TECHNICAL REPORT
Reprint 13
October 1977

Int. J. Man-Machine Studies (1978) **10**, 000–000

An economic basis for certain methods of evaluating probabilistic forecasts[†]

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

(*Received* 18 *October* 1977)

This paper deals with the question of selecting an appropriate measure of compatibility (also called Scoring Rule) between probabilistic models and empirical data. It is natural to require that if the model predicts the occurrence of an observed event with probability $p < 1$, then the compatibility measure should reflect the actual economical damage caused to the user who acts as though the event (which is about to happen) has only a $p < 1$ chance of occurring.

The paper establishes relations between the compatibility measure and the user's distribution of future pay-offs and shows that each of the commonly used measures (e.g. logarithmic, spherical, quadratic) represents a natural payoff emanating from an economic environment of a specific character. Using these relations the problem of selecting an appropriate measure of compatibility reduces to that of characterizing the economic impact of the forecast at hand.

1. Introduction: natural vs. universal scoring rules

The task of modeling requires a choice of an appropriate measure of compatibility to evaluate the degree at which each candidate model represents a faithful summary of the empirical data. For deterministic models the mean-square-error criterion is often employed to evaluate the models compatibility with the data. For stochastic (or probabilistic) models a new criterion should be employed since a probabilistic model determines not a specific outcome but an elastic constraint (distribution) over possible outcomes. More precisely, if a model predicts that an event ε will occur with probability $p(\varepsilon)$ and event ε is in fact observed, we wish to define an appropriate measure on the pair $(P(\varepsilon), \varepsilon)$, that would represent the quality of the model's prediction.

An identical problem also surfaces in decision analysis where the need arises to evaluate the quality of human predictors. For example, the effectiveness of probabilistic weather forcasters (e.g. "80% chance of rain") is clearly a function of the probabilistic report and the eventuality which actually takes place. In this context the compatibility measure became known as "scoring-rule" since it could be used to "score" and remunerate the human expert in accordance with the success of his forecast.

Following McCarthy (1956) we consider a situation in which a client pays a forecaster for predictions of a future event according to the following rules.

(i) The forecaster gives the client probabilities $p_i, \ldots . p_n$ for the events where $\Sigma p_i = 1$.
(ii) The client takes action on the basis of these probabilities and one of the possible events occurs.

[†]Supported in part by National Science Foundation Grant No. GJ-42732.

0020–7373/78/0000–0000$00.00/0

(ii) If the ith event~~s~~ occurs, the client pays the forecaster $R_i(p_1, \ldots, p_n)$, which is abbreviated $R_i(\mathrm{p})$.

(iv) Neither the forecaster nor the client can influence the predicted event, although the forecaster can make experiments to help predict it, and the client gets an amount which depends on both the action he takes and on the events which occur.

(v) It is assumed that the forecaster and the client both wish to maximize the expected value of their incomes.

The pay-off $R_i(\mathrm{p})$ has become known in the literature (Brown, 1970; Winkler, 1971; Shuford, Albert & Massengill, 1966) as a *Probability Scoring Rule* to be selected in a way that would induce the forecaster to follow a certain mode of behaviour. A scoring rule is said to be *admissible* (other names are *proper* or *reproducing*) if it tends to "keep the forecaster honest". That is, assuming that the forecaster perceives the probabilities of the possible events to be $\pi = (\pi_1, \ldots, \pi_n)$ then, regardless of the value of π, the forecaster expectation $R(\pi,\mathrm{p}) = \Sigma\pi_i R_i(\mathrm{p})$ is maximized if he reports $\mathrm{p} = \pi$. A scoring rule will be called *strictly admissible* if the report $\mathrm{p} = \pi$ is the only one which achieves the maximal expectation of $\bar{R}(\pi,\pi) = \Sigma\pi_i\, R_i(\pi)$.

There are several ways of generating admissible scoring rules. For example (McCarthy, 1956), every differentiable strictly convex function $S(\mathrm{p})$, which is homogeneous of the first degree can generate a strictly admissible scoring rule via $R_i(\mathrm{p}) = (\partial/\partial p_i)S(\mathrm{p})$. The expectation of an honest forecaster is then $\bar{R}(\pi,\pi) \triangleq S(\pi)$.

