In *Policy Analysis and Information Systems*, Special Issue on on "Mathematical Foundations," 3(1): 93-109, July 1979.

ENTROPY, INFORMATION AND RATIONAL DECISIONS*

Judea Pearl

School of Engineering and Applied Science
University of California, Los Angeles
California 90024, U.S.A.

ABSTRACT

This paper examines the adequacy of the entropy concept as an economically relevant measure of information in a decision-making framework. Some difficulties of relating the entropy to the traditional notions of information value are highlighted. Entropy-based concepts are shown to retain their economic significance in situations where binary inquiries are answerable at uniform cost, regardless of the payoffs. In these cases, entropy seems to govern the balance between the cost of acquiring useful information and the risks of acting uninformedly.

I. INTRODUCTION

The prospect that the subtle, intangible concept of information might lend itself to measurement in a way similar to the measurement of commodities by money (Renyi [1]) has captured the enthusiasm of many workers in various fields since the inception of information theory two decades ago (Shannon, [02]). Entropic measures of information, uncertainty, disorder and complexity have been applied, with various degrees of acceptance in such diverse fields as psychology (Miller [3]), philosophy (Carnap [4]), music (Goguen [5]), chess (Good [6]), and physics (Brillouin [7]). However, it is in the fields of economics and decision-analysis that controversies concerning the appropriateness of entropybased information measures have raged most intensely.

Beginning with Bagno [8] [9] and pursued by Theil [10] and many others, entropy has been taxed with the job of capturing the essential aspects in the distribution of many

*In memory of Jacob Marschak.

economic indicators. This approach has been surveyed and criticized by Horowitz et al [11]. In decision-analysis, both constrained maximization of prior entropies (Jaynes [12], Tribus [13]) and ranking of measurements by entropy considerations were advocated (Bremerman [14], Chien et al [15], Danskin [16]). This train of thought was criticized by Marschak [17], MacQueen et al [18] and White [19].

The purpose of this paper is to explore in detail the boundaries which delimit the applicability of the entropy concept in the general framework of decision making problems.

There are two major approaches to justifying the use of mathematical measures of qualitative phenomena - an axiomatic and a modeling approach. The axiomatic approach begins with a set of intuitively appealing axioms which any candidate measure should be expected to satisfy, then verifies that the measure it advocates satisfies the axiom and is, to some degree, unique. The modeling approach analyzes a simplified yet typical model of the class of phenomena under study and attempts to show that the measure advocated emerges as a natural feature of the model. Khinchin [20], following the axiomatic approach, proved that the entropy function

$$H(P) \overset{\Delta}{=} - \Sigma\, P(z) \log P(z) \overset{\Delta}{=} H(Z) \qquad (1)$$

is the only function (to within a positive multiplier) which satisfies a set of axioms tailored for coding theory. This set of axioms, in particular, the additivity property of H, was shown to be inappropriate in general decision situations (White [19]).

In this paper we pursue the modeling approach. We will start with the simplest decision-model possible, search for a formal interpretation of the notion of information value, gradually expand the model and examine at what point the entropy can be introduced into the picture as a significant determinant. Section II highlights and analyses the difficulties of tying entropy to pragmatic measures of information in decision problems; Section III recasts the basic structure of communication problems in decision-analytic terms and uncovers the conditions under which entropy-based concepts would retain their economic significance.

Our general conclusion is that the major confusion in this subject stems from conceiving entropy as a measure of the *uncertainty content* of various entities (e.g., signals, probabilities, etc.). A much clearer picture would obtain if, instead, entropy was regarded as a measure of the *effort necessary for removing uncertainty* using a given system of information gathering resources.

II. ENTROPY AND THE VALUE OF INFORMATIVE DECISIONS

We assume that the decision problem at hand can be characterized by the following elements:

$Z = \{z_1, z_2, \ldots z_{N_z}\}$ - a set of N_z mutually exclusive states of nature, one of which is sure to occur.

$A = \{a_1, a_2, \ldots a_{N_a}\}$ - a finite set of N_a mutually exclusive terminal actions from which one must be selected.

$P(Z) = \{P(z_1), P(z_2), \ldots, P(z_{N_z})\}$ set of prior probability weights on the states, where each P(z) measures the likelihood that event ztZ will occur next.

u(z,a) - a payoff matrix on A×Z, where each u(z,a) entry represents the terminal benefit or expected utility associated with the joint occurrence of $a \epsilon A$ and $z \epsilon Z$.

T - a set of tests or information sources where each test $t \epsilon T$ is characterized by a conditional probability matrix $P_t(y|z)$, giving the likelihood that an outcome $y \epsilon Y_t$ will be observed if t is examined and z is about to occur. The elements of T will be interchangeably referred to as tests, information sources, experiments, or inquiries.

