ati comparison.

Statistics & Decisions, Supplement Issue No. 1, 205–237 (1984)
© R. Oldenbourg Verlag, München 1984

INFORMATION GEOMETRY AND ALTERNATING MINIMIZATION PROCEDURES

I. Csiszár and G. Tusnády[*)]

Received: Revised version: July 4, 1983

Summary

Let P and Q be convex sets of finite measures, let $P_o \in P$ be arbitrary, and let for each $n \geq 0$ $Q_n \in Q$ minimize the Kullback-Leibler informational divergence $D(P_n \| Q)$ for $Q \in Q$ while $P_{n+1} \in P$ minimizes $D(P \| Q_n)$ for $P \in P$. We prove that $D(P_n \| Q_n)$ converges to the infimum of $D(P \| Q)$ on $P_o \times Q$ where P_o is the set of all $P \in P$ such that $D(P \| Q_n) < +\infty$ for some n. In special cases also the convergence of the sequences $\{P_n\}$ and $\{Q_n\}$ is proved. The basis

AMS subject classifications. Primary 60 A 10, 65 U 05.
Secondary 65 J 05, 62 B 10, 94 A 34, 40
Key words and phrases: alternating minimization, distance of two sets, Kullback-Leibler informational divergence, EM algorithm, maximum likelihood from incomplete data, decomposition of mixtures, channel capacity, rate distortion function, investment portfolio.

[*)] This paper was presented at the Colloquium on Information Theory, Budapest, August 1981.

of our approach is a general convergence criterion of a geometric flavor, also applicable to other problems.

Implications for iterative algorithms suggested in the literature for computing maximum likelihood estimates (in particular, for decomposition of mixtures), channel capacity, rate-distortion functions and optimum investment portfolios are discussed.

1. Introduction

For the numerical solution of various extremum problems arising in statistics and information theory (as well as in other branches of mathematics) alternating minimization procedures of the following form have been suggested. Let $d(P,Q)$ be an extended real valued function of two variables $P \in \mathcal{P}$, $Q \in \mathcal{Q}$ where $\mathcal{P}$ and $\mathcal{Q}$ are given sets. For P, P' in $\mathcal{P}$ and Q, Q' in $\mathcal{Q}$ write

$$(1.1) \qquad \begin{aligned} P \xrightarrow{1} Q' \quad &\text{iff} \quad d(P,Q') = \min_{Q\in\mathcal{Q}} d(P,Q) < +\infty \\ Q \xrightarrow{2} P' \quad &\text{iff} \quad d(P',Q) = \min_{P\in\mathcal{P}} d(P,Q) < +\infty . \end{aligned}$$

Here the numbers 1 and 2 indicate that the first, respectively second variable of d is fixed in the minimization; the numbers also distinguish this symbol from that of convergence. Now, we say that the sequences $\{P_n\}_{n=0}^{\infty}$ and $\{Q_n\}_{n=0}^{\infty}$ from $\mathcal{P}$ and $\mathcal{Q}$, respectively, are obtained by alternating minimization if $P_n \xrightarrow{1} Q_n \xrightarrow{2} P_{n+1}$, $n = 0,1,\dots$, where the "starting point" $P_0 \in \mathcal{P}$ is arbitrary.

In applications we have in mind, $\mathcal{P}$ and $\mathcal{Q}$ are sets of probability distributions or of arbitrary finite measures and $d(P,Q)$ is the Kullback-Leibler informational divergence $D(P\|Q)$. Often, the above setup arises as a device for solving an extremum problem in one variable. Then the existence of a

Q' (P') minimizing d(P,Q) for fixed P (Q), as well as their explicit form, is known by the very construction. Typical examples are the so-called EM algorithm for maximum likelihood estimation from incomplete data, cf. Dempster, Laird and Rubin (1977) and algorithms suggested by Arimoto (1972) and Blahut (1972) for computing channel capacity and rate-distortion functions, cf. Section 5.

In the theory of numerical methods various general theorems are known on the convergence of iterative procedures, cf. eg. Zangwill (1969). They mainly deal with functions defined on finite dimensional spaces and involve conditions of analytic nature such as compactness, convexity, differentiability. In this paper we give sufficient conditions of a geometric flavor for the convergence of alternating minimization procedures. We shall say that for a $P \in \mathcal{P}$ the "five points property" holds if for every $Q \in \mathcal{Q}$

$$(1.2) \quad d(P,Q) + d(P,Q_o) \geq d(P,Q_1) + d(P_1,Q_1) \quad \text{whenever} \quad Q_o \xrightarrow{2} P_1 \xrightarrow{1} Q_1 \; .$$

Here and in the sequel, the symbols P and Q (with or without indices) stand for elements of $\mathcal{P}$ and $\mathcal{Q}$, respectively, unless stated otherwise.

We prove in Section 2, cf. Theorem 2, that if the five points property holds either for every $P \in \mathcal{P}$ or for some $P \in \mathcal{P}$ which with some $Q \in \mathcal{Q}$ attains the minimum of d then

$$(1.3) \quad \lim_{n\to\infty} d(P_n,Q_n) = \inf_{P\in\mathcal{P},Q\in\mathcal{Q}} d(P,Q) \; ,$$

provided that the last infimum is not changed when replacing $\mathcal{P}$ by

$$(1.4) \quad \mathcal{P}_o = \{P : P \in \mathcal{P},\ d(P,Q_n) < +\infty \text{ for some } n\} \; .$$

Notice that $\mathcal{P}_o$ depends, in general, on the sequences $\{P_n\}$ and $\{Q_n\}$. Of course, if d is finite valued then $\mathcal{P}_o = \mathcal{P}$.

As shown in the Example at the end of Section 2, for closed convex subsets P and Q of a Hilbert space and the squared distance in the role of d, the five points property holds for every $P \in P$; this leads to a new proof of a theorem of Cheney and Goldstein (1959).

In Section 3 the five points property and thereby the convergence of the alternating minimization procedure is proved for the case when P and Q are convex sets of (not necessarily probability) measures and d is the Kullback-Leibler informational divergence D. This is applied in Section 4 to iteratively determine, in certain cases, the element of a convex set of measures which is closest in the divergence sense to a given measure. In particular, if Q is the convex hull of a finite set of measures $Q_1, \ldots, Q_k$ on a measurable space (X, X), and $P \ll \sum_{i=1}^{k} Q_i$ is a probability measure on (X, X), then we get the following result, cf. Theorem 5 (where the notation is somewhat different). Let $(c_1^o, \ldots, c_k^o)$ be any vector with positive components and define recursively

$$c_i^n = c_i^{n-1} \int \frac{dQ_i}{dQ^{n-1}} \quad , \text{ where } \quad Q^{n-1} = \sum_{i=1}^{k} c_i^{n-1} Q_i \ ;$$

then $c_i^n \to c_i^*$ $(i = 1, \ldots, k)$, and $Q^* = \sum_{i=1}^{k} c_i^* Q_i$ satisfies $P \xrightarrow{1} Q^*$.

In Section 5 various applications in statistics and information theory are given.

2. General sufficient conditions for the convergence of alternating minimization procedures

Let P and Q be arbitrary sets and $d(P,Q)$ be an extended real valued function on $P \times Q$ which does not take the value $-\infty$. For any pair of sets $A \subset P$, $B \subset Q$ define

$$d(\mathcal{A},\mathcal{B}) = \inf_{P\in\mathcal{A},Q\in\mathcal{B}} d(P,Q) \ .$$

Let $\{P_n\}_{n=0}^{\infty}$ and $\{Q_n\}_{n=0}^{\infty}$ be sequences from $\mathcal{P}$ and $\mathcal{Q}$, respectively, not necessarily obtained by alternating minimization. We clearly have

$$(2.1) \qquad d(P_n,Q_n) \geqq d(\mathcal{P}_o,\mathcal{Q}) \quad n = 0,1,\dots$$

where $\mathcal{P}_o$ has been defined in (1.4). We shall give criteria ensuring

$$\lim_{n\to\infty} d(P_n,Q_n) = d(\mathcal{P}_o,\mathcal{Q}) \ .$$

The following simple lemma will be used.