As an alternative, a scoring rule can be chosen which passes on to the forecaster some of the economical consequences of his report, as viewed by the client. For example, suppose that on the basis of the forecaster's prediction the client chooses the jth of the actions open to him and that his pay-off if the ith event occurs is a_{ij}. His expectation will be $g(\mathrm{p}) = \max_j \Sigma_i\, a_{ij}\, p_i$ if j is chosen optimally. If, eventually, event i occurs, the client's pay-off would be $a_{ij'(\mathrm{p})}$ where $j'(p)$ is the optimal action taken on the basis of the forecast (p). It is natural that the client would wish the forecaster to share his risks and so institute a scoring rule $R_i(p) = \alpha a_{ij'(\mathrm{p})}$ where α is some scaling factor. This type of score was called *naturally imputed* scoring rule by Raiffa (1964) who also showed that every such scoring rule is admissible.

In contrast to the natural rules, probabilistic forecasts are often evaluated by general standards, independent on the details of the client's decision problem. The three most popular scoring rules which seems to dominate both the practice of probabilistic forecasting (e.g. weather predictions, intelligen~~t~~ce information) and experimental research on human information processing are the following.

1. Logarithmic—$R_i(\mathrm{p}) = \log(p_i)$

2. Quadratic — $R_i(\mathrm{p}) = 1 + 2p_i - \sum_{i=1}^{n} p_i^2$

3. Spherical — $R_i(\mathrm{p}) = p_i/\left(\sum_{i=1}^{n} p_i^2\right)^{1/2}$

We call these rules *universal* to stress their apparent independence on any particular decision set-up or pay-off matrix a_{ij}.

The popularity of these universal rules stem primarily from their simplicity and secondary mathematical properties. For instance, the logarithmic scoring rule is the only one (for $n>2$) where the pay-off depends only on the probability assigned to the event which actually occurred (Shuford *et al.*, 1956). The quadratic score, on the other hand, is the only one with the property that the loss for reporting p when π applies is the same as that for reporting π when p applies (Savage, 1971).

An important consideration in the selection of a scoring rule is the way it influences the forecaster's allocation of resources in his effort to get information. Every scoring rule imposes its own preferential order on the inquiries or experiments which the forecasters might employ in his effort to sharpen his predictions. The logarithmic scoring rule, for example, would cause the forecaster to rank inquiries in order of Shannon's measure of channels Mutual Information (Pearl, 1974). While it is natural to require that the forecaster's ranking of inquiries matches that of his client it can be shown that such matching is only possible under a natural scoring rule or a rule which differs from the latter by at most a fixed amount (independent on p). Thus, no universal scoring rule exists which matches the forecaster's worth of information with that of all conceivable clients.

Although natural scoring rules provide the most appropriate measure of what it is worth to be given the probabilities p, and the most direct method of conveying to the forecaster the economical impact of his predictions, they are subject to several shortcomings which limit their application in both practical and laboratory environments. The most severe limitation is that probabilistic assessments are often needed long before the client's pay-off matrix a_{ij} becomes known. In many cases the client's perception of the problem structure becomes clear only after he obtains an assessment of the likelihood of future events. In some cases it is even desirable to conceal the pay-off matrix from the forecaster in order not to "contaminate" his likelihood assessment process with ulterior interest he might have in his client's actions.

A second limitation to using natural scores lies in the fact that unless the pay-off matrix a_{ij} contains an infinite set of actions, the resulting scores are not strictly admissible. Any finite pay-off matrix would partition the probability space p into equivalence regions since all forecast reports p which result in the same optimal action $j'(\text{p})$ would yield the same score to the forecaster. Thus, the forecaster finds no incentive to make his report p match π as accurately as possible as long as the two belong to the same equivalence region. Of course, in the particular economic set up described by a_{ij} the added accuracy is indeed superfluous, however, the forecast p could no longer be trusted as soon as a_{ij} undergoes a change.

