Correspondence

On the Practical Implication of Mutual Information for Statistical Decisionmaking

Fumio Kanaya and Kenji Nakagawa

Abstract—**A basic mathematical function that conjoins the two key conceptions of mutual information and Bayes risk is defined. Then based on that function, some asymptotic theorems that verify an important implication of mutual information in the context of practical Bayesian decisionmaking are proven.**

Index Terms—**Rate-distortion theory, statistical decision theory, Bayes risk, mutual information, rate-distortion function.**

I. Introduction and Motivation

Mutual information that describes the amount of information one random variable gives about a second random variable is one of the most fundamental information measures in information theory. In this discipline the concept of mutual information has emerged in conjunction with efficient coding of sources and their reliable transmission over noisy channels. On the other hand, it has also been known for some time that the concept of mutual information is related closely to the discipline of statistical decision theory because in the context of statistical decisionmaking, it is common to utilize any information that is provided by an observable random variable about the unknown true value of an underlying random parameter. In the light of certain common aspects emerging from the fundamental conception of mutual information, it is fairly reasonable to attempt to look at these two different disciplines from a unified perspective. This point is supported, for instance, by Berger's statement: "Rate distortion theory provides knowledge about how the frequency of faulty categorization will vary with the number and quality of the observations. More importantly, it also gives insight into what set of observations would provide the most information about the objects in question relative to the criterion of proper categorization and, therefore, is of potential value in the design of efficient pattern recognition devices" [1, p. 9]. This statement raises the possibility that the rate distortion theory in the field of source encoding is applicable to the basic statistical decision problem of pattern recognition.

However, to our knowledge, such speculations seem to have remained heuristic, and thus it seems worthwhile to justify them by mathematically rigorous arguments. This is the motivation of our work.

In the subsequent arguments we deal with the aforementioned interdisciplinary problem between information theory and statistical decision theory in rigorous mathematical formulations. First of all, we define a basic mathematical function to formulate a certain critical relationship that holds true between the amount of information which is contained in the observations available to a decisionmaker and the Bayes risk that is achievable in the decisions made by utilizing these observations. Then, after demonstrating some major properties of this basic function, we prove two main theorems of this correspondence: a decisionmaking theorem and its converse. By the former theorem one is guaranteed that when the length of the decisionmaking procedure is sufficiently large and a proper observation is used, the probability of the actual average loss being greater than a certain prescribed value goes asymptotically to zero. Proper observations can necessarily be found among those that contain at least the amount of information required by the basic function. Its converse theorem asserts the negation of the convergence to zero if the amount of information provided by the observations is less than a certain permissible minimum bound dictated by the basic function, whatever decisionmaking procedure one devises. Finally, we give one theorem that suggests that the speed of this convergence is exponential for a sufficiently large sample size.

Manuscript received May 31, 1989, revised January 3, 1991. This work was presented in part at the 1985 IEEE International Symposium on Information Theory, Brighton, England, June 1985 and at the 1988 IEEE International Symposium on Information Theory, Kobe, Japan, June 1988.

The authors are with NTT Transmission Systems Laboratories, 1-2356 Take Yokosuka-shi, Kanagawa 238-03, Japan.

IEEE Log Number 9144609.

II. Definition of the Basic Function and Its Properties

Let S be an unknown parameter taking values in the finite set $\Theta = \{\theta\}$, and let X be a random variable taking values in the finite set $\mathcal{X} = \{x\}$. It is assumed that a prior distribution $P(\theta)$ on Θ and a conditional probability matrix $W = \{W(x|\theta)\colon \theta \in \Theta, x \in \mathcal{X}\}$ are given. Hence, the joint distribution $\Pr[S = \theta, X = x] = P(\theta)W(x|\theta)$. After having observed x, one has to choose an action α from a finite set $\mathcal{A} = \{\alpha\}$. Without loss of generality, we assume that the loss function $\rho\colon \Theta \times \mathcal{A} \to R^+ \triangleq [0, \infty)$ satisfies the following condition:

$$\min_{\alpha \in \mathcal{A}} \rho(\theta, \alpha) = 0, \qquad \forall \theta \in \Theta. \tag{1}$$

Given a decision function $\psi\colon \mathcal{X} \to \mathcal{A}$, risk is defined by

$$r(\psi, P, W) = \sum_{\theta \in \Theta} \sum_{x \in \mathcal{X}} P(\theta)W(x|\theta)\rho(\theta, \psi(x)). \tag{2}$$

Let Ψ be the set of all possible decision functions. Then, the minimum attainable risk defined by the subsequent equation is referred to as the Bayes risk:

$$r(P, W) \triangleq \min_{\psi \in \Psi} r(\psi, P, W). \tag{3}$$

In the context of statistical decision problems, one can obtain information about θ by observing x. This information is then used for making decisions. The amount of information that is provided by the observations about the unknown parameter is given by the well-defined mutual information between S and X, that is,

$$I(P, W) = \sum_{\theta \in \Theta} \sum_{x \in \mathcal{X}} P(\theta)W(x|\theta) \log \frac{W(x|\theta)}{\sum_{\theta \in \Theta} P(\theta)W(x|\theta)}. \tag{4}$$

