KYBERNETIKA ČÍSLO 2, ROČNÍK 6/1970

Information–Theoretic Approach to Measurement Reduction Problems*

ALBERT PEREZ

Considering the measurement problem as a statistical decision problem subordinated to a more basic one there are applied some information-theoretic methods facilitating the task of defining what to be measured and with what accuracy in order to achieve the required overall decision efficiency with a set of measurement data as reduced as possible.

In this paper, the measurement problem is conceived as a statistical decision problem subordinated to a more basic one. In face of a decision task (including prediction, control, etc.), it is the need of efficient decision procedures which leads to the collection of as more of data as possible relevant to the above task, namely, through the measurement of some set of parameters serving to characterize the decision situation. In this context, it is possible to distinguish two aspects of the measurement problem. The first one, more usual, is related to the task of measuring something a priori given, namely, to the task of estimating the values of the above set of parameters with the accuracy needed by the original decision task. Given the boundedness of the measurement capabilities at our disposal, there is a tendency to reduce as far as possible the volume of data (and the measurement costs for their acquisition) sufficient for such an estimation. It is clear that the efficiency of a data reduction procedure closely depends on the degree of exploitation of the information on the set of parameters which may be extracted from the measurement data. The information-theoretic approach to data reduction problems developed by the author consists essentially to estimate in information-theoretical terms (not necessarily of Shannon's type) the loss of decision quality (measurement accuracy) implied by different versions of data reduction. The methods presented may, thus, facilitate the task of economizing measurement capabilities in measuring a given object (i.e. in estimating a given set of parameters) with the required accuracy. This is made on the base of what it would be possible to call "essential observables" in the sense that the set of

* Paper presented at IMECO Symposium in Pezinok near Bratislava, 7—10 October 1969.

data obtained by their observation is the more reduced one among all the similar sets containing, as to the given set of parameters, an information sufficient for their estimation with the required accuracy. This concerns the first aspect of the measurement problem as envisaged in this paper.

The second aspect of the measurement problem is related to the task of *what* is necessary to measure in order to characterize as economically as possible the decision situation as required by the original decision task (to which the measurement task is subordinated). This question arises, of course, only if the parameter set introduced above is not a priori given and, thus, a possibility of selection of this set exists. The methods of treating this problem of reducibility of the parameter space are to a certain extent similar to those applied in the study of reducibility of the observation space mentioned above.

1. PRELIMINARIES

As said in the introduction, the measurement problem is conceived as a statistical decision problem subordinated to a more basic one. In the sequel, by $X = X_1 \times \times \ldots \times X_s$ will be denoted the "complete" parameter space, Cartesian product of s component spaces-parameters, $X_1, X_2, \ldots, X_s$. Similarly, by $Y = Y_1 \times Y_2 \times \ldots \times Y_n$ will be denoted the original (unreduced) sample or observation space, Cartesian product of n components-observables, $Y_1, Y_2, \ldots, Y_n$. In general, to every realized value x of X there corresponds a probability distribution $P_{Y/x}$ on the observation space Y, indicating the probability of possible values taken by the "complex observable" Y. The system $\{P_{Y/x}, x \in X\}$ of conditional probability distributions represents, thus, the so-called "observation channel", characterizing, in particular, the measurement devices from the accuracy point of view. The dynamic aspect may be taken in account by considering suitable time-sequences of such channels.

Let, further, D be the decision space (i.e. the set of possible decisions) corresponding to the original (basic) decision problem. By $w(x, d)$, $x \in X$, $d \in D$, will be denoted the so-called weight or distance or loss function, indicating the "loss" or "inadequateness" implied by taking a decision $d \in D$ in a decision situation characterized by the value $x \in X$ of the "complex parameter". The latter is no directly observable, in general, but only through the observation channel $\{P_{X/x}, x \in X\}$, so that it can be only estimated on the base of the observed value y of the "complex observable" Y. As a consequence, the choice of the decision d in D may be made only on the base of the observed $y \in Y$, either directly or through the prior estimation (measurement) of the value x of the complex parameter X. Whatever be the case, the decision procedure may be represented by a function (called "decision function") $b(y)$ of $y \in Y$, taking its values on D. Sometimes it is useful to decide by applying a randomized decision function (mixed strategy) but here we consider only "pure" decision functions.