Lemma 1. Let a_n, b_n $(n = 0,1,\dots)$ be extended real numbers greater than $-\infty$ and c a finite number such that

$$(2.2) \qquad c + b_{n-1} \geqq b_n + a_n \, , \quad n = 1,2,\dots$$

and

$$(2.3) \qquad \limsup_{n\to\infty} b_n > -\infty \, , \quad b_{n_o} < +\infty \quad \text{for some } n_o \ .$$

Then

$$\liminf_{n\to\infty} a_n \leqq c \ .$$

If, in addition,

$$(2.4) \qquad \sum_{n=0}^{\infty} (c-a_n)^+ < +\infty$$

then

$$\sum_{n=n_o+1}^{\infty} |a_n - c| < +\infty$$

and consequently

$$\lim_{n\to\infty} a_n = c \ .$$

Proof. If $\sum_{n=0}^{\infty} (c - a_n)^+ = +\infty$ then $\liminf_{n\to\infty} a_n \leqq c$ obviously holds, hence it suffices to prove the second assertion. (2.2) implies, by induction, that $a_n < +\infty$, $b_n < +\infty$ for every $n > n_o$ if $b_{n_o} < +\infty$. Thus (2.2) gives

$$a_n - c \leq b_{n-1} - b_n \quad \text{if} \quad n > n_o .$$

Hence for every $N > n_o$

$$\sum_{n=n_o+1}^{N} (a_n - c) \leq b_{n_o} - b_N .$$

Here, by (2.4), the left side has a limit if $N \to \infty$, and the proof will be complete if we show that this limit is less than $+\infty$. This follows, however, from assumption (2.3) which implies that this limit has the finite upper bound $b_{n_o} - \limsup_{n\to\infty} b_n$. □

<u>Theorem 1</u>. For arbitrary sequences $\{P_n\}_{n=0}^{\infty}$ and $\{Q_n\}_{n=0}^{\infty}$ from $\mathcal{P}$ resp. $\mathcal{Q}$ such that

$$(2.5) \qquad d(P,Q) + d(P,Q_{n-1}) \geq d(P,Q_n) + d(P_n,Q_n) , \quad n = 1,2,\ldots,$$

either for every $P \in \mathcal{P}_o$ (cf. (1.4)) or for some $P \in \mathcal{P}_o$ such that $d(P,Q) = d(\mathcal{P}_o,\mathcal{Q})$ then

$$(2.6) \qquad \lim_{n\to\infty} d(P_n,Q_n) = d(\mathcal{P}_o,\mathcal{Q}) .$$

Under the first hypothesis the sequence $\{d(P_n,Q_n)\}_{n=0}^{\infty}$ is non-increasing, while under the second hypothesis

$$(2.7) \qquad \sum_{n=n_1}^{\infty} (d(P_n,Q_n) - d(\mathcal{P}_o,\mathcal{Q})) < +\infty$$

for some index n_1.

<u>Proof</u>. If $\mathcal{P}_o = \emptyset$ then $d(P_n,Q_n) = d(\mathcal{P}_o,\mathcal{Q}) = +\infty$ for every n thus (2.6) is true. If (2.5) holds for some $P \in \mathcal{P}_o$ then Lemma 1 applies to

$$a_n = d(P_n,Q_n) , \; b_n = d(P,Q_n) , \; c = d(P,Q) .$$

In fact, since $P \in \mathcal{P}_o$, we have $b_{n_o} < +\infty$ for some n_o and

$c < +\infty$, cf. (1.4). Then (2.5) with $n = n_0 + 1$ shows that $c = d(P,\mathcal{Q}) > -\infty$, and since $b_n \geq c$, this implies, in turn, that the condition $\limsup\limits_{n\to\infty} b_n > -\infty$ is also met.

Under our first hypothesis, Lemma 1 gives that

(2.8) $\quad \liminf\limits_{n\to\infty} d(P_n,Q_n) \leq d(P,\mathcal{Q})$

for every $P \in \mathcal{P}_0$. If we establish the monotonicity of $d(P_n,Q_n)$, this and (2.1) will imply (2.6). Supposing $d(P_{n-1},Q_{n-1}) < +\infty$, the substitution $P = P_{n-1}$ in (2.5) shows that $d(P_{n-1},Q_n) < +\infty$ and

$$d(P_n,Q_n) \leq d(P_{n-1},Q_{n-1})$$

as claimed. Since the last inequality is trivial if $d(P_{n-1},Q_{n-1}) = +\infty$, the assertion of Theorem 1 under the first hypothesis is proved.

Under the second hypothesis, (2.1) implies

$$d(P_n,Q_n) \geq d(\mathcal{P}_0,\mathcal{Q}) = d(P,\mathcal{Q}) \ ,$$

i.e., $a_n \geq c$ for every n. Hence the second part of Lemma 1 immediately gives (2.6) and (2.7). □

In the sequel we shall consider sequences $\{P_n\}_{n=0}^{\infty}$ and $\{Q_n\}_{n=0}^{\infty}$ obtained by alternating minimization, i.e., with the notation (1.1),

(2.9) $\quad P_0 \xrightarrow{1} Q_0 \xrightarrow{2} P_1 \xrightarrow{1} Q_1 \xrightarrow{2} \dots ,$

where the "starting point" $P_0 \in \mathcal{P}$ is arbitrary. Of course, such a pair of sequences need not exist for an arbitrary $P_0 \in \mathcal{P}$, or, conceivably, for no $P_0 \in \mathcal{P}$. The question of existence will not be entered here.

Clearly, for sequences satisfying (2.9)

(2.10) $\quad d(P_n,Q_n) \geq d(P_{n+1},Q_n) \geq d(P_{n+1},Q_{n+1}) \ , \quad n = 0,1,\dots .$

If the five points property (1.2) holds for some $P \in \mathcal{P}$ then condition (2.5) of Theorem 1 is met for the same P, for every pair of sequences constructed by alternating minimization. Now we formulate two conditions which together imply the five points property. These will be the conditions we shall check in the applications of Theorem 1.

Let $\delta(P,P')$ be a non-negative valued function on $\mathcal{P} \times \mathcal{P}$ such that $\delta(P,P) = 0$ for each $P \in \mathcal{P}$. Given d and δ, we shall say that for a $P \in \mathcal{P}$ the "three points property" holds, if

$$(2.11) \quad \delta(P,P_1) + d(P_1,Q_o) \leq d(P,Q_o) \quad \text{whenever} \quad Q_o \xrightarrow{2} P_1 \ .$$

Further, we say that for a $P \in \mathcal{P}$ the "four points property" holds if for every $Q \in \mathcal{Q}$

$$(2.12) \quad d(P,Q_1) \leq \delta(P,P_1) + d(P,Q) \quad \text{whenever} \quad P_1 \xrightarrow{1} Q_1 \ .$$

An example suggesting an intuitive geometric interpretation of the three and four points properties will be given at the end of this section.

Theorem 2. Let $\{P_n\}_{n=0}^{\infty}$ and $\{Q_n\}_{n=0}^{\infty}$ be sequences obtained by alternating minimization, cf. (2.9). Then

$$(2.13) \quad \lim_{n\to\infty} d(P_n,Q_n) = d(\mathcal{P}_o,\mathcal{Q})$$

providing either every $P \in \mathcal{P}_o$, or some $P \in \mathcal{P}_o$ with $d(P,\mathcal{Q}) = d(\mathcal{P}_o,\mathcal{Q})$, has the five points property (1.2), where $\mathcal{P}_o$ is defined by (1.4). The five points property (1.2) is implied by the three and four points properties (2.11), (2.12). Further, if the latter properties hold for some $P \in \mathcal{P}_o$ with $d(P,\mathcal{Q}) = d(\mathcal{P}_o,\mathcal{Q})$ then, in addition to (2.13), we also have (for this P)

$$(2.14) \quad \delta(P,P_{n+1}) \leq \delta(P,P_n) \ , \quad n = 0,1,\dots \ .$$

Proof. As it has already been noted, the first assertion is immediate from Theorem 1. To prove that (2.11) and (2.12) imply the five points property (1.2), it suffices to consider

or ie $P \in \mathcal{P}$ then
ne same P, for every
ing minimization. Now
r imply the five
ions we shall check

function on $\mathcal{P} \times \mathcal{P}$
iven d and δ, we
ints property" holds,if
never $Q_o \xrightarrow{2} P_1$.

our points property"

ver $P_1 \xrightarrow{1} Q_1$.

metric interpretation
ill be given at the

be sequences ob-
2. Then

$P \in \mathcal{P}_o$ with
perty (1.2), where
property (1.2) is
erties (2.11), (2.12).
r some $P \in \mathcal{P}_o$ with
.13), we also have

. .

he first assertion
t (2.11) and (2.12)
suffices to consider

Q_o 's with $d(P,Q_o) < +\infty$. Then (2.11) implies $\delta(P,P_1) < +\infty$ and (1.2) follows by adding (2.11) and (2.12), since $d(P_1,Q_1) \le d(P_1,Q_o)$. Finally, if P has the three and four points properties, substitute $Q_n \xrightarrow{2} P_{n+1}$ for $Q_o \xrightarrow{2} P_1$ in (2.11) and $P_n \xrightarrow{1} Q_n$ for $P_1 \xrightarrow{1} Q_1$ in (2.12). Thus we get for every $Q \in \mathcal{Q}$

$$\delta(P,P_{n+1}) + d(P_{n+1},Q_n) \le d(P,Q_n) \le \delta(P,P_n) + d(P,Q)$$

thus

$$\delta(P,P_{n+1}) + d(P_{n+1},Q_n) \le \delta(P,P_n) + d(P,\mathcal{Q}).$$

If here $d(P,\mathcal{Q}) = d(\mathcal{P}_o,\mathcal{Q})$ then $d(P_{n+1},Q_n) \ge d(P,\mathcal{Q})$, thus (2.14) follows. □

Remarks. Of course, in applications of Theorem 2 one wants to have the limit relation (1.3) rather than (2.13). To this end, the starting point $P_o \in \mathcal{P}$ (or $Q_o \in \mathcal{Q}$) of the iteration should be "properly" selected, i.e., in such a way that for the set $\mathcal{P}_o$ defined by (1.4) $d(\mathcal{P}_o,\mathcal{Q}) = d(\mathcal{P},\mathcal{Q})$ should hold. One might be interested also in the convergence of the very sequences $\{P_n\}$ and $\{Q_n\}$ to limits P^* and Q^* such that $d(P^*,Q^*) = d(\mathcal{P},\mathcal{Q})$ (in some "natural" topology on $\mathcal{P}$ and $\mathcal{Q}$). In applications we shall consider, this is often easy to show if $d(\mathcal{P},\mathcal{Q})$ is attained by a unique pair (P^*,Q^*). The last assertion of Theorem 2 will be useful to prove convergence results $P_n \to P^*$, $Q_n \to Q^*$ in the harder case when $\min d(P,Q)$ may be attained for several pairs (P,Q), cf. the next example and Theorem 3.