0018-9448/91/0700-1151$01.00

Now we are ready to define a basic mathematical function $R(P, L)$. For a fixed nonnegative real number L, denote by $\mathscr{W}_L$ the set of all L-admissible conditional probability matrices, i.e.,

$$\mathscr{W}_L \triangleq \{W: r(P,W) \le L\}, \qquad \forall L \ge 0. \tag{5}$$

Then, the basic function $R(P, L)$ is the function that is given by the solution of the following extremum problem:

$$R(P,L) \triangleq \min_{W \in \mathscr{W}_L} I(P,W), \qquad \forall L \ge 0. \tag{6}$$

It is evident from assumption (1) that the set $\mathscr{W}_L$ is nonempty for all $L \ge 0$. It is also apparent that the set $\mathscr{W}_L$ is compact, since $\mathscr{W}_L$ is the inverse image of a closed interval $[0, L]$ under a continuous real-valued function $r(P,W)$ of W. Thus, the existence of the basic function $R(P, L)$ is ensured for all $L \ge 0$ because the continuous real-valued function $I(P,W)$ must assume a minimum in any nonempty compact set of W.

Apparently in definition $R(P, L)$ is very similar to the well-known rate distortion function in the field of source encoding. However, $R(P, L)$ is more difficult to evaluate because the Bayes risk $r(P,W)$ is not a simple average distortion, but is itself the solution of a different functional optimization problem that is essential to any decisionmaking procedure.

We are now in the position to demonstrate fundamental properties of $R(P, L)$. First of all we give the following lemma.

Lemma 1: For each decision function $\psi \in \Psi$, let $R(P, \psi, D)$ be the rate-distortion function associated with the hypothetical distortion measure $d_\psi: \Theta \times \mathscr{X} \to R^+$ that is defined as

$$d_\psi(\theta, x) \triangleq \rho(\theta, \psi(x)), \qquad \forall \theta \in \Theta, \quad \forall x \in \mathscr{X}. \tag{7}$$

That is, let $R(P, \psi, D)$ be the solution to the following optimization problem:

$$\begin{cases} \mathscr{W}_D(\psi) \triangleq \left\{W: \sum_{\theta \in \Theta} \sum_{x \in \mathscr{X}} P(\theta) W(x|\theta) d_\psi(\theta, x) \le D\right\}, \\ R(P,\psi,D) = \min_{W \in \mathscr{W}_D(\psi)} I(P,W). \end{cases} \tag{8}$$

Then,

$$R(P,L) = \min_{\Psi \in \Psi} R(P,\psi,L), \qquad \forall L \ge 0, \quad \forall P \text{ on } \Theta. \tag{9}$$

Proof: Let $\mathscr{V}_L$ be the union of a collection of sets $\{\mathscr{W}_L(\psi): \psi \in \Psi\}$, i.e.,

$$\mathscr{V}_L \triangleq \bigcup_{\psi \in \Psi} \mathscr{W}_L(\psi). \tag{10}$$

The proof immediately follows from the proposition that equation $\mathscr{V}_L = \mathscr{W}_L$ holds for the set $\mathscr{W}_L$ that is previously defined by (5). Since it is apparent that $\mathscr{V}_L \supset \mathscr{W}_L$, proof is completed if we demonstrate the opposite relation.

Given any conditional probability matrix $U \in \mathscr{V}_L$, it follows from the definition of $\mathscr{V}_L$ that there exists a nonempty subset $\Delta \subset \Psi$ consisting of every decision function ψ that satisfies the following:

$$\sum_{\theta \in \Theta} \sum_{x \in \mathscr{X}} P(\theta) U(x|\theta) d_\psi(\theta, x) \le L. \tag{11}$$

Since it is obvious that Δ is a finite set, a member δ must be found in it that satisfies the following minimal condition:

$$\sum_{\theta \in \Theta} \sum_{x \in \mathscr{X}} P(\theta) U(x|\theta) d_\delta(\theta, x) \le \sum_{\theta \in \Theta} \sum_{x \in \mathscr{X}} P(\theta) U(x|\theta) d_\psi(\theta, x), \qquad \forall \psi \in \Delta. \tag{12}$$

Thus, considering the definition of $d_\psi(\theta, x)$ and of Δ, we obtain

$$L \ge \sum_{\theta \in \Theta} \sum_{x \in \mathscr{X}} P(\theta) U(x|\theta) d_\delta(\theta, x) = \min_{\psi \in \Psi} \sum_{\theta \in \Theta} \sum_{x \in \mathscr{X}} P(\theta) U(x|\theta) \rho(\theta, \psi(x)). \tag{13}$$

Since this implies that $U \in \mathscr{W}_L$, it is proved that $\mathscr{V}_L \subset \mathscr{W}_L$.

This lemma assures us that the function $R(P, L)$ can be computed by determining the minimum among a finite-number of rate-distortion functions, because $|\Psi|$ is equal to $|\mathscr{A}|^{|\mathscr{X}|}$, which is obviously finite. Here we denote the cardinality of a set by $|\cdot|$. Note that in the aforementioned case, respective distortion measures are induced from one and only one loss function $\rho(\theta, \alpha)$ through decision functions as defined by (7). Next, we demonstrate an important property of the basic function $R(P, L)$ that reveals the possibility of its evaluation without explicitly undertaking the often very difficult and impractical solution of the Bayes risk for a decision problem.