The *risk* corresponding to the parameter value x and to the decision function b is

given by

$$r(x, b) = \int_Y w(x, b(y))\, \mathrm{d}P_{Y/x}(y)\,. \tag{1}$$

If P_X is the so-called a priori probability distribution on X, the *average risk* corresponding to the decision function b is given by (cf. (1))

$$r(b) = \int_X r(x, b)\, \mathrm{d}P_X(x)\,. \tag{2}$$

If P is the simultaneous probability distribution on the Cartesian product $X \times Y$, generated by P_X and the channel $\{P_{Y/x}, x \in X\}$, the so-called *Bayes risk* is given by (cf. (2))

$$r_0(P) = r(b_0) = \min_{b \in B} r(b)\,. \tag{3}$$

(In the sequel we shall suppose that always there exists an optimal or Bayes decision function b_0 in the set B of all possible decision functions.)

Let us, now, recall the definition of some infomation-theoretic concepts applied in the sequel.

The Shannon's information on X contained in Y is given by

$$I(X, Y) = I(P) = \int_{X \times Y} \log \frac{\mathrm{d}P(x, y)}{\mathrm{d}(P_X \times P_Y)}\, \mathrm{d}P(x, y) \tag{4}$$

where $\mathrm{d}P/\mathrm{d}(P_X \times P_Y)$ is the density of P with respect to the product measure of its marginals P_X and P_Y. If this density does not exist, then $I(X, Y) = \infty$.

As well known, $I(X, Y) \geqq 0$ with equality if, and only if, X and Y are stochastically independent, i.e. if, and only if, $P = P_X \times P_Y$. Further, by reducing X to X' (for instance, in our case, by retaining only some components X_i of X and by rejecting the rest, represented in the sequel by X'' so that $X = X' \times X''$) and, similarly, by reducing Y to Y', it holds

$$I(X', Y') \leqq I(X, Y)\,, \tag{5}$$

the sign of equality taking place if, and only if, the transformations (reductions) $S(X) = X'$ and $T(Y) = Y'$ are "sufficient" with respect to $\{P, P_X \times P_Y\}$ in the sense of mathematical statistics.

The concept of information may be considered as a special case of the concept of generalized entropy of one probability measure P with respect to another probability measure Q defined on the same measurable space. Thus, the Shannon's generalized entropy $H(P, Q) = \infty$ unless P is absolutely continuous with respect to Q, the cor-

responding density being $\mathrm{d}P/\mathrm{d}Q = u$. Then

$$H(P, Q) = \int \log u \, \mathrm{d}P = \int u \log u \, \mathrm{d}Q \,. \tag{6}$$

If the convex function $u \log u$ is replaced by a more general continuous and strict convex function $f(u)$, we obtain a more general concept, the so-called *generalized f-entropy* of P with respect to Q

$$H_f(P, Q) = \int f(u) \, \mathrm{d}Q \,, \tag{7}$$

where by u we denote the ratio of the densities of P and Q with respect to a dominating measure R:

$$u = \frac{\mathrm{d}P/\mathrm{d}R}{\mathrm{d}Q/\mathrm{d}R} \,. \tag{7a}$$

Except of the well-known additivity property, the other fundamental properties are essentially conserved on passing from the Shannon's to the generalized f-entropies. If, namely, the "statistical hypotheses" P and Q are transformed to P' and Q', respectively, due to the fact that their common sample space is not observed directly but through some observation channel (which may be also some ordinary transformation of the sample space), then it holds (cf. (5))

$$H_f(P', Q') \leqq H_f(P, Q) \,, \tag{8}$$

the sign of equality taking place if, and only if, the observation channel in question is (in some generalized sense) "sufficient" with respect to the system $\{P, Q\}$. This relation, thus, exprimes the well-known fact that by reducing the space or the accuracy of the observations we cannot increase the discernability. For more details see references [1, 2, 3, 4, 5].