Example. Let $\mathcal{P}$ and $\mathcal{Q}$ be closed convex subsets of a Hilbert space and set $d(P,Q) = \|P - Q\|^2$, $\delta(P,P') = \|P - P'\|^2$. Then the three and four points properties (2.11) and (2.12) hold for every $P \in \mathcal{P}$. This is an elementary geometric consequence of the fact that the triangles Q_oP_1P and P_1Q_1Q have angles $\ge \pi/2$ at P_1 and Q_1, respectively. Since d is finite valued, we have $\mathcal{P}_o = \mathcal{P}$ and Theorem 2 gives

$$\lim_{n\to\infty} d(P_n,Q_n) = d(\mathcal{P},\mathcal{Q}) \tag{2.15}$$

for any pair of sequences $\{P_n\}$ and $\{Q_n\}$ obtained by alternating minimization. A theorem of Cheney and Goldstein (1959) states that if $\mathcal{P}$ (say) is compact then $P_n \to P^*$, $Q_n \to Q^*$ where $d(P^*,Q^*) = d(\mathcal{P},\mathcal{Q})$. A new proof of this can be obtained from our results as follows. Let P^* be the limit point of some subsequence $\{P_{n_i}\}$ of $\{P_n\}$. Then (2.15) implies $d(P^*, \mathcal{Q}) = d(\mathcal{P},\mathcal{Q})$, and (2.14) gives that the sequence $\|P^* - P_n\|^2$, $n = 1,2,\dots$ is non-increasing. Since $P_{n_i} \to P^*$, this sequence then must converge to zero, which means that actually $P_n \to P^*$. Of course, this implies that Q_n also converges to some Q^*, and $d(P^*,Q^*) = d(\mathcal{P},\mathcal{Q})$ follows from (2.15).

3. Information distance of convex sets of measures

Let $(X,\mathcal{X})$ be an arbitrary measurable space. Throughout this section, $\mathcal{P}$, $\mathcal{Q}$ will be sets of finite measures on $(X,\mathcal{X})$. (By measure we shall always mean a finite measure not identically 0, unless stated otherwise.) We suppose that both $\mathcal{P}$ and $\mathcal{Q}$ are convex, i.e., for arbitrary P_o, P_1 in $\mathcal{P}$ resp. Q_o, Q_1 in $\mathcal{Q}$ and $0 < t < 1$ the measures

$$(3.1) \qquad P_t = (1-t)P_o + tP_1 \ , \quad Q_t = (1-t)Q_o + tQ_1$$

also belong to $\mathcal{P}$ resp. $\mathcal{Q}$.

Let the role of $d(P,Q)$ be played by the Kullback-Leibler informational divergence

$$(3.2) \qquad D(P\|Q) = \begin{cases} \int \log p\,dP & \text{if} \quad P \ll Q \\ +\infty & \text{if} \quad P \not\ll Q \end{cases}$$

where $p = \frac{dP}{dQ}$. (Concerning a different definition cf. the remarks at the end of this section.)

We shall show that Theorem 2 on alternating minimization procedures applies to this case, by proving that the three and four points properties (2.11) and (2.12) hold for every $P \in \mathcal{P}$, if $\delta(P,P')$ is defined as

(3.3) $\delta(P,P') = D(P\|P') + P'(X) - P(X)$.

While $D(P\|Q) \geq 0$ in the most familiar case when P and Q are probability measures (with equality iff $P = Q$), for arbitrary measures $D(P\|Q) < 0$ is also possible. The functional δ defined by (3.3) is, however, always non-negative and vanishes iff $P = P'$, as one sees from the inequality $-\log t \geq 1 - t$.

Let us recall the notation (1.1) which we henceforth use for $D(P\|Q)$ in the role of $d(P,Q)$. With an obvious extension, notation like $P \xrightarrow{1} Q'$ or $Q \xrightarrow{2} P'$ will also be used when P or Q is an arbitrary measure not necessarily in $\mathcal{P}$ or $\mathcal{Q}$.

The following lemma establishes the three points property. The case of probability measures has been covered already in Csiszár (1975) Theorem 2.2. The general case is completely analogous; still, for the reader's convenience, we give the simple proof.

Lemma 2. Let $\mathcal{P}$ be a convex set of measures and let Q_o be another measure on $(X,\mathcal{X})$. Then $Q_o \xrightarrow{2} P_1$ implies

(3.4) $D(P\|P_1) + P_1(X) - P(X) + D(P_1\|Q_o) \leq D(P\|Q_o)$

for every $P \in \mathcal{P}$.

Proof. By assumption, $D(P_1\|Q_o) = D(\mathcal{P}\|Q_o) < +\infty$. $D(P\|Q_o) < +\infty$ may also be assumed since else (3.4) is trivial. Write

(3.5) $p_1 = \dfrac{dP_1}{dQ_o}$, $p = \dfrac{dP}{dQ_o}$.

Since the measures $P_t = (1 - t)P + tP_1$ belong to the convex set $\mathcal{P}$ for each $0 < t \leq 1$, it follows that

$$f(t) = D(P_t\|Q_o)$$

attains its minimum at $t = 1$. Thus

(3.6) $0 \geq \dfrac{f(1) - f(t)}{1 - t} = \int \frac{1}{1-t}[p_1 \log p_1 - p_t \log p_t]dQ_o$

where $p_t = (1-t)p + tp_1$. Here the integrand is a difference quotient of the convex function $p_t \log p_t$ of t, hence it is non-increasing as $t \uparrow 1$. By monotone convergence, the limit as $t \uparrow 1$ in (3.6) can be exchanged with the integration, yielding

$$0 \geq \int \frac{d}{dt}(p_t \log p_t)\,|_{t=1}\, dQ_o = \int (1 + \log p_1)(p_1 - p)\,dQ_o \ .$$

This is equivalent to (3.4). □

The next lemma establishes the four points property.

Lemma 3. Let $\mathcal{Q}$ be a convex set of measures and let P_1 be another measure on $(X, \mathcal{X})$. Then $P_1 \xrightarrow{1} Q_1$ implies

(3.7) $\quad D(P\|Q_1) \leq D(P\|P_1) + P_1(X) - P(X) + D(P\|Q)$

for every measure P on $(X, \mathcal{X})$ and every $Q \in \mathcal{Q}$.

Proof. The proof is similar to that of Lemma 2. The assumptions imply that the function

$$g(t) = D(P_1\|Q_t) \ , \quad 0 < t \leq 1$$

attains its minimum at $t = 1$, where $Q_t = (1-t)Q + tQ_1$.

Denote by $\bar{Q}$ and $\overline{Q_1}$ the absolutely continuous component with respect to P_1 of Q and Q_1, respectively, and write

$$q = \frac{d\bar{Q}}{dP_1} \ , \quad q_1 = \frac{d\bar{Q}_1}{dP_1} \ .$$

Since $D(P_1\|Q_1) = D(P_1\|\mathcal{Q}) < +\infty$, we have $P_1 \ll Q_1$ and consequently, $q_1 > 0$ P_1 - a.e. Further, $P_1 \ll Q_1$ implies $P_1 \ll Q_t$ for every $0 < t \leq 1$. Since the Radon-Nikodym derivative of the absolutely continuous component of Q_t with respect to P_1 is $q_t = (1-t)q + tq_1$, it follows by (3.2) that

$$0 \geq \frac{g(1) - g(t)}{1 - t} = \int \frac{1}{1-t}(-\log q_1 + \log q_t)\,dP_1 \ .$$

One sees as in the proof of Lemma 2 that the integrand converges non-increasingly to $\frac{d}{dt}(-\log q_t)\,|_{t=1} = 1 - \frac{q}{q_1}$. Thus by monotone convergence we get

$$(3.8)\qquad 0 \le \int (1 - \frac{q}{q_1})\, dP_1 \ .$$

Now, to prove (3.7) we may suppose that $P \ll P_1$, $P \ll Q$, for else the right side is $+\infty$. Then $P_1 \ll Q_1$ implies also $P \ll Q_1$, further, with $p = \frac{dP}{dP_1}$, we have P - a.e. $pqq_1 > 0$ and

$$\frac{dP}{dQ_1} = \frac{p}{q_1}\ , \qquad \frac{dP}{dQ} = \frac{p}{q}\ .$$

Thus, using the inequality $-\log t \ge 1 - t$, we obtain

$$D(P\|P_1) + D(P\|Q) - D(P\|Q_1) = \int [\log p + \log \frac{p}{q} - \log \frac{p}{q_1}]dP =$$

$$= \int -\log \frac{q}{pq_1}\, dP \ge \int (1 - \frac{q}{pq_1})\, dP = P(X) - \int \frac{q}{q_1}\, dP_1\ .$$

On account of (3.8), this completes the proof. □

Combining our results thus far, we obtain

Theorem 3. Let $\mathcal{P}$ and $\mathcal{Q}$ be convex sets of measures on $(X,\mathcal{X})$ and let $\{P_n\}_{n=0}^{\infty}$, $\{Q_n\}_{n=0}^{\infty}$ be sequences from $\mathcal{P}$ resp. $\mathcal{Q}$ obtained by alternating minimization of $d(P,Q) = D(P\|Q)$, starting from some $P_0 \in \mathcal{P}$, cf. (2.9). Then

$$\lim_{n\to\infty} D(P_n\|Q_n) = D(\mathcal{P}_0\|\mathcal{Q})\ ,$$

where

$$(3.9)\qquad \mathcal{P}_0 = \{P : D(P\|Q_n) < +\infty \ \text{for some}\ n\}\ .$$

Further, if X is a finite set and $\mathcal{P}$ and $\mathcal{Q}$ are closed in the topology of pointwise convergence then P_n converges to some $P^* \in \mathcal{P}_0$ such that $D(P^*\|\mathcal{Q}) = D(\mathcal{P}_0\|\mathcal{Q})$.

