Pi(\mathscr{F} \mid \mathscr{K})$ is almost completely convex. The almost complete convexity of $\Pi'(\mathscr{F} \mid \mathscr{K})$ follows in the same way. □

REMARK. If the functions $f \in \mathscr{F}$ are not bounded on S, $\Pi(\mathscr{F})$ and $\Pi'(\mathscr{F})$ defined by (3.7) are not completely convex, in general, since the left side of (4.17) need not exist even if $\int f\,d\nu(\omega, \cdot)$ exists and is positive for every $\omega \in \Omega$. On the other hand, if a sequence $\{K_i\}_{i=1}^{\infty}$ with the properties (3.8) exists such that $\cup_{i=1}^{\infty} K_i = S$ then Lemma 4.3 implies that $\Pi(\mathscr{F})$ and $\Pi'(\mathscr{F})$ are almost completely convex.

PROOF OF THEOREM 4. By assumption, the ψ-image Q of P_X is convex-tight, i.e., there exist compact and convex subsets $K_1 \subset K_2 \subset \cdots$ of V such that

$$Q(\cup_{i=1}^{\infty} K_i) = P_X(\{s\colon \psi(s) \in \cup_{i=1}^{\infty} K_i\}) = 1.$$

Without any loss of generality, we assume that $\psi(s) \in \cup_{i=1}^{\infty} K_i$ for every $s \in S$. Let Π, Π' and Π'' denote the subsets of Λ consisting of those pm's whose ψ-image belongs to $\Pi_0(\text{cl } C)$, $\Pi_0(C)$ and $\Pi_0(\text{int } C)$, respectively, cf. (2.27). Clearly, the event (2.35) is the same as $\hat{P}_n \in \Pi'$.

It follows from Theorem 3 and Lemma 3.3 that both $D(\Pi \parallel P_X)$ and $D(\Pi' \parallel P_X)$ are equal to

$$D(\Pi_0(\text{cl } C) \parallel Q) = D(\Pi_0(\text{int } C) \parallel Q) = D$$

where D is given by (2.30), and the generalized I-projection P^* of P_X on Π is given by (2.36). Hence Theorem 4 follows from Theorem 1 if we show that Π is almost completely convex and Π'' is τ_0-open.

By Lemma 4.3 and (3.36), $\Pi_0(\text{cl } C)$ is almost completely convex. Hence, by the definition of Π and Definition 2.3, Π is almost completely convex, as well. The τ_0-openness of Π'' follows from the weak *-continuity of the mapping $\tilde{P} \to E(\tilde{P})$ in the space of pm's on V satisfying $\tilde{P}(K_n) = 1$, cf. Choquet (1969, page 115). The latter means that to any $\tilde{P}_0$ with $\tilde{P}_0(K_n) = 1$ and 0-neighborhood N in V there exist continuous functions $f_1, \cdots, f_k$ on K_n and positive numbers $\varepsilon_1, \cdots, \varepsilon_k$ such that

$$E(\tilde{P}) \in E(\tilde{P}_0) + N \quad \text{if} \quad \tilde{P}(K_n) = 1$$

$$\text{and} \quad \left| \int f_i\,d\tilde{P} - \int f_i\,d\tilde{P}_0 \right| < \varepsilon_i, \quad i = 1, \cdots, k. \tag{4.18}$$

Now, if $P_0 \in \Pi''$ then its ψ-image $\tilde{P}_0$ satisfies $E(\tilde{P}_0) \in \text{int } C$ and $\tilde{P}_0(K_n) = 1$ for

some n. Then, with N such that $E(\tilde{P}_0) + N \subset \text{int } C$, (4.18) implies the existence of a τ_0-neighborhood $U_0(P_0, \mathscr{P}, \varepsilon) \subset \Pi''$, cf. (2.15), since it is clearly possible to choose $\mathscr{P}$ and ε such that the ψ-image $\tilde{P}$ of each $P \in U_0(P_0, \mathscr{P}, \varepsilon)$ satisfies the hypothesis in (4.18). Thus Π'' is, indeed, τ_0-open, and the proof of Theorem 4 is complete. □

REFERENCES

[1] Bahadur, R. R. and ZABELL, S. L. (1979). Large deviations of the sample mean in general vector spaces. *Ann. Probab.* **7** 587–621.