C(t) - a cost function, where C(t) measures the cost of acquiring access to the outcome of test t.

The objective of a decision-maker operating in this environment is to design a plan of sequential testing, followed by a terminal action which would maximize the expected payoff minus the cost of testing.

Intuitively we expect the more informative decision maker, i.e., the one who can predict which state will occur with greater accuracy, to secure himself a higher overall benefit. An examination of some difficulties in making this notion of informativeness more precise follows, especially in regard to coupling it with the entropy concept.

1.a. Informativeness, Value of Information and Information Sources

If our state of knowledge concerning an event can be characterized by the set of probabilities P(Z), then it is natural to ask what it is worth to be given these probabilities. The answer clearly would depend on both the payoff, u(z,a), and the detailed content and cost of the test space T. If all these details are available one can, in principle, find an optimal solution to the sequential testing problem above, calculate the overall expected utility given P(Z) and interpret the result as the value of the information represented by P(Z). Yet, aside from the difficulties of finding such a solution, it is very unlikely that the answer would closely match our intuitive notion of *informativeness.* For these reasons simpler and more direct approaches have been attempted to capture the notion of information value.

A good place to start simplifying the decision problem above is to eliminate the test set altogether and assume that a terminal action is to be taken on the basis of P(Z) and u(z,a) alone. The maximum expected utility in this situation is given by:

$$U(P) = \max_{a} \sum_{z} P(z)\, u(z,a) \tag{2}$$

which might be interpreted as the value of the information contained in P(Z). Equation (2) does not mean, of course, that $U(P_1) > U(P_2)$ implies that a person with a probability $P_1(Z)$ is guaranteed a higher expected benefit than a person with probability $P_2(Z)$. It does imply that Person 1 perceives the opportunities offered by the situation at hand to be more attractive than those perceived by Person 2. Therefore, a more appropriate name for the quantity U(P) would be the *Perceived Value of Belief.*

But aside from the fact that U(P) does not have even a remote resemblance to H(P) (being piecewise linear and highly dependent on u(z,a)), it also fails to capture our intuitive notion of *informativeness* or *certainty*. For example, the majority of people would agree that regardless of any payoffs, the probability vector (1,0,0) represents a "more certain" individual than the vector (1/3,1/3,1/3). Similarly, we would agree that the following set of vectors represents, from left to right, an increasing degree of uncertainty:

$$(1,0,0)\ (2/3,1/3,0)\ (1/2,1/2,0)\ (1/2,1/4,1/4)\ (1/3,1/3,1/3) \tag{3}$$

However, this order is not retained by all U(P) measures; given any two vectors P_1 and P_2, one can always find a matrix u(z,a) that would yield either $U(P_1) > U(P_2)$ or (using another matrix) $U(P_1) < U(P_2)$. Hence, we must abandon efforts to link the relation "more certain than" to the value of terminal decisions.

As is well known (DeGroot [21]), the order of the vectors in (3) would be reproduced by any function of P which is both concave and symmetric (e.g., H(Z) or $\sum_z [P(z)]^{1/2}$). One may wonder, however, whether any such function can be shown to represent a significant economic factor within the framework of terminal decision problems with arbitrary payoffs.

An attempt may be made to generate such a function in a slightly different direction, not as a representative of terminal benefits, but as a measure of the decision maker's willingness to purchase additional information instead. It is intuitively suggestive that the "more certain" individual normally would be less willing to pay for additional information; consequently, the question arises if the need for more information is a symmetric concave function of P, regardless of u? The answer is, no. In the face of a terminal decision, the value V(t) of a given test t is given by the expected increase in the posterior utility beyond U(P):

$$\begin{aligned} v_p(t) &= E_y\, U[P(Z|y)] - U(P) \\ &= \sum_y \sum_z P(z)\, P(y|z) \max_a \left[\sum_{z'} P_t(z'|y)\, u(z',a)\right] - U(P) \end{aligned} \tag{4}$$

$V_P(t)$ is again highly dependent on u(z,a) and P(y|z); it is concave in P but not necessarily symmetric. Marschak [22] brings an example where the highest value of a perfect information source occurs at P = (1/2,1/3,1/6), not at P = (1/3,1/3,1/3).

The relation "more informative than," when applied to information sources (probability matrices), is much better understood than when applied to states of knowledge (probability vectors). From Blackwell's theorem [23], we know that if we interpret the former relation as the requirement that $V_P(t_1) > V_P(t_2)$ holds for all u and P, then it forms a partial order on the test set which agrees with our intuition (e.g., t_1 is not "less informative than" t_1 cascaded with t_2).

The assignment of the relation 'more informative' or 'more certain' to states of knowledge is less fortunate. We could not find in the context of terminal decisions an economically meaningful function which could capture our intuition that these relations form payoff-independent partial orders on P. Needless to say, attempts to couple the entropy with the terminal value of information would face similar difficulties.