Remark. If X is finite and P_0 is positive for exactly those $x \in X$ to which there exist $P \in \mathcal{P}$ and $Q \in \mathcal{Q}$ with $P(x)Q(x) > 0$, then $\mathcal{P}_0 = \{P : D(P\|\mathcal{Q}) < +\infty\}$, and consequently

$D(\mathcal{P}_o \| \mathcal{Q}) = D(\mathcal{P} \| \mathcal{Q})$.

Proof. Lemmas 2 and 3 mean that the three and four points properties hold for every $P \in \mathcal{P}$. This, by Theorem 2, proves the first assertion.

If X is a finite set then $D(P\|Q) < +\infty$ iff $P \ll Q$. Hence one sees that if $\mathcal{P}$ is closed, so is $\mathcal{P}_o$. Thus there is a convergent subsequence $\{P_{n_i}\}$ of $\{P_n\}$ with limit $P^* \in \mathcal{P}_o$, say, and for some further subsequence $\{n_i'\}$ of $\{n_i\}$ also $\{Q_{n_i'}\}$ converges to some $Q^* \in \mathcal{Q}$. Since $D(P\|Q)$ is lower semicontinuous, we then have

$$(3.10) \quad D(P^* \| Q^*) \leq \lim D(P_{n_i'} \| Q_{n_i'}) = D(\mathcal{P}_o \| \mathcal{Q}) \ ,$$

where, of course, the strict inequality is impossible. Thus $D(P^* \| \mathcal{Q}) = D(\mathcal{P}_o \| \mathcal{Q})$ and by (2.14) in Theorem 2 it follows that $\delta(P^*, P_n)$ is monotone non-increasing in n, where δ is defined in (3.3). In particular, $\lim\limits_{n\to\infty} \delta(P^*, P_n)$ exists, and then $P_{n_i} \to P^*$ implies that this limit equals 0. Hence, in turn, we can conclude that $P_n \to P^*$. □

The convergence $P_n \to P^*$ (proved for finite X) immediately imply the convergence of $Q_n(x)$ for every $x \in X$ with $P^*(x) > 0$. Notice the essential role of the continuity argument in the proof involving the function δ defined by (3.3). This causes the proof break down for infinite X. On the other hand, if it were known that a unique pair (P^*, Q^*) attaining the minimum $D(\mathcal{P} \| \mathcal{Q})$ exists then $P_n \to P^*$, $Q_n \to Q^*$ could be proved under much weaker conditions (assuming now that $D(\mathcal{P}_o \| \mathcal{Q}) = D(\mathcal{P} \| \mathcal{Q})$). E.g., if on a metric space with its Borel σ-algebra the usual weak convergence of measures were considered, compactness of both $\mathcal{P}$ and $\mathcal{Q}$ would already be sufficient for $P_n \to P^*$, $Q_n \to Q^*$. This follows by the argumen leading to (3.10), since $D(P\|Q)$ is known to be lower semicontinuous for the weak convergence of measures.

For practical computations based on Theorem 3, it is desirable to have some bound on the difference $D(P_n \| Q_n) - D(\mathcal{P}_o \| \mathcal{Q})$, for the purpose of determining when to stop the iteration. If a $P \in \mathcal{P}_o$ attains $D(\mathcal{P}_o \| \mathcal{Q})$, the five points property (1.2) (with $d = D$) for Q_{n-1}, P_n, Q_n in the role of Q_o, P_1, Q_1 implies

$$D(\mathcal{P}_o \| \mathcal{Q}) + D(P \| Q_{n-1}) \geq D(P \| Q_n) + D(P_n \| Q_n) ,$$

i.e.,

$$D(P_n \| Q_n) - D(\mathcal{P}_o \| \mathcal{Q}) \leq D(P \| Q_{n-1}) - D(P \| Q_n) = \int \log \frac{dQ_n}{dQ_{n-1}} dP .$$

While this bound involves the unknown P, it leads to

$$(3.11) \quad D(P_n \| Q_n) - D(\mathcal{P}_o \| \mathcal{Q}) \leq \log \sup_x \frac{dQ_n}{dQ_{n-1}}$$

which is already a useful bound in many cases. In particular, if X is a finite set and the sequence Q_n is convergent then the right side of (3.11) tends to 0 as $n \to \infty$.

Remarks. In some applications, one may be interested in minimizing instead of $D(P \| Q)$ some related functional. E.g., the integral with respect to P of some given function $c(x)$ may be added to $D(P \| Q)$. Notice that this particular case is covered by the results in this section, since

$$D(P \| Q) + \int c(x) dP = D(P \| \tilde{Q})$$

where $\tilde{Q}$ is defined by letting $\frac{d\tilde{Q}}{dQ} = \exp(-c(x))$. Another choice of interest for at least one application (the computation of channel capacity per unit cost, see Section 5) is

$$(3.12) \quad d(P,Q) = \frac{D(P \| Q)}{\int c(x) dP} ,$$

where $c(x)$ is a given positive valued function. One can show similarly to the proof of Lemma 2 that the three points property holds also in this case, with

$$\delta(P,P') = \frac{D(P \| P') + P'(X) - P(X)}{\int c(x) dP} ,$$

while the four points property for these d and δ is an immediate consequence of Lemma 3. Choosing, in particular, $c(x) = 1$, (3.12) gives the alternative definition of informational divergence suggested in Rényi (1961).

4. Minimizing information distance from a single measure

Let $(X,\mathcal{X})$ and $(Y,\mathcal{Y})$ be measurable spaces and T be a measurable mapping of $(X,\mathcal{X})$ into $(Y,\mathcal{Y})$. The T-image of a measure Q on $(X,\mathcal{X})$ will be denoted by Q^T and for a set $\mathcal{Q}$ of measures on $(X,\mathcal{X})$ we shall write

$$(4.1) \qquad \mathcal{Q}^T = \{Q^T : Q \in \mathcal{Q}\} .$$

It may happen that to measures P on $(X,\mathcal{X})$ one can "easily" find $Q \in \mathcal{Q}$ with $P \xrightarrow{1} Q$ (cf. (1.1), where now $d(P,Q) = D(P\|Q)$), while to measures $\tilde{P}$ on $(Y,\mathcal{Y})$ one cannot easily find $Q^T \in \mathcal{Q}^T$ with $\tilde{P} \xrightarrow{1} Q^T$. Similarly, given a set $\mathcal{P}$ of measures on $(X,\mathcal{X})$ it may be "easy" to find $P \in \mathcal{P}$ with $Q \xrightarrow{2} P$ (to given Q), without having a direct way of finding $P^T \in \mathcal{P}^T$ with $\tilde{Q} \xrightarrow{2} P^T$ to a given $\tilde{Q}$ on $(Y,\mathcal{Y})$. The next theorem shows how alternating minimization can be used to these problems. We shall use the well-known inequality

$$D(P\|Q) \geq D(P^T\|Q^T)$$

valid for any two measures P and Q on $(X,\mathcal{X})$, where in case $P \ll Q$ with

$$(4.2) \qquad \frac{dP}{dQ}(x) = \frac{dP^T}{dQ^T}(T(x)) \quad \text{for every } x \in X$$

the equality holds.