2. FIRST ASPECT OF THE MEASUREMENT PROBLEM: REDUCTION OF THE OBSERVATION SPACE Y

As said in the introduction, the boundedness of the measurement capabilities in face of the task to measure a given object, i.e. to estimate with a given accuracy the value of a given "complex parameter" X, represented by the set of parameters-components $X_1, X_2, \ldots, X_s$, (cf. section 1), leads to the tendency to reduce as far as possible the set of "observables" $Y_1, Y_2, \ldots, Y_n$, (cf. section 1), as well as the accuracy of their values, sufficient for the estimation of X with the needed accuracy. In the language of σ-algebras of subsets of the spaces X and Y it would be possible to express all this in a more suitable and precise form (cf. references above). For the sake of simplicity,

however, we shall try to expose here our ideas in terms of sets of observables and parameters.

Let, thus, reduce the original set of observables Y to Y' by rejecting some of its components Y_i. As a consequence, the observation channel $\{P_{Y/x}, x \in X\}$, (cf. section 1), is transformed to the reduced observation channel $\{P_{Y'/x}, x \in X\}$, the marginal probability distribution P_Y becomes $P_{Y'}$ and the simultaneous probability distribution P becomes P' on $X \times Y'$.

Let $\tilde{P}'_{Y/x}$ be, for every $x \in X$, defined by the relation $\mathrm{d}\tilde{P}'_{Y/x} = (\mathrm{d}P_{Y'/x}/\mathrm{d}P_{Y'})\,\mathrm{d}P_Y$, (in the sequel, we assume the existance of all the densities used; this is assured by the finiteness of the Shannon's informaiton we suppose throughout the paper), and, thus, representing an extension of $P_{Y'/x}$ from Y' to Y. Similarly, let $\tilde{P}'$ be the corresponding extension of P' from $X \times Y'$ to $X \times Y$. It is possible to prove, (cf. [2]), that the Bayes risk as well as the Shannon's information, defined by (3) and (4), are preserved by such an extension, i.e.

$$r_0(P') = r_0(\tilde{P}'), \tag{9}$$

$$I(P') = I(\tilde{P}'). \tag{10}$$

Moreover, for the Shannon's generalized entropy (cf. (6)) it holds

$$H(P, \tilde{P}') = I(P) - I(P') = \min_{Q} H(P, Q), \tag{11}$$

where Q is an arbitrary extension of P' from $X \times Y'$ to $X \times Y$ preserving the Shannon's information.

On the base of [2–5], it is possible to write under very general conditions that the risk increment (cf. (1)), the average risk increment (cf. (2)), and the Bayes risk increment (cf. (3)), on passing from a probability law P to a probability law Q (for instance, change of the measurement device, reduction of the set of observables, and so on), satisfy the inequalities:

$$r(x, b, Q) - r(x, b, P) \leqq \sqrt{[2H(P_{Y/x}, Q_{Y/x})\, R(x, w^2, Q)]}, \tag{12}$$

$$r(b, Q) - r(b, P) \leqq \sqrt{[2H(P, Q)\, R(w^2, Q)]}, \tag{13}$$

$$r_0(Q) - r_0(P) \leqq \sqrt{[2H(P, Q)\, R_0(w^2, Q)]}, \tag{14}$$

where

$$R(x, w^2, Q) = \int (w(x, b(y)) - r(x, b, Q))^2\, \mathrm{d}Q_{Y/x}(y),$$

$$R(w^2, Q) = \int (w(x, b(y)) - r(b, Q))^2\, \mathrm{d}Q(x, y),$$

$$R_0(w^2, Q) = \int (w(x, b_Q(y)) - r_0(Q))^2\, \mathrm{d}Q(x, y).$$

Here b_Q is a Bayes decision function corresponding to Q.

In (12), (13), (14), the quantities $2H(P_{Y/x}, Q_{Y/x})$ and $2H(P, Q)$ may be replaced (sometimes, with advantage) respectively by the quantities $H_2(P_{Y/x}, Q_{Y/x}) - 1$ and $H_2(P, Q) - 1$, where $H_2(P, Q)$ is the so-called generalized entropy of second order of P with respect to Q,

$$H_2(P, Q) = \int u^2 \, \mathrm{d}Q \,. \tag{15}$$

Here u is the density of P with respect to Q if it exists, otherwise is the ratio of densities introduced in the definition (8) of the generalized f-entropy. In the present case $f(u) = u^2$.