Universal scoring rules are free from these weaknesses by virtue of their being strictly admissible. Thus, a forecast report obtained under a universal score can be used with trust by many clients confronted with widely different decision situations. From a conceptual viewpoint too, it is desirable to make $\bar{R}(\pi,\pi)$, which measures the value of the information contained in π, a strictly convex function of π. Only a strictly convex measure of information has the property that its value increases whenever results of a relevant[†] experiment become known, thus matching our intuitive notion that it is always a good idea to look at the outcome of an experiment if it is free.

The purpose of this paper is to demonstrate that strictly admissible universal scoring rules are not completely void of economical rationale and that the latter is not less "natural" than the natural scoring rules of the foregoing discussion. We shall show that each of the commonly used scoring rules (e.g. logarithmic, spherical, quadratic) represents

[†] A relevant experiment is one whose outcomes are not entirely independent on the events to be predicted.

a natural pay-off emanating from a simple decision situation with a unique economical character.

2. Analysis

The basic gambling model we construed consists of the same forecaster-client relations as in the introduction with one added feature: the client payoffs (a_{ij}) are not known at the time the prediction is given. Rather than dealing with fixed pay-offs a_{ij} we now have random variables x_{ij} whose distribution represents the likelihood that a pay-off level x_{ij} will eventually be realized from each event-action combination. We assume that at the time the forecast is given only the distribution of future pay-offs is known to the client; at a later time, when the actual values of the pay-off matrix become known, the client may utilize the forecast given to him earlier, and choose an action which maximizes his expected return.

Uncertainties concerning future payoffs were introduced by Murphy (1966) in the context of meteorological forecasting. There, the ratio between the cost of protecting against an adverse weather condition and the loss anticipated from such a condition (without protection) was regarded as a random variable. The aspect of uncertain payoffs, however, seems to prevail almost all situations involving probabilistic coding of partial knowledge. 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 are required to generate probabilistic estimates; the forecaster in the phrasing of his statements and the student in the way he structures his knowledge.

To facilitate an analytic treatment we limit the model to a simple situation with only two events and two actions. Let the client pay-off matrix be represented by the following table:

		Actions	
Probability	Events	a_1	a_0
$p_1 = p$	E_1	0	y
$p_2 = 1-p$	E_0	x	0

The random variables x and y represent the pay-offs connected with acting a_1 when E_0 occurs and acting a_0 when E_1 occurs, respectively. x and y can both assume positive and negative values but are unknown when the forecast is obtained. Upon receiving the forecast p = $(p, 1-p)$ the client commits to a linear decision rule $d_p(x,y)$ concerning future pay-offs which maximizes his expected return:

$$d_p(x,y) = \begin{cases} \text{act } a_1 & \text{if } x(1-p) \geq y\,p, \\ \text{act } a_0 & \text{if } x(1-p) \leq y\,p. \end{cases} \tag{1}$$

Note that for certain combinations of (x,y) (e.g. $x > 0$, $y < 0$) the choice of best action could be determined without reference to p. The forecast p draws its worth from those occasions only where it serves to ~~determine~~ select an appropriate action ~~resolutions~~.

For a fixed pay-off pair (x,y), the decision rule $d_{\mathrm{p}}(x,y)$ results in the following returns to the client. If E_1 occurs he receives:

$$r_1[d_p(x,y)] = \begin{cases} 0 & d_p(x,y) = a_1, \\ y & d_p(x,y) = \underset{c}{a_{\cancel{1}}}, \end{cases} \tag{2}$$

while if E_0 occurs he receives:

$$r_0[d_{\mathrm{p}}(x,y)] = \begin{cases} x & d_p(x,y) = a_1, \\ 0 & d_p(x,y) = a_0. \end{cases} \tag{3}$$

The expected returns, $R_1(p)$ and $R_0(p)$, depends on the joint distribution of x and y. Choosing the simplest model that leads to nontrivial results we assume that the economical environment underlying the emergence of the gambles (x,y) is symmetric with respect to E_1 and E_0, and that x and y are independent, identically distributed, continuous random variables with joint density $f(x,y) = f(x)\,f(y)$. In the case of weather forecasting, for example, $f(x,y)$ would represent that fraction of the population whose weather-dependent pay-offs lie in the neighborhood of (x,y).