<u>Theorem 4</u>. (i) Given a measure $\tilde{P}$ on $(Y,\mathcal{Y})$ and a convex set $\mathcal{Q}$ of measures on $(X,\mathcal{X})$, define

$$(4.3) \qquad \mathcal{P} = \{P : P^T = \tilde{P}\} .$$

Starting from some $Q_o \in \mathcal{Q}$, let $Q_o \xrightarrow{2} P_1 \xrightarrow{1} Q_1 \xrightarrow{2} P_2 \xrightarrow{1} \ldots$ be obtained by alternating minimization of $D(P\|Q)$, $P \in \mathcal{P}$,

$Q \in \mathcal{Q}$, where the step $Q_{n-1} \xrightarrow{2} P_n$ is given, cf. (4.2), by

$$\frac{dP_n}{dQ_{n-1}}(x) = \frac{d\tilde{P}}{dQ^T_{n-1}}(T(x)) \quad \text{for every} \quad x \in X . \tag{4.4}$$

Then

$$\lim_{n\to\infty} D(\tilde{P}\| Q^T_n) = D(\tilde{P}\| \mathcal{Q}^T) \tag{4.5}$$

iff for $\mathcal{P}_o$ defined by (3.9) we have

$$D(\mathcal{P}_o\|\mathcal{Q}) = D(\mathcal{P}\|\mathcal{Q}) . \tag{4.6}$$

(ii) Given a convex set $\mathcal{P}$ of measures on $(X,\mathcal{X})$ and a measure $\tilde{Q}$ on $(Y,\mathcal{Y})$, set

$$\mathcal{Q} = \{Q : Q^T = \tilde{Q}\} . \tag{4.7}$$

Starting from some $P_o \in \mathcal{P}$, let $P_o \xrightarrow{1} Q_o \xrightarrow{2} P_1 \xrightarrow{1} Q_1 \rightarrow \cdots$ be obtained by alternating minimization of $D(P\|Q)$, $P \in \mathcal{P}$, $Q \in \mathcal{Q}$, where the step $P_n \xrightarrow{1} Q_n$ is determined, cf. (4.2), by

$$\frac{dP_n}{dQ_n}(x) = \frac{dP^T_n}{d\tilde{Q}}(T(x)) \quad \text{for every} \quad x \in X . \tag{4.8}$$

Then (4.6) is necessary and sufficient for

$$\lim_{n\to\infty} D(P^T_n\|\tilde{Q}) = D(\mathcal{P}^T\|\tilde{Q}) . \tag{4.9}$$

Proof. (4.4) and (4.8) guarantee, respectively,

$$D(P_n\|Q_{n-1}) = D(\tilde{P}\|Q^T_{n-1})$$

in case (i) and

$$D(P_n\|Q_n) = D(P^T_n\|\tilde{Q}) ,$$

in case (ii). Since

$$\lim_{n\to\infty} D(P_n\|Q_n) = D(\mathcal{P}_o\|\mathcal{Q})$$

by Theorem 3, the proof is complete. □

The intuitive assumption underlying Theorem 4, i.e., that an element of $\mathcal{Q}$ $(\mathcal{P})$ minimizing $D(P\|Q)$ for given P (Q) can be "easily" found, is fulfilled in the important case described in the next lemma.

Lemma 4. Suppose that $(X,\mathcal{X}) = (Z,\mathcal{Z}) \times (Y,\mathcal{Y})$, where $Z = \{1,\dots,k\}$ and $\mathcal{Z}$ is the family of all subsets of Z. Let $\mu_1,\dots,\mu_k$ be given measures on $(Y,\mathcal{Y})$, and let $\mathcal{R}$ be the set of all measures on $(X,\mathcal{X})$ of form

$$R = \sum_{i=1}^{k} c_i \delta_i x \mu_i \ , \quad c_i \geq 0 \ , \quad \sum_{i=1}^{k} c_i = 1 \ ,$$

where δ_i is the point mass in $i \in Z$. Then to any given measure S on $(X,\mathcal{X})$ with $D(S\|\mathcal{R}) < +\infty$ or $D(\mathcal{R}\|S) < +\infty$, respectively, the measure $R \in \mathcal{R}$ with $S \xrightarrow{1} R$ or $S \xrightarrow{2} R$, respectively, is uniquely determined by

$$(4.10) \quad c_i = S(i)/S(X) \ ,$$

and

$$(4.11) \quad c_i = \frac{1}{e} S(i) \exp\{- \frac{D(\mu_i\|\nu_i) + \lambda}{\mu_i(Y)}\} \ ,$$

respectively, where $S(i) = S(\{i\} \times Y)$, ν_i is defined by

$$\nu_i(B) = \frac{S(\{i\} \times B)}{S(i)} \ ,$$

and λ is determined by the condition $\Sigma c_i = 1$.

Proof. For $R = \sum_{i=1}^{k} c_i \delta_i x \mu_i$ we have

$$D(S\|R) = \int \log\frac{dS}{dR} dS = \sum_{i=1}^{k} S(i) \log\frac{S(i)}{c_i} + \sum_{i=1}^{k} S(i) \int \log\frac{d\nu_i}{d\mu_i} d\nu_i$$

$$D(R\|S) = \int \log\frac{dR}{dS} dR = \sum_{i=1}^{k} c_i \mu_i(Y) \log\frac{c_i}{S(i)} + \sum_{i=1}^{k} c_i \int \log\frac{d\mu_i}{d\nu_i} d\mu_i \ .$$

Hence the assertions follow by standard calculus. □

Using Theorem 4 and Lemma 4, one can find to any measures $\widetilde{S}, \mu_1,\dots,\mu_k$ on $(Y,\mathcal{Y})$ a convex combination $\widetilde{R} = \sum_{i=1}^{k} c_i\mu_i$ for which $D(\widetilde{S}\|\widetilde{R})$ or $D(\widetilde{R}\|\widetilde{S})$ is minimized, by an explicitly defined iterative procedure. For brevity, we shall consider only the first case, assuming - without any loss of generality - that $\widetilde{S} = \widetilde{P}$ is a probability measure.

Theorem 5. Let $\widetilde{P}, \mu_1,\dots,\mu_k$ be given measures on a measurable space $(Y,\mathcal{Y})$ and let $\widetilde{\mathcal{Q}}$ be the set of all measures of form

$$(4.12) \quad \widetilde{Q} = \sum_{i=1}^{k} c_i\mu_i \ , \quad c_i \geq 0 \ , \quad \sum_{i=1}^{k} c_i = 1 \ .$$

We suppose that $\widetilde{P}(Y) = 1$ and $D(\widetilde{P}\|\widetilde{\mathcal{Q}}) < +\infty$. Starting from some vector $\underline{\underline{c}}_o = (c_1^o,\dots,c_k^o)$ with positive components of sum 1, let $\underline{\underline{c}}_n = (c_1^n,\dots,c_k^n)$, $n = 1,2,\dots$ be defined recursively by

$$(4.13) \quad c_i^n = c_i^{n-1}\int \frac{d\mu_i}{d\widetilde{Q}_{n-1}}\, d\widetilde{P} \ , \quad \widetilde{Q}_n = \sum_{i=1}^{k} c_i^n \mu_i \ .$$

Then $\lim_{n\to\infty} \underline{\underline{c}}_n = \underline{\underline{c}}^*$ exists and

$$D(\widetilde{P}\|\widetilde{\mathcal{Q}}) = \lim_{n\to\infty} D(\widetilde{P}\|\widetilde{Q}_n) = D(\widetilde{P}\|\widetilde{Q}^*) \quad \text{for} \quad \widetilde{Q}^* = \sum_{i=1}^{k} c_i^*\mu_i \ ,$$

further,

$$D(\widetilde{P}\|\widetilde{Q}_n) - D(\widetilde{P}\|\widetilde{\mathcal{Q}}) \leq \max_{1\leq i\leq k} \log \frac{c_i^{n+1}}{c_i^n} \ .$$

Proof. We apply Theorem 4 (i) to $(X,\mathcal{X}) = (Z,\mathcal{Z}) \times (Y,\mathcal{Y})$ as in Lemma 4, letting T be the projection of X onto Y, and letting $\mathcal{Q}$ be the set $\mathcal{R}$ of Lemma 4, i.e., the set of all measures of form

$$(4.14) \quad Q = \sum_{i=1}^{k} c_i\delta_i \times \mu_i \ , \quad c_i \geq 0 \ , \quad \Sigma\, c_i = 1 \ .$$

For Q as in (4.14) Q^T equals the $\widetilde{Q}$ of (4.12).

$\mathcal{P} = \{P : P^T = \tilde{P}\}$ is now the set of all probability measures on $(X,\mathcal{X})$ having Y-marginal equal to $\tilde{P}$, and the iteration step $Q_{n-1} \xrightarrow{2} P_n$ is given, cf. (4.4), by

$$(4.15)\quad \frac{dP_n}{dQ_{n-1}}(i,y) = \frac{d\tilde{P}}{d\tilde{Q}_{n-1}}(y) \quad .$$

The step $P_n \xrightarrow{1} Q_n$ is now determined, cf. (4.10), by

$$(4.16)\quad c_i^n = P_n(i) = P_n(\{i\} \times Y), \quad Q_n = \sum_{i=1}^{k} c_i^n \delta_i x \mu_i \quad .$$

Combining these steps, we obtain (4.13). Thus by Theorem 4 (i), to prove

$$(4.17)\quad D(\tilde{P}\|\tilde{Q}_n) \to D(\tilde{P}\|\tilde{\mathcal{Q}})$$

for $\tilde{Q}_n$ defined in (4.13) it suffices to show that for every $P \in \mathcal{P}$, $D(P\|Q) < +\infty$ for some $Q \in \mathcal{Q}$ implies $D(P\|Q_o) < +\infty$ (for in this case (4.6) is clearly satisfied). Notice that for every Q and Q_n as in (4.14), (4.16), $Q \ll Q_n$ holds iff $c_i = 0$ whenever $c_i^n = 0$ and then

$$(4.18)\quad \frac{dQ}{dQ_n}(i,y) = \frac{c_i}{c_i^n} \quad .$$

This implies that

$$(4.19)\quad D(P\|Q_n) = \int \log\left(\frac{dP}{dQ}\,\frac{dQ}{dQ_n}\right)dP = D(P\|Q) + \sum_{i=1}^{k} P(i)\log\frac{c_i}{c_i^n}$$

where $P(i) = P(\{i\} \times Y)$. Since $c_i^o > 0$ for $i = 1,\dots,k$, this proves that $D(P\|Q_o) < +\infty$ as claimed.