1.b. Standard Payoff Structures

Any attempt to link entropy-based concepts to the value of information must cope with two difficulties, first it must remove the dependency of U(P) on u(z,a) and, second, it must provide a rationale for the appearance of the logarithm function. A simple method of constructing payoff-independent information measures is to limit the analysis to special cases of symmetric payoff structures which contain a minimal number of parameters and, at the same time, are typical of a large class of problem situations. The most commonly used such structure is the square matrix:

$$u(z,a) = \begin{array}{ll} 1 & a = z \\ 0 & a \neq z \end{array}$$

representing the task of predicting z with unit reward for correct prediction and no reward for all errors. Under this standard we obtain:

$$U(P) = \max_{z} P(z)$$

and

$$V_P(t) = \sum_{y} \max_{z} \ [P(y\,|\,z)\ P(z)] \ - \ \max_{z} P(z)$$

Thus, while this standard structure yields a symmetric concave function for U(P), it is still far from resembling H(P).

The question one may raise at this point is: can we find a standard payoff structure for which U(P) is characterized by the entropy H(P)? Any such attempt must deal with a basic disparity between the two; while H(P) is a strictly concave function of P, U(P) (see equation 2) is piecewise linear in P as long as u(z,a) represents a finite set of actions. Therefore, no finite action payoffs exist which lead to $U(P) = \alpha H(P) + \beta$.

Kelly [24] and more recently, Cover [25] have explored the construction of entropy from a continuous action set. Kelly, motivated by the desire "to take some real-life situation which seems to possess the essential features of a Communcation problem, and to analyze it without the introduction of an arbitrary cost function" has considered a gambler who has to decide in advance what portion r of his capital Q to bet on one of the two outcomes of the bet. In this case, the monetary payoffs, are given by rQ if the gambler wins and -rQ if he loses. At this point, Kelly assumes that the gambler wishes to maximize not the expected monetary gains but the (expected) logarithm of the ratio between the capital after and before the bet, i.e., $u = \log(1+r)$ if the gambler wins and $u = \log(1-r)$ if he loses. Under this assumption Kelly shows that U(P) is proportional to -H(P). A similar device was used by Cover [25]. However, although the continuous action model considered is realistic (many real-life decisions require the allocation of continuous resources) and successfully eliminates the piece-wise linearity of U(P), it still falls short of making a convincing argument for the entropy, as the introduction of the logarithm utility function is rather ad-hoc.

Several workers, discouraged by the difficulties above, use entropy merely as a convenient, heuristic estimate for U(P). In the field of pattern recognition, where the payoff matrix in (5) is a realistic representation of the systems' performances, it has often been the practice to rank pattern features (playing the role of information sources) by entropy-based measures (Andrews [26]). A feature is considered more valuable if it leads to a higher difference between the prior entropy and expected posterior entropy pertaining to the set of classes. This practice was justified by the argument that the true performance measure, i.e., the probability of mis-classification $P_e = 1\text{-}U(P)$, can be bounded by entropy-based measures. Specifically, P_e can be upper bounded by (Ben Bassat et al [27]):

$$P_e \leqslant 1/2\ H(P)$$

and lower bounded by:

$$H(P) \leqslant (1\text{-}P_e)\log(1\text{-}P_e) - P_e \log \frac{P_e}{N_Z - 1}$$

(unless stated otherwise, logarithms are to the base 2). Thus, in situations where these bounds are tight and under a probability of error performance criterion, entropy could provide realistic estimates to the value of information and information sources.

2.3 Uncertain Payoffs

The piecewise linearity of U(P) can also be "smoothed out" within a finite action model if the payoffs u(z,a) are considered random variables whose exact values become known to tne decision maker just before he is about to select a terminal action. At the phase of data gathering, however, only a distribution over a possible range of payoffs is available to him. The expected utility, in this case, is obtained by taking the expectation of expression (2) with respect to the given payoff distribution.

This model is not unrealistic. There is hardly a decision problem where the exact values of the payoffs are known at the time that information is collected. The weather predictor, for example, has only an aggregate knowledge of the stakes which his clients, the radio listeners, have in future weather conditions. A college student, gathering knowledge in preparation for his professional career, has only a vague notion of the nature of the circumstances where his knowledge will stand a critical trial. Yet both he and the weather forecaster must be able to assess the value of the information which they already possess in order to select information sources judicously.