[2] BAHADUR, R. R., ZABELL, S. L. and GUPTA, J. C. (1980). Large derivations, tests, and estimates, In: *Asymptotic Theory of Statistical Tests and Estimation* 33–64, Proc. Adv. Internat. Symp., Chapel Hill, NC, 1979. Academic, New York.

[3] BÁRTFAI, P. (1974). On a conditional limit theorem. In: *Progress in Statistics* **1** 85–91. Colloquia Math. Soc. J. Bolyai **9**: European Meeting of Statisticians, Budapest, 1972. North Holland, Amsterdam.

[4] BÁRTFAI, P. (1977). On the multivariate Chernoff theorem. Preprint of the Math. Inst. of the Hung. Acad. Sci. Budapest.

[5] BOROVKOV, A. A. (1967). Boundary-value problems of random walks and large derivations in function spaces. *Theor. Probab. Appl.* **12** 575–595.

[6] ČENCOV, N. N. (1972). *Statistical Decision Rules and Optimal Inferences* (in Russian). Nauka, Moscow.

[7] CHOQUET, G. (1969). *Lectures on Analysis*, Vol. 2. Benjamin, New York.

[8] CSISZÁR, I. (1962). Informationstheoretische Konvergenzbergriffe im Raum der Wahrscheinlichkeitsverteilungen. *Publ. Math. Inst. Hung. Acad. Sci.* **7** 137–158.

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

[10] CSISZÁR, I. (1970). Some problems concerning measures on topological spaces and convolutions of measures on topological groups. In: *Les Probabilités sur les Structures Algebriques*, 75–96. Colloques Internationaux du CNRS, **186**, CNRS, Paris.

[11] CSISZÁR, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Ann. Probab.* **3** 146–158.

[12] CSISZÁR, I. and KÖRNER, J. (1981). *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Academic, New York.

[13] DONSKER, M. D. and VARADHAN, S. R. S. (1975–1976). Asymptotic evaluation of certain Markov process expectations for large time I–III. *Comm. Pure Appl. Math.* **28** 1–47, 279–301 and **29** 389–461.

[14] GROENEBOOM, P., OOSTERHOFF, J. and RUYMGAART, F. H. (1979). Large deviation theorems for empirical probability measures. *Ann. Probab.* **7** 553–586.

[15] JUPP, P. E. and MARDIA, K. V. (1983). A note on the maximum entropy principle. *Scand. J. Statist.* **10** 45–47.

[16] KÖTHE, G. (1960). *Topologische Lineare Räume.* Springer, Berlin.

[17] KULLBACK, S. and LEIBLER, R. A. (1951). On information and sufficiency. *Ann. Math. Statist.* **22** 79–86.

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

[19] HOADLEY, A. B. (1967). On the probability of large deviations of functions of several empirical cdf's. *Ann. Math. Statist.* **38** 360–381.

[20] HOEFFDING, W. (1965). Asymptotically optimal tests for multinomial distributions. *Ann. Math. Statist.* **36** 369–401.

[21] LANFORD, O. E. (1973). Entropy and equilibrium states in classical statistical mechanics. In: *Statistical Mechanics and Mathematical Problems.* Lecture Notes in Physics **20** 1–113. Springer, Berlin.

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

[23] SANOV, I. N. (1957). On the probability of large deviations of random variables. *Mat. Sb.* **42** 11–44 (in Russian). English translation in *Sel. Transl. Math. Statist. Probab.* (1961) **1** 213–244.
[24] STONE, M. (1974). Large deviations of empirical probability measures. *Ann. Statist.* **2** 362–366.
[25] TOPSOE, F. (1979). Information theoretical optimization techniques. *Kybernetika* **15** 8–27.
[26] VAN CAMPENHOUT, J. M. and COVER, T. M. (1981). Maximum entropy and conditional probability. *IEEE Trans. Inform. Theory.* **IT-27** 483–489.
[27] VASICEK, O. A. (1980). A conditional law of large numbers. *Ann. Probab.* **8** 142–147.
[28] ZABELL, S. L. (1980). Rates of convergence for conditional expectations. *Ann. Probab.* **8** 928–941.

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