The client's expected return of acting in accordance with the forecast p, assuming E_1 occurs, is given by:

$$\begin{aligned} R_1(p) &= \iint_{x,y} f(x)\,f(y)\;r_1[d_{\mathrm{p}}(x,y)]\mathrm{d}x\mathrm{d}y \\ &= \int_{y=-\infty}^{\infty} y f(y) \int_{-\infty}^{yp/1-p} f(x)\mathrm{d}x = \int_{-\infty}^{\infty} y f(y)\;F(yp/1-p)\mathrm{d}y \end{aligned} \tag{4}$$

where $F(\cdot)$ stands for the cumulative distribution associated with density f. Likewise, the expected return in case E_0 occurs is:

$$R_0(p) = \int_{-\infty}^{\infty} x f(x)\,F\left(\frac{1-p}{p}x\right)\ \mathrm{d}x = R_1(1-p). \tag{5}$$

In line with the philosophy of natural scoring rules the client ought to pass on to the forecaster the economical worth of acting in accordance with the forecast. Hence, (4) and (5) represent a natural scoring rule reflecting (in a condensed way) an underlying economic environment characterized by $f(\cdot)$.

We now show how simple density functions give rise to several familiar scoring rules.

Example 1. A uniform density

$$f(y) = \begin{cases} \dfrac{1}{2a} & -a \lesssim y \leq a, \\ 0 & |y| > a, \end{cases} \tag{6}$$

yields:

$$R_1(p) = \begin{cases} \dfrac{a}{6}\dfrac{1-p}{p} & p \leq \dfrac{1}{2}, \\ \dfrac{a}{12}\left[3 - \dfrac{(1-p)^2}{p^2}\right] & p \geq \dfrac{1}{2}. \end{cases} \tag{7}$$

Example 2. An exponential density

$$f(y) = \begin{cases} \lambda e^{-\lambda y} & y \geq 0, \\ 0 & y < 0, \end{cases} \tag{8}$$

yields the quadratic scoring rule:

$$R_1(p) = \frac{1}{\lambda}[1 - (1-p)^2]. \tag{9}$$

A quadratic scoring rule also results from a double sided exponential density $f(y) = \frac{\lambda}{2}e^{-\lambda|y|}$.

Example 3. A normal density

$$f(y) = \frac{1}{\sqrt{2\pi\sigma}}e^{-y^2/2\sigma^2} \tag{10}$$

yields the spherical scoring rule:

$$R_1(p) = \frac{\sigma^2}{\sqrt{2\pi}}\frac{p}{\sqrt{p^2+(1-p)^2}}. \tag{11}$$

As can be expected, the greater the variance σ^2, the higher the value of the forecast p, and the higher the score $R_1(p)$.

Example 4. A Cauchy density

$$f(y) = \begin{cases} \dfrac{2/\pi}{1+y^2} & y \leq 0, \\ 0 & y > 0, \end{cases} \tag{12}$$

yields the logarithmic scoring rule:

$$R_1(p) = \frac{2}{\pi}\log p. \tag{13}$$

The logarithmic divergence of $R_1(p)$ at $p \to 0$ can now be given an economic vindication, the mean assets the client anticipates losing if he follows the advice $p = 0$ and E_1 occurs indeed approaches infinity due to the slow decay of the Cauchy density.

We now examine how the salient features of the underlying density f influences the shape of the resultant scoring rule.

STRICT ADMISSIBILITY ~~MONOTONICITY~~

Every pay-off distribution $f(x,y)$ gives rise to an admissible scoring rule. This is evident from the fact that [see equation (4)] $R_1(p)$ is simply a linear superposition of the scores $r_1[d_p(x,y)]$ which constitute a natural (and therefore admissible) scoring rule for any fixed (x,y).