Pearl [28] has analyzed simple models with uncertain payoffs for the purpose of finding an economic basis for probability scoring rules. Imagine a decision situation with two states $z \epsilon \{0,1\}$, and two actions $a \epsilon \{0,1\}$, and let the payoff matrix be:

$$u(a,z) = \begin{cases} 0 & a = z \\ x_1 & z = 1,\ a = 0 \\ x_2 & z = 0,\ a = 1 \end{cases}$$

Now assume that x_1 and x_2 are two independent random variables, both distributed according to a probability density function f(x) It can be shown that if f(x) is given by the Cauchy density

$$f(x) = \begin{cases} \dfrac{2/\pi}{1+x^2} & x \leqslant 0 \\ 0 & x > 0 \end{cases}$$

the expected utility becomes:

$$U(P) = \frac{2}{\pi} H(P)$$

Other densities would, of course, given rise to different value functions. For example, an exponential density:

$$f(x) = \begin{cases} \lambda e^{-\lambda x} & x \geqslant 0, \\ 0 & x < 0, \end{cases}$$

yields a quadratic value function:

$$U(P) = \frac{1}{\lambda} [1-p(1-p)]$$

where $p = P(1)$.

Thus, the use of the entropy as a measure of information values would be justified in situations where the decision-maker perceives Cauchy-like payoff distributions in future decisions. It gives Shannon's entropy some realistic economical interpretation outside the field of communication.

This result, however, was limited to the case of a 2 x 2 payoff matrix. Attempts to extend it to arbitrary matrices have met with substantial difficulties; ad-hoc assumptions regarding the statistical dependence between the various payoff variables must be made in order for U(P) to assume an entropy-like form.

III. INFORMATION THEORY AND THE ECONOMICS OF INFORMATION GATHERING POLICIES

The difficulties thus encountered in justifying the use of entropy-based measure in the analysis of simple decision situations behoove us to take a fresh look at the communication problems which gave rise to information theory in the late 1940's. After all, the problem faced by the communication engineer can also be regarded as an exercise in rational decision making and so, by carefully examining the emergence of the entropy function in communication related problems one may hope to uncover the more general conditions under which entropy would serve as a natural measure of the economical value of information.

Many decision analysts have discarded information theory on the ground that it is preoccupied with the overly narrow objective of transmitting information symbols with maximum fidelity, with no regard to the meaning or purpose carried by the symbols. With this narrow view, the essential ingredients of decision-making, i.e., selecting actions and buying information, seem to disappear. We will carry out our analyses in a wider context, preserving all the features making up a general decision problem and show that some information-theoretic concepts still retain their significance when the test space is suitably constrained.

3.1 Standardization of the Test Space T

The essential element necessary to bring a general decision problem (Z, A, P, u, T, C) in line with the conditions prevailing in common communication problems is the uniform cost structure assumed by the test space T and its associated cost function C(t). First assume that T contains the entire set of binary tests on Z. In other words, each test is characterized by a two column stochastic matrix: $T = \{P(y|z) | y \in \{0,1\}\}$. Second, assume that the usage of each test consumes a fixed cost of c units, i.e., C(t) = c. We will denote this test space by the symbol T_b. The problem now faced by the decision-maker is to devise a strategy for selecting tests and actions so as to maximize the expected value of the terminal payoffs minus the cost of testing.

The assumption limiting T_b to contain binary tests only is not too severe, and serves primarily as a standard to distinguish complex questions from simple 'yes' or 'no' type questions. The answer to an arbitrary question could be regarded as a concatenation of binary answers.

The other assumption though, that all binary queries are answerable at equal cost, is unique to information transmission problems. Equal cost represents the fact that the major cost of information transmission is the message length and that each bit in the message occupies the same time duration regardless of what it represents. In many common economical problems the cost of queries is certainly not uniform, as some queries simply require much higher expenditure of resources than others. For example, to determine whether the majority of voters favor candidate A is much more costly than to determine whether a given individual favors candidate A. In some other applications, as when the main cost of obtaining information lies in the need to store the answers in a computer memory, the uniform cost structure provides a reasonable model for the information acquisition system.

3.2 Entropy as the Perceived Cost of Removing Uncertainty

In order to see how the entropy enters the economical environment defined above, let us begin by making additional assumptions regarding the payoff matrix u(z,a) (these assumptions will be relaxed later on in the paper). Assume that for every state z there exists a unique best action a_z, i.e., $u(z, a_z) = \min_{a \in A} u(z, a)$, and $a_{z_1} = a_{z_2} \Rightarrow z_1 = z_2$. Assume further that the losses associated with selecting a suboptimal action are much higher than the information cost unit c, i.e., for all z and $a \neq a_z$ $u(z, a_z) - u(z, a) \gg c$. Under these high stakes conditions the decision-maker should not risk acting suboptimally but rather purchase all the information needed for complete state identification. Any residual uncertainty regarding z would result in losses greater than the cost of removing that uncertainty and, so, the decision-makers objective must be to plan an information gathering strategy that identifies z at a minimum expected cost.