The convergence of the vectors $\underline{\underline{c}}_n$ follows from Theorem 3 if Y is finite, and can be established by the same method in general. To this end, notice first that $D(\tilde{P}\| \sum_{i=1}^{k} c_i\mu_i)$ is a lower semicontinuous function of $\underline{c}$ on the simplex $c_i \geq 0$, $\sum_{i=1}^{k} c_i = 1$, as one checks with Fatou's lemma (discontinuity may occur only on the boundary). Now, for a convergent subsequence $\underline{\underline{c}}_{n_j} \to \underline{\underline{c}}^*$, say, define

$$Q^* = \sum_{i=1}^{k} c_i^* \delta_i \times \mu_i \ , \quad \tilde{Q}^* = \sum_{i=1}^{k} c_i^* \mu_i \ ,$$

and let $P^* \in \mathcal{P}$ be defined by $Q^* \xrightarrow{2} P^*$. From (4.2), the lower semicontinuity of $D(P \| \sum_{i=1}^{k} c_i \mu_i)$, and (4.17) it follows that

$$D(P^* \| Q^*) = D(\tilde{P} \| \tilde{Q}^*) = D(\tilde{P} \| \tilde{Q}) = D(\mathcal{P} \| \mathcal{Q}) \ . \tag{4.20}$$

This means, in particular, that in addition to $Q^* \xrightarrow{2} P^*$ also $P^* \xrightarrow{1} Q^*$ holds so that by Lemma 4

$$c_i^* = P^*(i) = P^*(\{i\} \times Y) \ .$$

This and (4.16) imply that

$$D(\underline{\underline{c}}^* \| \underline{\underline{c}}_n) = \sum_{i=1}^{k} c_i^* \log \frac{c_i^*}{c_i^n} \le D(P^* \| P_n) \ . \tag{4.21}$$

Further, by the three points property (3.4) and by (4.19)

$$D(P^* \| P_n) + D(P_n \| Q_{n-1}) \le D(P^* \| Q_{n-1}) =$$

$$= D(P^* \| Q^*) + \sum_{i=1}^{k} P^*(i) \log \frac{c_i^*}{c_i^{n-1}} = D(P^* \| Q^*) + D(\underline{\underline{c}}^* \| c_{n-1}) .$$

Comparing this with (4.21) and taking into account (4.20) and (4.15) we obtain

$$D(\underline{\underline{c}}^* \| \underline{\underline{c}}_n) \le D(\underline{\underline{c}}^* \| \underline{\underline{c}}_{n-1}) - [D(\tilde{P} \| Q_{n-1}) - D(\tilde{P} \| \tilde{Q})] \le \tag{4.22}$$

$$\le D(\underline{\underline{c}}^* \| \underline{\underline{c}}_{n-1}) \quad .$$

Since the assumption $\underline{\underline{c}}_{n_j} \to \underline{\underline{c}}^*$ implies $D(\underline{\underline{c}}^* \| \underline{\underline{c}}_{n_j}) \to 0$, by (4.22) we have proved that actually $D(\underline{\underline{c}}^* \| \underline{\underline{c}}_n) \to 0$, i.e., $\underline{\underline{c}}_n \to \underline{\underline{c}}^*$. Also the last assertion of the Theorem follows from (4.22):

$$D(\tilde{P} \| Q_{n-1}) - D(\tilde{P} \| \tilde{Q}) \le D(\underline{\underline{c}}^* \| \underline{\underline{c}}_{n-1}) - D(\underline{\underline{c}}^* \| c_n) \le$$

$$\max_{1\le i\le k} \log \frac{c_i^n}{c_i^{n-1}} \quad .$$

Remark: The algorithm of Theorem 5 is widely used in the literature for decomposition of mixtures, cf. the discussion in the next section (at the end of part A).

5. Applications

(A) Maximum likelihood from incomplete data

Let $\mathcal{S} = (X,\mathcal{X},\mathcal{Q})$ be a statistical space, i.e. let $(X,\mathcal{X})$ be a measurable space and $\mathcal{Q}$ a set of probability distributions on $(X,\mathcal{X})$. Let μ be some σ-finite measure dominating $\mathcal{Q}$ and write $q = \frac{dQ}{d\mu}$ (it is assumed that fixed versions of the densities q are used). Let $\underline{\underline{x}} = (x_1,\dots,x_N)$ be an i.i.d. sample from $\mathcal{S}$; then the log-likelihood function is

$$(5.1) \qquad L(Q) = \frac{1}{N}\sum_{i=1}^{N} \log q(x_i) = \int \log q \, dP_{\underline{\underline{x}}}$$

where $P_{\underline{\underline{x}}}$ is the empirical distribution of the sample, i.e.,

$$P_{\underline{\underline{x}}} = \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i}$$

(here δ_x stands for the probability measure concentrated in x).

We say that we are given incomplete data from $\mathcal{S}$ if instead of $\underline{\underline{x}} = (x_1,\dots,x_N)$ only the sample $\underline{\underline{y}} = (y_1,\dots,y_N)$ is available, where

$$y_i = T(x_i) \qquad i = 1,\dots,N ,$$

for some measurable mapping T of $(X,\mathcal{X})$ into another measurable space $(Y,\mathcal{Y})$. Let the set $\mathcal{Q}^T$ be defined by (4.1) and let $\tilde{\mu}$ be a dominating measure for $\mathcal{Q}^T$, with densities $q^T = \frac{dQ^T}{d\tilde{\mu}}$, $Q^T \in \mathcal{Q}^T$. Then the likelihood function based on

widely used in the
cf. the discussion

data

e, i.e. let $(X,\mathcal{X})$
ability distribu-
measure dominating
fixed versions of
$\ldots,x_N)$ be an
hood function is

the sample, i.e.,

re concentrated

from $\mathcal{S}$ if
e $\underline{\underline{y}} = (y_1,\ldots,y_N)$

into another
be defined by (4.1)
, with densities
nction based on

the incomplete data $\underline{\underline{y}}$ is

$$L(Q^T) = \frac{1}{N}\sum_{i=1}^{N} \log q^T(y_i) = \int \log q^T(y)\, dP_{\underline{\underline{y}}} \tag{5.2}$$

where, of course, $P_{\underline{\underline{y}}} = P^T_{\underline{\underline{x}}}$.

We are interested in finding a maximum likelihood estimate of Q on the basis of $\underline{\underline{y}}$, i.e., a $Q \in \mathcal{Q}$ maximizing (5.2), providing that a maximum likelihood estimate of Q on the basis of $\underline{\underline{x}}$ could be "easily" found. More exactly, we assume that to every measure ν on $(X,\mathcal{X})$ a $Q \in \mathcal{Q}$ maximizing $\int \log q\, d\nu$ can be "easily" found. Now we show how Theorem 4 can be applied to this problem.

Let us suppose first that Y is a finite set. Then, choosing for $\tilde{\mu}$ the counting measure, (5.2) becomes

$$L(Q^T) = \frac{1}{N}\sum_{i=1}^{N} \log Q^T(y_i) = \int \log Q^T(y)\, dP_{\underline{\underline{y}}} \ . \tag{5.3}$$

Since this differs but in a constant term from $-D(P_{\underline{\underline{y}}}\|Q^T)$, the maximum likelihood estimate of Q^T achieves $D(P_{\underline{\underline{y}}}\|\mathcal{Q}^T)$. Hence, by Theorem 4 (i), one can use the alternating minimization procedure to maximizing $L(Q^T)$, providing $\mathcal{Q}$ is convex. Starting from some $Q_o \in \mathcal{Q}$ define the sequences $\{P_n\}$ and $\{Q_n\}$ with densities $\{p_n\}$, $\{q_n\}$ such that (in accordance with (4.4), where now $\tilde{P} = P_{\underline{\underline{y}}}$)

$$p_{n+1}(x) = \frac{P_{\underline{\underline{y}}}(T(x))}{Q_n^T(T(x))}\, q_n(x)$$

and Q_{n+1} minimizes $D(P_{n+1}\|Q)$. By Theorem 4 (i) we then have $D(P_{\underline{\underline{y}}}\|Q_n^T) \to D(P_{\underline{\underline{y}}}\|\mathcal{Q}^T)$, i.e.,

$$L(Q_n^T) \to \sup_{Q\in\mathcal{Q}} L(Q^T) \ . \tag{5.4}$$

As to the determination of Q_{n+1}, suppose that $\int \log p_{n+1} dP_{n+1}$ is finite (a sufficient condition for this is that $\int \log q\, dQ$ be finite for every $Q \in \mathcal{Q}$). Then we can write

$$D(P_{n+1} \| Q) = \int \log p_{n+1}\, dP_{n+1} - \int \log q\, dP_{n+1}$$

where the second term equals

$$(5.5) \qquad L(Q|Q_n) = \int \log q \cdot \frac{P_{\underline{\underline{y}}}(T(x))}{Q_n^T(T(x))}\, dQ_n =$$

$$= \frac{1}{N} \sum_{i=1}^{N} \int_{T^{-1}(y_i)} \frac{\log q}{Q_n^T(y_i)}\, dQ_n \; .$$

It follows that $D(P_{n+1} \| Q)$ is maximized by the same Q which maximizes $L(Q|Q_n)$; this can, by assumption, be easily found, since (5.5) is of form $\int \log q\, d\nu$.