Strict admissibility can be demonstrated by showing that

$$\frac{d\bar{R}}{dp}(\pi,p) = 0$$

has a unique solution at $p = \pi$. From (4) we write:

$$\frac{\mathrm{d}}{\mathrm{d}p}\bar{R}(\pi,p) = \frac{\mathrm{d}}{\mathrm{d}p}[\pi\, R_1(p) + \overset{(1-\pi)}{\cancel{(1-p)}}R_0(p)]$$

$$= \frac{\pi}{(1-p)^2}G(p) - \frac{(1-\pi)}{p^2}G(1-p) \qquad (14)$$

where:

$$G(p) = \int_{-\infty}^{\infty} y^2 f(y)\, \overset{f(yp/1-p)}{\cancel{p(1-p)}}\, \mathrm{d}y. \qquad (15)$$

A change of variables $y = [(1-p)/p]\cancel{x}$ in (15) yields the relation:

$$G(p) = \frac{1-p^3}{p}G(1-p). \qquad (16)$$

and so, the stationary points of $\bar{R}(\pi,p)$ must satisfy

$$\frac{\pi}{(1-\pi)}G(p) = \frac{p}{(1-p)}G(p). \qquad (17)$$

This equation has a unique solution $p = \pi$ whenever $G(p)$ is non-zero. An inspection of equation (15) shows that for $0 < p < 1$ the integrand is positive on some finite interval whenever $f(y)$ is non-zero on a finite interval containing $y = 0$. Furthermore, taking the derivative of (14) gives

$$\frac{\mathrm{d}^2}{\mathrm{d}p^2}\bar{R}(\pi,p) = -\frac{G(\pi)}{(1-\pi)^3} < 0, \qquad (18)$$

showing that the stationary point $p = \pi$ is a maximum of $\bar{R}(\pi,p)$.

We conclude that every $f(y)$ which is non-zero on an interval around $y = 0$ gives rise to a strictly admissible scoring rule.

STRICT MONOTONICITY

Taking the derivative of (4) with respect to p, gives:

$$R_1'(p) = \frac{1}{(1-p)^2}G(p). \qquad (19)$$

Hence, under the previous condition guaranteeing a positive $G(p)$, $R'_1(p)$ is strictly positive for all $p < 1$.

END-POINTS CHARACTERISTICS

The behavior of $R_1(p)$ near $p = 0$ can be related to f by examining equation (4). Successive differentiations of (4) yield:

$$R_1(0) = F(0)\, E(y), \qquad (20)$$

$$R_1'(0) = f(0)\, E(y^2), \qquad (21)$$

$$R_1''(0) = 2E(y^2) + f'(0)\, E(y^3). \qquad (22)$$

Clearly, $R''_1(p)$ can be either positive or negative.

Near $p = 1$, equation (4) gives:

$$R_1(1) = E[\max(0,y)], \tag{23}$$

$$R_1'(1) = 0 \qquad \text{if } E(y^2) \text{ exists}, \tag{24}$$

$$R_1''(1) = -f(0)\,E(y^2). \tag{25}$$

Relations (20) to (25) are only valid when f possesses the appropriate order moments. For example, the logarithmic score seems to violate equation (24) as the Cauchy density does not possess a second moment.

We are not sure at the present whether equation (4) can be inverted to give a density f for any given R. That is, it is not clear whether every admissible scoring rule can be modeled as a natural byproduct of some economic environment f. The fact, however, that the common scoring rules can be generated by simple densities indicate that the model is not too restrictive.

3. Discussion

This paper demonstrates that an economic interpretation to standard information measures can be cast in a relatively simple betting context. Using the element of uncertainty concerning the magnitudes of the stakes in future confrontations it is possible to generate strictly convex information measures in a decision setup with only a finite number of actions. This construct reduces the problem of selecting an appropriate scoring rule (or information measure) to that of characterizing the anticipated economic impact of the forecast at hand.

Consider, for example, the classical parameter estimation problem of finding the "best" estimate p of a probabilistic parameter π, given a set of observations. Most teachers of statistics (the author among them) find it rather awkward to convince students that the square error criterion $(\pi-p)^2$, has, aside from its mathematical convenience, an authentic significance to justify its textbook popularity. Even demonstrating that $(\pi-p)^2$ is, under a quadratic scoring rule, the forecaster's loss of reporting p while π applies, offers but a minor comfort. The choice of the quadratic scoring rule in itself seems artificial. Based on equation (9), however, one can state: "The minimum square error estimator is optimal for decision situations with exponentially distributed future pay-offs." In a given estimation problem it is easier to assess whether such distribution is a reasonable one tha~~t~~n it is to speculate on whether the square error criterion is an appropriate loss function.