Further, in the case of the reduction of the set of observables from Y to Y', Q in (12), (13) and (14) will be replaced by $\tilde{P}'$, introduced above. According to (9)–(11), in particular from (14) we throw:

$$0 \leqq r_0(P') - r_0(P) \leqq \sqrt{[2(I(P) - I(P'))\, R_0(w^2, P')]} \,. \tag{16}$$

In [2] we introduced the concept of ε-sufficiency, generalizing the concept of sufficiency and of sufficient statistics in mathematical statistics, in terms of the Shannon's information as follows: A reduction of the sample (observation) space is **ε-sufficient** if the corresponding loss of information satisfies the inequality

$$I(P) - I(P') \leqq \varepsilon \quad (\varepsilon \geqq 0) \,. \tag{17}$$

As for the searching of a minimal sufficient statistics, there is possible to establish algorithms for the searching of a minimal ε-sufficient statistics for $\varepsilon > 0$ given. Note that the loss of information is zero if, and only if, the corresponding transformation is sufficient (cf. (5)). In a similar way, it is possible to introduce the concept of ε-sufficiency in terms of generalized f-entropies (cf. (7) and (8)).

If the weight function w is of the type "0 or 1" so that the average risk is the so-called probability of error, it is possible to derive many interesting estimates (cf. [4, 5]) from the following general inequality

$$e_Q f(e_P/e_Q) + (1 - e_Q) f((1 - e_P)/(1 - e_Q)) \leqq H_f(P, Q) \tag{18}$$

for every continuous function $f(u)$ strictly convex on $[0, \infty)$. Here e_Q and e_P are the minimal error probabilities corresponding to the probability laws Q and P, respectively. As before, in the case of reduction Q will be replaced by $\tilde{P}'$. Note that $e_{P'} = e_{\tilde{P}'}$.

Very important is the special case of discriminating two statistical hypotheses P_0 and Q_0 on the base of a sequence $Y_1, Y_2, \ldots, Y_n$ of independent observables equally distributed according to the probability distribution P_0 (if the first hypothesis is realized) or Q_0 (if the second hypothesis is realized). It is well known that, if $P_0 \neq Q_0$, the minimal probability of error e_n of their discrimination on the base of n observations converges to zero when n converges to infinity. The question arises what is the rate of this convergence.

It is possible to prove (cf. [6, 7]) that e_n converges to zero for $n \to \infty$ as $[H_{a_0}(P_0, Q_0)]^n$ (asymptotically), where

$$H_{a_0}(P_0, Q_0) = \min_{0 \leqq a \leqq 1} H_a(P_0, Q_0) \tag{19}$$

with

$$H_a(P_0, Q_0) = \int u_0^a \, \mathrm{d}Q_0 \,. \tag{20}$$

Here u_0 is defined analogously as u in the definitions (7) or (15). $H_a(P_0, Q_0)$ is the so-called generalized entropy of order a of P_0 with respect to Q_0 (exceptionally, even for the case $0 \leqq a \leqq 1$, in spite of the fact that then the function $f(u) = u^a$ is not convex but concave). Its value lies between 0 and 1 for a between zero and one, and as a function of a is a convex one, having, thus, always a minimum in the interval $[0, 1]$.

3. SECOND ASPECT OF THE MEASUREMENT PROBLEM: REDUCTION OF THE PARAMETER SPACE X

As said in the introduction, this aspect of the measurement problem is relevant only if the parameter set X, which serves to characterize the decision situation, is not a priori given and, thus, there exists a possibility of selection of this set in order to characterize as economically as possible the decision situation of the original decision task (to which the measurement task is subordinated). Imagine, for instance, that instead of using a measuring device covering the "complete" set of parameters X, there is applied a less sharp one covering only a subset X' of X. The loss of decision quality, thus, resulting may be estimated in a similar way as the loss resulting by a reduction of the observation space (cf. section 2). It is sufficient to take for P and Q the probability laws on $X \times Y$ corresponding to the original and to the reduced or modified case, respectively. If, in particular, the observation channel corresponding to the reduced case may be considered as a rounded off version of the observation channel corresponding to the original case, in the sense that the former, $\{P'_{Y/x'}, x' \in X'\}$, is the conditional expectation of the latter, $\{P_{Y/x}, x \in X\}$, with respect to X'', where $X = X' \times X''$, i.e. if, for every measurable subset E of Y,

$$P'_{Y/x'}(E) = \int_{X''} P_{Y/x}(E) \, \mathrm{d}P_{X''/x'}(x'') \,, \quad (x = (x', x'')) \,, \tag{21}$$

then as Q we shall take an extension $\tilde{P}''$ of P'' (defined by $P_{X'}$ and the rounded off channel above, constructed in a similar way as $\tilde{P}'$ in the preceding section). If the estimates are made in terms of the Shannon's generalized entropy, it again holds that $H(P, \tilde{P}'') = I(P) - I(P'') = I(X, Y) - I(X', Y)$ = loss of Shannon's information due to the reduction of X to X'.