Any information gathering strategy consists of a procedure for selecting an information source from T_b, testing its outcome and on that basis, deciding on the next test until the sequence of outcomes from the selected tests is sufficient to determine z unequivocally. Such a sequential testing procedure can be represented by a binary tree where the terminal nodes represent the identified states, the non-terminal nodes represent the various tests chosen, and the arcs represent the tests outcomes. Thus, the task of designing an optimal information gathering policy reduces to that of configuring a tree with minimum expected path length (the probability of traversing a given path being given by the probability of the state terminating that path). This minimization problem was solved by Huffman and the resultant test procedure became known as the Huffman coding scheme [29]. Let the optimal expected path length obtained by Huffman's procedure be designated by L_h. One of the earliest results of information theory states that L_h satisfies the inequality:

$$H(Z) \leqslant L_h < H(Z) + 1$$

Thus, the entropy H(Z) provides a good approximation to the optimal mean path length.

Correspondingly, c times the entropy can be taken as a good approximation to the expected minimum cost of buying the information required for optimal decision-making. If U° is the expected utility associated with perfect knowledge:

$$U^{\circ} = \Sigma_z P(z) \max_a u(z,a) ,$$

then the overall expected utility U associated with the decision problem at hand, including the cost of information, is bounded by:

$$U^{\circ} - c[H(Z) - 1] < U \leqslant U^{\circ} - cH(Z)$$

Note that unlike the attempts made in section II, the entropy no longer measures the terminal economical losses caused by uncertainty but rather the cost of eliminating that uncertainty (given a uniform cost test space).[1] Thus, the economical significance of this measure stems from the fact that a decision-maker with a high entropic uncertainty is expected to spend more money on information gathering devices than the more knowledgeable (low entropy) decision maker. If the probabilities P(Z) are subjective in nature, there is of course, no guarantee that Z can in fact be identified at an estimated mean cost of cH(Z); in this case cH(Z) represents the minimum mean information cost as perceived by the decision-maker.

Note, also, that we are now in possession of a meaningful interpretation of the relation "more certain than", which caused insurmountable difficulties in Section 2.1. Uncertainty can be captured not only by the entropy function, but also by the minimum expected text-costs under high-stakes conditions and any symmetric (not necessarily uniform) cost function C(t) (i.e., $C(t_1) = C(t_2)$ if t_1 differs from t_2 by a permutation of states or outcomes). The minimum expected test-costs under these conditions would be a concave symmetric function of P and could, therefore, reproduce the partial order expressed by the term "more uncertain than."

It is important to mention that the bounds on L_h can be significantly tightened in case the decision problem involves not a single event but a sequence of independent and identically distributed events z(1), z(2) ... z(N), $z(i) \epsilon Z$. If the test space is correspondingly augmented to include all binary tests on N-sequences (at equal cost), and the problem circumstance permits withholding action until such tests are performed, then the bounds above can be tightened to yield:

$$c\,H(Z) \leqslant \overline{C} < c[H(Z) + \frac{1}{N}]$$

where $\overline{C}$ is the expected cost per event. Information theory is primarily interested in coding long sequences of data, and therefore, the entropy becomes an almost exact measure of the mean code length. In fact, most information-theoretical results are valid only asymptotically, for $N \rightarrow \infty$. In most economical decision problems, on the other hand, observations must be made separately on each individual event and actions must be selected before the next event takes place. In such cases, the entropy only provides an estimate for the mean number of tests, the quality of which improves with increasing N_z.

In the remaining parts of this paper, we shall treat this approximation as an equality with the understanding that the results may vary within the ranges defined by the bounds above.

3.3 Channel Capacity – The Cost Reduction Potential of an Information Source

Suppose that the decision-maker, still having access to a uniform cost binary test space T_b, is offered an extra information source t, not necessarily binary. How much should he be willing to pay for the option of using t?

Clearly, if $t \epsilon T_b$, then he should not purchase it for a price higher than c, because he can always find an equivalent source within T_b for a cost c. However, t may be worth less than c. If the outcomes y of t are only loosely related to the states z than t may be replaced by a more effective member of T_b, at a price c.

In general, the worth of any information source should now be judged not by its potential for improving decisions, but rather by that for reducing the necessary cost of identifying z, i.e., reducing H(Z). If the outcome y is observed, then the expected residual cost of identifying z is cH(Z|y), and the overall expected residual cost after consulting t (denoted by cH(Z|Y)) would be cE_y H(Z|y). Therefore, the value of t is given by the reduction in the expected cost:

$$V_p(t) = cI(Z,Y) \triangleq c[H(Z) - H(Z|Y)]$$

$$= c\, \Sigma_z\, \Sigma_y\, P(z,y) \log \frac{P(z|y)}{P(z)}$$

$$= c\, \Sigma_z\, \Sigma_y\, P(z)\ P(y|z) \log \frac{P(y|z)}{\Sigma_z\, P(y|z)}$$

Note that I(Z,Y) (known as Shannon's mutual information) depends on both the matrix P(y|z) and the prior probabilities P(z), implying that two different users may rank information sources in different order. I(Z,Y) can also be shown to be a strictly concave function of P(z) and a strictly convex function of P(y/z). The former implies that the perceived value of any information source averaged over a group of decision-makers is lower than the value of the source based on the group's averaged priors. The latter implies that if the corresponding outcomes of several information sources are mixed by a random device, then the worth of the mixed source is less than the average worth of the individual sources.