Notice that (5.5) can also be written as

$$(5.6) \qquad L(Q|Q_n) = E_{Q_n}(L(Q)|\underline{\underline{y}}) \; ,$$

i.e., as the conditional expectation of the log-likelihood (5.1) given the sample $\underline{\underline{y}}$ if the underlying distribution is Q_n.

The iteration consisting of the steps of maximizing $L(Q|Q_n)$ in Q and then letting the maximizing Q be Q_{n+1} is widely used in statistical practice for maximizing the log-likelihood $L(Q^T)$, which is justified if (5.4) holds. In this generality, this iterative method was suggested by Dempster, Laird and Rubin (1977) under the name "EM algorithm", where a long list of earlier papers using this method in special cases is also provided, dating back to Hartley (1958).

The above reasoning proves the desirable convergence property (5.4) of the EM algorithm whenever $\mathcal{Q}$ is convex and Y is finite; further, in this case (5.4) easily implies the convergence of $Q_n^T(y)$ for every y in the sample $\underline{\underline{y}}$. The restriction to finite Y will be removed soon. The sequence $L(Q_n^T)$ generated by the EM algorithm is monotone non-decreasing also for non-convex $\mathcal{Q}$. This fact, proved in Dempster, Laird and Rubin (1977), follows from our approach for free. On the other hand, the desired convergence relation (5.4) does not hold in general. In the literature known to us the most complete discussion of the problem of convergence is Wu (1983).

When Y is not finite, the EM algorithm can still be fitted into our framework as follows. Suppose that the dominating measure μ of $\mathcal{Q}$ can be represented by a σ-finite measure $\tilde{\mu}$ on $(Y,\mathcal{Y})$ and a family of σ-finite measures $\eta(\cdot|y)$ on $(X,\mathcal{X})$ in the sense that $\eta(A|\cdot)$ is $\mathcal{Y}$-measurable for every $A \in \mathcal{X}$ and

$$\mu(A) = \int \eta(A|y)\tilde{\mu}(dy) \quad ,$$

where $\eta(\cdot|y)$ is concentrated on $T^{-1}(y)$. Then $Q^T \ll \tilde{\mu}$ for every $Q \in \mathcal{Q}$, with density

$$q^T(y) = \int q(x)\,\eta(dx|y) \quad .$$

Further,

$$q(x|y) = \begin{cases} \dfrac{q(x)}{q^T(y)} & \text{if } \; T(x) = y \\ 0 & \text{else} \end{cases}$$

is the conditional density of Q given $T(x) = y$ with respect to $\eta(\cdot|y)$.

Now, given the sample $\underline{\underline{y}} = (y_1,\dots,y_N)$, let $\tilde{Y}$ be the set of those $y \in Y$ which are in the sample and let a measure $\hat{\mu}$ concentrated on $T^{-1}\tilde{Y}$ be defined by

$$\hat{\mu} = \sum_{y\in\tilde{Y}} \eta(\cdot|y) \quad .$$

For each $Q \in \mathcal{Q}$ let $\hat{Q}$ be the measure with $\hat{\mu}$-density equal to the μ-density of Q. Then

$$\hat{Q}^T(y_i) = q^T(y_i) \quad i = 1,\dots,N$$

and therefore the log-likelihood function (5.2) has the form (5.3) with Q replaced by $\hat{Q}$. It follows that the arguments given for the case of finite Y apply to maximizing (5.2) also in the general case, if we replace $\mathcal{Q}$ by $\hat{\mathcal{Q}} = \{\hat{Q} : Q \in \mathcal{Q}\}$; the elements of $\hat{\mathcal{Q}}$ need not be probability measures, but this affects nothing. Notice, in particular, that (5.5) (with Q and Q_n replaced by $\hat{Q}$ and $\hat{Q}_n$) becomes

$$L(\hat{Q}|\hat{Q}_n) = \frac{1}{N}\sum_{i=1}^{N} \int_{T^{-1}(y_i)} \frac{\log q(x)}{q_n^T(y_i)}\, q_n \, d\eta(\cdot|y_i)$$

which is the same as the conditional log-likelihood function (5.6). Hence maximizing $L(\hat{Q}|\hat{Q}_n)$ with respect to $\hat{Q}$ is the same as maximizing $L(Q|Q_n)$ with respect to Q, as required by the EM algorithm.

An important field of applications of the EM algorithm is the decomposition of mixtures, particularly in the case when the number of components is large so that other numerical methods cf. Zangwill (1969) are not feasible. Formally, the problem is a special case of the model in this section which is obtained by setting $X = Z \times Y$, $T(z,y) = Y$ (where $Z = \{1,\dots,k\}$, and k is the "number of components"), and letting $\mathcal{Q}$ to be the set of measures (4.14), where $\mu_1,\dots,\mu_k$ are known probability measures and $c_1,\dots,c_k$ are unknown parameters. (The case when the μ_i 's contain unknown parameters has also been investigated in the literature, including the references below, but is not covered by our convergence theorem.) To our knowledge, the iteration (4.13) for getting the ML estimate of the parameters $c_1,\dots,c_k$ was first suggested by Hasselblad (1966); another relevant early reference is Šlezinger (1968). An extensive list of references can be found in Grim (1982). The convergence results on the algorithm (4.13) available in the literature do not seem to yield convergence of the weight vectors $\underline{\underline{c}}_n$ without the so-called identifiability condition. (The latter means that the representation (4.12) of elements of $\tilde{\mathcal{Q}}$ is unique, in this case the optimizing vector $\underline{\underline{c}}^*$ is unique and $\underline{\underline{c}}_n \to \underline{\underline{c}}^*$ easily follows.) Thus our Theorem 5 appears to be a stronger convergence result than those known priviously, even though the strengthening, i.e., the convergence of the very weight vectors $\underline{\underline{c}}_n$, is more of theoretical than of practical interest. As a sample of practical applications we mention remote sensing (Peters and Coberly (1976)), tomography (Shepp and Vardi (1982)) and cluster analysis of congenital malformations (Czeizel, Telegdi and Tusnády (1984)).

(B) Channel capacity. A memoryless channel with finite input alphabet X and finite output alphabet Y is determined by a stochastic matrix $W : X \to Y$, i.e., a family of distributions $\{W(\cdot|x)\}_{x\in X}$ on Y. We follow the notation of the book Csiszár and Körner (1981) and refer to the same book for the information-theoretic significance of the concepts below. The capacity of this channel is $C(W) = \max I(P,W)$ where

$$(5.7) \qquad I(P,W) = \sum_{x,y} P(x)W(y|x) \log \frac{W(y|x)}{PW(y)}$$

and the maximum is taken for all distributions P on X; here

$$PW(y) = \sum_{x} P(x)W(y|x) \quad .$$

Arimoto (1972) and Blahut (1972) suggested an iterative algorithm for computing $C(W)$, based on the observation that it can be written as a double maximum:

$$(5.8) \qquad C(W) = \max_{P,\Phi} \sum_{x,y} P(x)W(y|x) \log \frac{\Phi(x|y)}{P(x)} \quad .$$

Here Φ ranges over all stochastic matrices $\Phi : Y \to X$. For fixed P, the maximum in Φ equals $I(P,W)$ and it is attained by

$$\Phi(x|y) = \frac{P(x)W(y|x)}{PW(y)} \quad .$$

The maximizing P for a fixed Φ can also be readily given by an explicit formula. The mentioned algorithm consists in alternating minimization with respect to P and Φ, starting from some P_o such that $P_o(x) > 0$ for every $x \in X$.

Now, let $\mathcal{P}$ resp. $\mathcal{Q}$ be the set of all measures on $X \times Y$ of the form $P(x)W(y|x)$ resp. $\Phi(x|y)W(y|x)$ where P is any distribution on X and $\Phi : Y \to X$ is any stochastic matrix. Notice that $\mathcal{P}$ consists of probability distributions, while $\mathcal{Q}$ does not. Clearly, (5.8) is equivalent to

$$C(W) = -D(\mathcal{P}\|\mathcal{Q}) \quad ,$$

and the Arimoto-Blahut algorithm is just the alternating minimization procedure for $D(P\|Q)$, $P \in \mathcal{P}$, $Q \in \mathcal{Q}$. Hence Theorem 3 contains the result proved by Arimoto (1972) that the iteration converges to C(W) and, in addition, the distributions P_n on X constructed in course of the iteration converge to some P^* such that $I(P^*,W) = C(W)$. The first part of this result follows from Theorem 3 also for countable alphabets X and Y (understanding C(W) as the supremum of I(P,W), for now the maximum need not be attained) while the convergence of the P_n 's no longer follows.

A variant of the capacity computing algorithm has been suggested by Jimbo and Kunisawa (1979) for computing

$$\max_P \frac{I(P,W)}{c(P)} \quad ,$$

the so-called capacity per unit cost. Here

$$c(P) = \sum_{x \in X} P(x)\, c(x) \quad ,$$

where c(x) is a given positive function interpreted as the cost of transmission (or duration) of the symbol $x \in X$. Their method is equivalent to the alternating minimization procedure for $D(P\|Q)/c(P)$, $P \in \mathcal{P}$, $Q \in \mathcal{Q}$ where $\mathcal{P}$ and $\mathcal{Q}$

are the same as above. By the last remark in Section 3, the three and four points properties are valid for $d(P,Q) = D(P\|Q)/c(P)$, hence our general results cover also this case yielding the convergence theorem of Jimbo and Kunisawa (1979).