Let us observe that to the rounded off observation channel $\{P'_{X/x'},\ x' \in X'\}$, defined by (21), there corresponds as weight or loss function $w'(x', d)$, $x' \in X'$, $d \in D$, a rounded off version of the original weight function $w(x, d) = w(x', x''; d)$ given by

$$w'(x', d) = \int_{X''} w(x', x''; d)\, \mathrm{d}P_{X''/x'}(x'') \,. \tag{22}$$

It is clear that similar estimation methods may be applied without any difficulty also in the case of a simultaneous reduction of the parameter and observation spaces. However, the situation becomes more complicated if it is desired, together with the parameter space, to reduce parallelly the decision space D. Some results in this direction are contained in the paper [8].

(Received November 27th, 1969.)

REFERENCES

[1] I. Csiszár: Eine informationstheoretische Ungleichung und ihre Anwendung auf den Beweis der Ergodizität von Markoffschen Ketten. Publications of the Mathematical Institute of the Hungarian Academy of Sciences *VIII* (1963), Series A, 85–108.

[2] A. Perez: Information, ε-Sufficiency and Data Reduction problems. Kybernetika *1* (1965), 4, 297–323.

[3] A. Perez: Information Theory Methods in Reducing Complex Decision Problems. Transactions of Fourth Prague Conference on Information Theory (1965), Prague 1967, 55–87.

[4] A. Perez: Information-Theoretic Risk Estimates in Statistical Decision. Kybernetika *3* (1967), 1, 1–21.

[5] A. Perez: Risk Estimates in Terms of Generalized f-Entropies. Proceedings of the Colloquium on Information Theory organized by the Bolyai Mathematical Society, Debrecen, (Hungary) 299–315, 1967.

[6] H. Chernoff: A measure of asymptotic efficiency for tests of a hypothesis based on a sum of observations. Ann. Math. Stat. *23* (1952), 4.

[7] I. Vajda: On the convergence of informaiotn contained in a sequence of observations. Proceedings of the Colloquium on Information Theory, Debrecen, (Hungary), 1967, 489–501.

[8] A. Perez: On the Reducibility of a Set of Statistical Hypotheses. Paper presented at the 37th Session of the International Statistical Institute, London, September 1969.

Informačně-teoretický přístup k problémům redukce v měření

ALBERT PEREZ

V tomto článku je problém měření považován jakožto statistický rozhodovací problém podřízený danému (původnímu) rozhodovacímu problému. Přitom jsou uvažovány dva aspekty měřící úlohy. První aspekt se vztahuje k odhadu hodnot a priori daného systému parametrů, který slouží k charakterizaci rozhodovací situace původního problému. Snahou je redukovat co nejvíce objem dat nutných pro tento odhad, jakož i objem nákladů, spojených s jejich získáním pomocí měření. K usnadnění této úlohy lze použít informačně-teoretické metody zavedené autorem, které spočívají v odhadu ztráty rozhodovací efektivnosti, způsobené redukcí výběrového (tj. pozorovacího) prostoru.

Druhý aspekt úlohy měření, uvažovaný v tomto článku se vztahuje k fundamentální otázce *co* (tj. jaký systém parametrů) je nutno měřit k tomu, aby bylo možno co nejúsporněji charakterizovat rozhodovací situaci, odpovídající původnímu rozhodovacímu problému (ke kterému je úloha měření podřízena). Tato otázka vzniká, pochopitelně, jen když zmíněný systém parametrů není a priori dán tak, že existuje možnost výběru tohoto systému. V článku je poukázáno i na metody usnadňující tento výběr.

Dr. Albert Perez, DrSc., Ústav teorie informace a automatizace ČSAV, Vyšehradská 49, Praha 2.