We may now wish to ask what is the highest price that a user may possibly be willing to pay for a given information source? The answer is obtained by maximizing I(Z,Y) over all prior probabilities, and the resulting quantity

$$\mathrm{Cap}(t) = \max_{P(Z)} I(Z,Y)$$

was termed the channel *capacity*, It is clearly an intrinsic property of the information source or the channel P(y|z). In communication theory, the capacity represents the maximum information transfer capability of a given (noisy) channel, and is therefore of utmost importance. This results from the communication engineer's ability (having the option of coding the input messages before they enter a transmission channel) to adjust the probabilities P(z) to make I(Z,Y) achieve its highest value Cap(t). Such an option is not available in the decision-making environment we are considering; the prior P(z) is determined by the decision-maker's previous experience and background knowledge, and cannot be tampered with.

It is interesting to note, though, that the highest price for t would not always be offered by the most ignorant (i.e., $P(z) = 1/N_z$) user, since the expression for I(Z,Y) is not invariant under state permutation. If, however, the outcomes (columns) of the matrix P(Y|Z) can be partitioned into subsets in such a way that in each submatrix each row is a permutation of each other row and each column is a permutation of each other column, then I(Z,Y) achieves its highest value under equally distributed state probability (complete ignorance).

3.4 Tradeoffs Between Performance Degradation and Information Costs

So far, we have assumed that the unit information cost c is much lower than the terminal stakes involved, and that the condition of high stakes necessitates the removal of all uncertainties concerning state identity. We now wish to relax these two assumptions. It quite often happens that the cost of acquiring information is of the same order of magnitude as the terminal payoffs, and moreover, the choice of the optimal action may not require a complete state identification, as two or more states may require the same optimal action.

In the more general case, one may be willing to settle for a certain degradation in terminal payoffs if that would cut substantially the cost of information: the problem that then arises is finding the proper balance between the two. Mathematically, we can formulate the problem as follows: consider a decision problem with an arbitrary u(z,a), $T = T_b$, C(t) = c, and let U° be the maximum expected utility achievable with complete knowledge of z. What is the minimum expected cost of information which would guarantee an expected payoff of at least U°-D, where D is the maximally tolerable degradation in terminal performance?

The formulation above no longer ignores the economical significance of information symbols but rather brings into focus the tight coupling between the information purchased and its consequences, the payoff matrix u(z,a). In this sense, we now possess a more representative model of a typical decision-making problem. The solution to this problem was developed by information theorists under the topic of *rate-distortion theory.*

Rate distortion theory was introduced by Shannon [30] as early as 1959, and although it has become a major focus of research in more recent years (Berger [31]) it remains largely unknown outside the field of information-theory. Some applications of rate-distortion theory to pattern-recognition and computational complexity can be found in references [32], [33], and [34].

In order to see the role played by the entropy in the general problem of arbitrary payoffs, assume that the decision-maker adopts an arbitrary information gathering policy S for selecting and inspecting information sources from T_b and that S costs him an expected cost C(S), enabling him to produce an expected utility of at least $U^\circ - D$. The action $a \epsilon A$ finally taken by the decision-maker after observing the outcomes of the sources chosen by S is a random variable whose dependence on the state z can be characterized by a conditional probability matrix $P_S(a|z)$, satisfying:

$$U^\circ - \sum_z \sum_a P(z)\, P_S(a|z)\, u(z,a) \leqslant D$$

We will call any conditional probability matrix satisfying the inequality above D- admissible and designate the set of all such matrices by $\underline{P}_D$.