The value of the information contained in probabilistic assertions has been the subject of many discussions. Attempts to give the Shannon's entropy $H(\mathrm{p}) = -\Sigma p_i \log p_i$ a unique economical interpretation outside the field of communication have remained all but convincing (Marschak, 1972; White, 1975). In fact, McCarthy (1956) and others (DeGroot, 1962) have shown that any convex function $S(\mathrm{p})$ could represent the worth of probabilistic knowledge under specially configured circumstances. Equation (15) can be used to give Shannon's entropy a more natural interpretation. $H(\mathrm{p})$ is simply the expected reward $\bar{R}(\mathrm{p},\mathrm{p})$ of an honest forecaster employed under a logarithm scoring rule and, if the client accepts the forecaster's report p, it is also the economical worth of the knowledge contained in p to a client with a Cauchy-like distribution of future payoffs. Similarly, the use of $H(\mathrm{p})$ as a measure of *approximation* for identification of probabilistic automata

(Gaines, 1977) would be justified in situations where the modeller perceives Cauchy-like pay-off distributions in future decisions.

The construct of uncertain pay-off distribution allows the empirical psychologist to relate peculiar behavior observed under a certain scoring rule to assertions concerning human behavior in a corresponding economic environment. Likewise, mental procedures found optimal under a certain scoring rule may explain behavior in the corresponding environment.

For example, assume a system with only a limited memory is allowed to inspect a long list of N independent truth statements. At a later time the list is removed and the system is asked to provide a probabilistic estimate of the truth of each statement on the original list. An information theoretic study shows (Pearl, 1977) that under a logarithmic scoring rule the optimal mnemonic strategy would be to devote the entire available memory to an exact record of a portion N' of the propositions, ignore the remaining N–N' propositions, answer with certainty ($p = 0,1$) questions regarding the former and with $p=\frac{1}{2}$ those regarding the latter. The optimality of this behavior is unique to logarithmic scoring rules; a quadratic rule, in contrast, could encourage the distribution of memory resources over the entire list and the use of intermediate values of probability beside 1, $\frac{1}{2}$, 0.

Such theoretical findings could not be related to human information-processing behavior unless one finds models of natural environments which tend to create in humans the perceptions of operating under a logarithmic or a quadratic rule. Equations (9) and (13) imply that such perceptions and their associated mnemonic strategies are natural in decision situations where the anticipation of action-dependent pay-offs is of a certain character.

References

Brown, T. A. (1970). Probabilistic forecasts and reproducing scoring systems. *RM-6299-ARPA*. Santa Monica, California: RAND Corporation, June.

DeGroot, M. H. (1962). Uncertainty, information, and sequential experiments. *Annals of Mathematical Statistics*, **33**, 404–419.

Gaines, B. R. (1977). System identification, approximation and complexity. *International Journal of General Systems*, **3**, (3), 145–174.

McCarthy, J. (1956). Measures of the value of information. *Proceedings of the National Academy of Sciences*, [illegible], 654–655.

Marschak, J. (1972). Optimal systems for information and decision. In *Techniques of Optimization*. New York and London: Academic Press, Inc.

Murphy, A. H. (1966). A note on the utility of probabilistic predictions and the probability score in the cost-loss ratio decision situation. *Journal of Applied Meteorology*, **5**, 534–537.

Pearl, J. (1974). On the management of probability assessors. *UCLA-ENG-PAPER-0375*, February.

Pearl, J. (1977). On summarizing data using probabilistic assertions, *IEEE Transactions on Information Theory IT-23*, No. 4, July, 459–465.

Raiffa, H. (1964) Assessments of probabilities. Unpublished report, January.

Savage, L. J. (1971). Elicitation of personal probabilities and expectations. *Journals of the American Statistical Association*, 66(336).

Shuford, E. H. A., Massengill, H. E. (1966). Admissible probability measurement procedures. *Psychometrica*, **31** (2), June, 125–145.

White, D. J. (1975). Entropy and decision. *Operational Research Quarterly*, 26(I), 15–23.

Winkler, R. L. (1971). Probabilistic prediction: some experimental results. *Journal of the American Statistical Association*, **66** (366).