The capacity computing algorithm fits into our framework also in another way, letting $\mathcal{P}$ resp. $\mathcal{Q}$ be the set of all distributions on X resp. of all stochastic matrices $\Phi : Y \to X$, and considering directly

$$d(P,\Phi) = \sum_{x,y} P(x)W(y|x)\log\frac{\Phi(x|y)}{P(x)}\ , \ P \in \mathcal{P},\ \Phi \in \mathcal{Q}\ .$$

For this function, the three and four points properties hold with $\delta(P,P') = D(P\|P')$ for every $P \in \mathcal{P}$, thus Theorem 2 applies. This shows, in particular, that the convergence results remain valid even for channels with abstract output alphabet, as long as the input alphabet X is finite. Even this finiteness assumption can be dispensed with if convergence to $C(W)$ is all what is required, without any convergence statement on the P_n 's.

(C) <u>Rate-distortion functions</u>. Let X and Y be finite sets and $\rho(x,y)$ be a non-negative valued function on $X \times Y$. The rate-distortion function of a distribution P on X is

$$R(\Delta) = \min_{W:\rho(P,W)\leq\Delta} I(P,W) \quad (\Delta \geq 0)$$

where $I(P,W)$ is defined by (5.7),

$$\rho(P,W) = \sum_{x,y} P(x)W(y|x)\rho(x,y)\ ,$$

and the minimization refers to all stochastic matrices $W : X \to Y$ satisfying the indicated constraint. For evaluating the function $R(\Delta)$, the standard approach is to introduce a Lagrange multiplier δ, thereby reducing the problem to the evaluation of the function

$$(5.9) \quad G(\delta) = \min_{W}[I(P,W) + \delta\rho(P,W)] =$$

$$= \min_{W} \sum_{x,y} P(x)W(y|x) \log \frac{W(y|x)}{PW(y)\exp[-\delta\rho(x,y)]} .$$

An iterative algorithm for evaluating $G(\delta)$, $\delta \geq 0$, was suggested by Blahut (1972). It is based on the observation that $G(\delta)$ can be written as a double minimum:

$$G(\delta) = \min_{W,Q} \sum_{x,y} P(x)W(y|x) \log \frac{W(y|x)}{Q(y)\exp[-\delta\rho(x,y)]} ,$$

where Q ranges over the distributions on Y. For fixed W, the minimum is attained when $Q = PW$, and the minimizing W for fixed Q can also be given explicitly. The mentioned algorithm consists in alternating minimizations with respect to W and Q. Denoting by $\mathcal{P}$ resp. $\mathcal{Q}$ the set of measures on $X \times Y$ of form $P(x)W(y|x)$ (for some $W : X \to Y$) resp. $P(x)Q(y)\exp[-\delta\rho(x,y)]$ (for some distribution Q on Y), it is obvious that Theorem 3 applies to this case. It gives the result proved by Csiszár (1974) that Blahut's iteration does converge to $G(\delta)$, and, in addition, the stochastic matrices $W_n : X \to Y$ obtained in course of the iteration converge to some $W^* : X \to Y$ which attains the minimum in (5.9). Perhaps it is worth pointing out that, in contrast with the previous case, now it is the set X whose finiteness is not needed for the result.

(D) Investment portfolio with maximum expected log return. Let $X_1, X_2, \ldots, X_k$ be non-negative valued random variables with finite expectations. X_j is interpreted as the yield (per one dollar investment) of the j'th one of k given stocks where $X_j > 1$ means gain while $X_j < 1$ means loss, and an investor is supposed to invest fractions $c_1, \ldots, c_k$ of his total invested capital into these stocks. The vector $\underline{c} = (c_1, \ldots, c_k)$ with non-negative components of sum 1 is called the investment portfolio. Cover (1981) suggested an iterative algorithm for computing the portfolio yielding maximum expected log return, i.e., maximizing $E \log \sum_{j=1}^{k} c_j X_j$

(supposing that the joint distributuion of the X_j 's is known). For literature substantiating that this is the right optimality criterion we refer to Cover (1981). Cover's algorithm is given by

$$(5.10)\qquad c_j^{n+1} = c_j^n \, E \frac{X_j}{\sum_{j=1}^{k} c_j^n X_j} \quad , \qquad n = 0,1,\ldots$$

where $(c_1^o,\ldots,c_k^o)$ can be any vector with positive components.

Now let Y be the set of all k-dimensional vectors $y = (r_1,\ldots,r_k)$ with non-negative components. Let the measure $\widetilde{P}$ on (the Borel σ-algebra of) Y be the joint distribution of $X_1,\ldots,X_k$, and let the measures $\mu_1,\ldots,\mu_k$ on Y be defined by

$$\frac{d\mu_i}{d\widetilde{P}}(r_1,\ldots,r_k) = r_i \qquad i = 1,\ldots,k \ .$$

Then

$$E \log \sum_{j=1}^{k} c_j X_j = \int (\log \sum_{j=1}^{k} c_j r_j) d\widetilde{P} = -D(\widetilde{P} \| \widetilde{Q})$$

where

$$\widetilde{Q} = \sum_{j=1}^{k} c_j \mu_j \quad .$$

Thus maximizing the expected log return is the same as minimizing $D(\widetilde{P} \| \widetilde{Q})$. Obviously, Cover's algorithm (5.10) is the same as (4.13) applied to the present setup. In particular, Theorem 5 contains the result of Cover (1981) that the iteration (5.10) is convergent to a portfolio $\underline{\underline{c}}^*$ yielding maximum expected log return.

REFERENCES

Arimoto, S.: An algorithm for computing the capacity of arbitrary discrete memoryless channels. IEEE Trans. Inform. Theory 18, 14 - 20, (1972).

Blahut, R.E.: Computation of channel capacity and rate-distortion functions. IEEE Trans. Inform. Theory 18, 460 - 473, (1972).

Cheney, W. and Goldstein, A.A.: Proximity maps for convex sets. Proc. Amer. Math. Soc. 10, 448 - 450, (1959).

Cover, T.: An algorithm for maximizing expected log investment. Technical report No. 46. Stanford University, Dept. Statist., (1981). To appear in IEEE Trans. Inform. Theory.

Csiszár, I.: On the computation of rate distortion function. IEEE Trans. Inform. Theory 20, 122 - 124, (1974).

Csiszár, I.: I-divergence geometry of probability distributions and minimization problems. Annals of Probability 3, 146 - 158, (1975).

Csiszár, I. and Körner, J.: Information Theory: Coding Theorems for Discrete Memoryless Systems. Academic, New York, (1981).

Czeizel, E., Telegdi, L. and Tusnády, G.: Multiple Congenital Abnormalities. Publishing House of the Hungarian Academy of Sciences, Budapest, (1984).

Dempster, A.P., Laird, N.M. and Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. J. Roy. Statist. Soc. Ser. B, 39, 1 - 22, (1977).

Grim, J.: On numerical evaluation of maximum likelihood estimates for finite mixtures of distributions. Kybernetika 18, 173 - 190, (1982).

Hartley, H.O.: Maximum likelihood estimation from incomplete

capacity of
ls. IEEE Trans.

y and rate-
form. Theory 18,

ps for convex
- 450, (1959).

ted log invest-
ord University,
IEEE Trans.

ortion function.
124, (1974).

ility distribu-
als of Proba-

ry oding
tems. Academic,

ltiple Congenital
he Hungarian
).

Maximum likeli-
algorithm. J. Roy.
977).

likelihood
tributions.

from incomplete

data. Biometrics 14, 174 - 194, (1958).

Hasselblad, V.: Estimation of parameters for a mixture of normal distributions. Technometrics 8, 432 - 444, (1966).

Jimbo, M. and Kunisawa, K.: An iteration method for calculating the relative capacity. Information and Control 43, 216 - 223, (1979).

Peters, C. and Coberly, W.A.: The numerical evaluation of the maximum likelihood estimate of mixture distributions. Commun. Statist. Theor. Meth. A5, 1127 - 1135, (1976).

Rényi, A.: On measures of entropy and information. Proc. 4th Berkeley Symposium, Vol. 1, 547 - 561. Univ. of Calif. Press, Berkeley, (1961).

Shepp, L.A. and Vardi, Y.: Maximum likelihood reconstruction for emission tomography. Preprint, Bell Laboratories, Murray Hill NJ, (1982).

Šlezinger, M.I.: Interconnection of learning and unsupervised learning in pattern recognition (in Russian). Kibernetika (Kiev), 1968, No. 2, 81 - 88, (1968).

Wu, Chien-Fu: On the convergence properties of the EM algorithm. Ann. Statist. 11, 95 - 103, (1983).

Zangwill, W.: Nonlinear Programming: A Unified Approach. Prentice-Hall, Englewood Cliffs, (1969).

Mathematical Institute of the
Hungarian Academy of Sciences
H-1053 Budapest, Reáltanoda u. 13 - 15
Hungary