Imagine now that a hypothetical forecaster attempts to predict the state z by examining the final actions taken by the decision-maker. The forecaster would be able to determine the exact state if he purchases additional information at a cost of:

$$cH_S(Z|A) = -c \sum_Z \sum_a P(z)\, P_S(z|a) \log P(z|a)$$

thus enabling the decision-maker and the forecaster as a team to identify z at a total cost of $C(S) + cH_S(Z|A)$. However, since the complete identification of z must cost at least cH(Z) we have:

$$C(S) \geqslant c\, [H(Z) - H_S(Z|A)] = cI_S(Z,A)$$

Taking the minimum of C(S) over all policies which yield D-admissible performances gives the minimum cost C(D) of the information needed to achieve such a performance. Thus:

$$C(D) \overset{\Delta}{=} \min_{S:\, P_S(a|z) \in \underline{P}_D} C(S) \geqslant c \min_{P(a|z) \in \underline{P}_D} I(Z,A)$$

The function defined by the minimization of I(Z,A) is called the rated-distortion function R(D) which, in communication theory, gives the minimum expected code length (in bits per symbol) required to produce a mean *distortion* not exceeding D. Shannon proved also that codes do exist which achieve a mean distortion D with a mean code length arbitrarily closed to R(D), provided that very large blocks of data are coded simultaneously. In our application though, where each observation sequence serves a separate decision, only the inequality

$$C(D) \geqslant cR(D)$$

carries economical significance.

For an arbitrary probability vector P(Z) and an arbitrary payoff matrix u(z,a) the exact calculation of R(D) is not a trivial matter. Analytic expressions for R(D) are available in only a few special cases (e.g., $P(z) = 1/N_z$ and $u(z,a) = \delta(z,a)$ yields $R(D) = \log N_z + D \log D + (1-D) \log (1-D) - D \log (N_z - 1)$. However, good bounds to R(D) can be readily obtained using the following procedure (Berger [31]): let M(s) be any function of s satisfying:

$$M(s) \geqslant \max_{a} \sum_{I} 2^{sd(z,a)}$$

where

$$d(z,a) = U^{\circ} - u(z,a)$$

The rate-distortion function is lower bounded by $R_L(D)$ where:

$$R_L(D) = H(Z) + sD(s) - \log M(s) \qquad -\infty < S \leqslant 0 \quad s \quad (1962).$$

and

$$D(s) = \frac{d}{ds} \log M(s) \qquad -\infty < s \leqslant 0$$

The last two equations provide a parametric representation for a curve $R_L(D)$ lying on or below R(D), with s as the variable parameter.

From an economic viewpoint, the merit of a given information gathering strategy S is judged by the expected terminal utility U(S) minus the expected information cost C(S).[2] Using the lower bound $R_L(D)$ above, we can find an absolute upper limit on U(S) − C(S) for all information gathering strategies:

$$\max_{S} [U(S) - C(S)] \leqslant U^{\circ} - \min_{D} [cR_L(D) + D]$$
$$\leqslant U^{\circ} - c\,H(Z) + c \log \max_{a} \sum_{z} 2^{-d(z,a)/c}$$

$U^{\circ} - cH(Z)$ is the maximum expected utility when the user resolves to purchase all the information required for state identification. The logarithmic term represents a potentially extra gain due to the user's willingness to make suboptimal decisions in order to save information costs. It reduces to zero under high-stakes conditions $c \ll d(z,a)$.

Thus, by a simple inspection of the payoff matrix and prior to any optimization procedures, one can determine a ceiling on the overall expected utility of a decision situation, including both terminal payoffs and information costs.

IV. CONCLUSION

The major cause of misunderstanding about the meaning of the entropy measure and its indiscriminate usage stem from the misconception that entropy was developed to measure the benefits of information. In fact, even in communication theory proper, where more information invariably implies more benefits, the entropic-measure of uncertainty has never been meant to measure the evils of uncertainty but rather the cost of its removal. Likewise, it is not the assumption of equal penalty for all errors which keeps entropy-based concepts from breaking the confines of communication problems and becoming more universally applicable, but rather the assumption of equal cost for all (binary) tests.

The paper shows that in situations where this latter assumption represents a reasonable approximation, entropy plays a significant role under all payoff structures. Conversely, when the assumption of uniform test-costs is not valid, only loose connections can be established between entropic measures and the pragmatic value of information.

ACKNOWLEDGEMENT

This work is greatly indebted to the ever inspiring memory of the late Jacob Marschak, who first introduced me to this subject. I thankfully acknowledge the fruitful discussions I had with Norman Dalkey.

Portions of this work were supported by the National Science Foundation grants No. MCS-78-07468 and MCS-75-18739, and the Office of Naval Research grant No. N00014-78-C-0372.

REFERENCES

1. Renyi, A., Statistics based on information theory. Presented at the European Meeting of Statisticians, 1966. Quoted in [17].
2. Shannon, C.E., The mathematical theory of communication. *Bell System Technical Journal* (two papers, reproduced in the book of same title, by Shannon and Weaver, University of Illinois Press, 1949).
3. Miller, G.A., *The Psychology of Communication.* New York, Basic Books. 1967.
4. Carnap, R., *The Continuum of Inductive Methods.* Chicago: University of University of Chicago Press (1952).
5. Goguen, J.A., Complexity of Hierarchically Organized Systems, *Int. J. General Systems,* Vol. 3, No. 4. pp. 233-251 (1977).
6. Good, I.J., Dynamic Probability, Computer Chess and the Measurement of Knowledge, in Michie, D., and Elcock, E.W. (eds.). *Machine Representation of Knowledge;* Dordrecht: D. Reidel (1976).
7. Brillouin, L., *Science and Information Theory,* 2nd ed., London: Academic Press, Inc. 1962.

8. Bagno, S., The Communication Theory Model and Economics, IRE National Convention Records, Part IV, (1955).
9. Bagno, S., *The Angel and the Wheat (Communication Theory and Economics)* New York: Jonah Publishing Co. (1963).
10. Theil, H., *Economics and Information Theory.* New York, Rand McNally, 1967.
11. Horowitz, A.R., and Horowitz, I., The Real and Illusory Virtues of Entropy-Based Measures for Business and Economic Analysis, *Decision Sciences 7,* pp. 121-136 (1976).
12. Jaynes, E.T., "Prior Probabilities," *IEEE Trans. Syst. Sci. Cyb.* SSC-4, 227-241. (1968).
13. Tribus, M., *Rational Descriptions, Decisions and Design* (Pergamon Press, New York). 1968.
14. Bremermann, H.J., "Pattern Recognition, Functionals, and Entropy," *IEEE Trans. on Bio-Medical Engineering,* pp. 201-207 (July 1968).
15. Chien, Y.T. and K.S. Fu, "Selection and Ordering of Feature Observations in a pattern Recognition System," *Information and Control,* vol. 12, pp. 394-415 (1968).
16. Danskin, J., Reconnaissance I and II, *Opl Res. Q.* 10, 285, (1962).
17. Marschak, J., "Economics of Information Systems." In M.D. Intriligator, ed., *Frontiers of Quantitative Economics.* Amsterdam: North-Holland Publishing Company, pp. 32-107, (1971).
18. MacQueen, J. and Marschak, J., Partial Knowledge, Entropy, and Estimation, *Proc. Nat. Acad. Sci.* USA, Vol. 72, No. 10, pp. 3819-3824 (Oct. 1975).
19. White, D.J., Entropy and Decision. *Operational Research Quarterly,* 26(I), 15-23, (1975).
20. Khinchin, A.I., *Mathematical Foundations of Information Theory.* Dover Publications, New York. (1957).
21. DeGroot, M.H., Uncertainty, information, and sequential experiments. *Annals of Mathematical Statistics,* 33, 404-419, (1962)
22. Marschak, J., Remarks on the Economics of Information. *Contributions to Scientific Management.* Western Data Processing Center, University of California, Los Angeles, CA, 79-98, 1960.
23. Blackwell, D. and Girshick, A., *Theory of Games and Statistical Decisions.* New York, Wiley, 1954.
24. Kelly, J., "A New Interpretation of Information Rate," *Bell Syst. Tech. J.,* 35, 917-926, 1956.
25. Cover, T., "Universal Gambling Schemes and the Complexity Measures of Kolmogorov and Chaitin," *Technical Report* No. 12, Department of Statistics, Stanford University, Stanford, CA, Oct. 14, 1974.
26. Andrews, H.C., *Introduction to Mathematical Techniques in Pattern Recognition,* New York: John Wiley & Sons, Inc. (1972).
27. Ben-Bassat, M., and Raviv, J., Renyi's Entropy and the Probability of Error, *IEEE Trans. on Inf. Theory,* Vol. IT-24, No. 3, pp. 324-331, (May 1978).
28. Pearl, J., An Economic Basis for Certain Methods of Evaluating Probabilistic Forecasts, *Int. J. Man-Machine Studies, 10,* pp. 175-183, (1978).
29. Gallager, R.G., *Information Theory and Reliable Communication.* New York: Wiley, 1968.
30. Shannon, C.E., "Coding Theorem for a Discrete Source with Fidelity Criterion," in *IRE National Convention Record,* part 4, pp. 142-163, 1959.
31. Berger, T., *Rate Distortion Theory.* Englewood Cliffs, New Jersey: Prentice-Hall, 1971.
32. Pearl, J., An Application of Rate-Distortion Theory to Pattern Recognition and Classification, *Pattern Recognition,* Vol. 8, pp. 11-22 (1976).
33. Pearl, J., "Theoretical Bounds on the Complexity of Inexact Computations," *IEEE Trans. Inform. Theory,* Vol. IT-22, pp. 580-586, Sept. 1976.
34. Pearl, J., On Summarizing Data Using Probabilistic Assertions, *IEEE Transactions on Information Theory* IT-23, No. 4, 459-465, July 1977.

FOOTNOTES

[1] This distinction is highly related to what Marschak [22] called demand and supply values of information.

[2] We assume here that testing costs do not fluctuate much beyond the linear range of the utility function.