Theorem 1: Let $R(P, \rho, D)$ be the rate distortion function that regards Θ and $\mathscr{A}$ as a hypothetical source alphabet and reproduction alphabet, respectively, with the loss function $\rho(\theta, \alpha)$ adopted as the hypothetical distortion measure between these two alphabets. Then the following equality holds:

$$R(P,L) = R(P,\rho,L), \qquad \forall L > 0, \quad \forall P \text{ on } \Theta. \tag{14}$$

This theorem can be derived as a rather immediate consequence of both the last lemma and Stjernvall's dominance theory [2]. Prior to proof of this theorem, we will state the relevant definitions and theorem given by Stjernvall in appropriate terminology.

Definition 1: Let $d(x, y)$ be a distortion measure between finite alphabets $\mathscr{X} = \{1, 2, \cdots, |\mathscr{X}|\}$ and $\mathscr{Y} = \{1, 2, \cdots, |\mathscr{Y}|\}$. We can define the vectors as

$$\boldsymbol{d}(y) = (d(1,y), d(2,y), \cdots, d(|\mathscr{X}|, y)), \qquad \forall y \in \mathscr{Y}.$$

Definition 2: Let the convex hull of a set $\mathscr{A}$ be denoted by the symbol $\operatorname{conv}\mathscr{A}$. The characteristic set $\mathscr{B}(d)$ of the distortion measure $d(x, y)$ can then be defined as

$$\mathscr{B}(d) = \operatorname{conv}\mathscr{C}(d) + \mathscr{K},$$

where

$$\mathscr{C}(d) = \{\boldsymbol{d}(y): y \in \mathscr{Y}\}$$

and

$$\mathscr{K} = \{\boldsymbol{k}: k_x \ge 0, \forall x \in \mathscr{X}\}.$$

Stjernvall's Theorem [2, p. 800]: Assume the two distortion measures $d_1: \mathscr{X} \times \mathscr{Y} \to R^+$ and $d_2: \mathscr{X} \times \mathscr{Z} \to R^+$ satisfy the condition $\mathscr{B}(d_1) \subset \mathscr{B}(d_2)$. Then, for all distortion $D > 0$ and all proability distributions P on $\mathscr{X}$

$$R(P, d_1, D) \ge R(P, d_2, D),$$

where $R(P, d, D)$ denotes the rate-distortion function associated with the distortion measure d.

Now we prove Theorem 1.

Proof of Theorem 1: Let the distortion measure defined by the loss function $\rho(\theta, \alpha)$ be denoted by $d_0: \Theta \times \mathscr{A} \to R^+$ and the distortion measure defined by $\rho(\theta, \psi(x))$ for any given ψ be denoted by $d_\psi: \Theta \times \psi(\mathscr{X}) \to R^+$, where $\psi(\mathscr{X})$ denotes the image of $\mathscr{X}$ under the mapping $\psi: \mathscr{X} \to \mathscr{A}$. Then it is apparent from Definition 2 that $\mathscr{B}(d_\psi) \subset \mathscr{B}(d_0)$, since $\psi(\mathscr{X}) \subset \mathscr{A}$ for all $\mathscr{X}$ and all $\psi \in \Psi$. Thus in the light of Stjernvall's Theorem, for

all $\mathscr{X}$ and all $\psi \in \Psi$,

$$R(P,\psi,D) \geq R(P,\rho,D), \qquad \forall D > 0, \quad \forall P \text{ on } \Theta. \quad (15)$$

Now, consider some $\mathscr{X}$ with $|\mathscr{X}| \geq |\mathscr{A}|$. It is then obvious that a decision function ϕ must exist for which $\phi(\mathscr{X}) = \mathscr{A}$ holds. We now prove that at least for this ϕ

$$R(P,\phi,D) = R(P,\rho,D), \qquad \forall D > 0, \quad \forall P \text{ on } \Theta. \quad (16)$$

For the purpose of this proof let W^o be the optimal W that achieves $R(P,\rho,D)$. Then

$$\sum_{\theta \in \Theta} \sum_{\alpha \in \mathscr{A}} P(\theta) W^o(\alpha|\theta) \rho(\theta,\alpha) \leq D, \quad (17)$$

$$R(P,\rho,D) = \sum_{\theta \in \Theta} \sum_{\alpha \in \mathscr{A}} P(\theta) W^o(\alpha|\theta) \log \frac{W^o(\alpha|\theta)}{PW^o(\alpha)}, \quad (18)$$

where $PW^o(\alpha) = \sum_\theta P(\theta) W^o(\alpha|\theta)$. For each $x \in \phi^{-1}(\alpha)$, we define $W^*(x|\theta)$ by

$$W^*(x|\theta) \triangleq \frac{1}{|\phi^{-1}(\alpha)|} W^o(\alpha|\theta). \quad (19)$$

Then we have

$$W^o(\alpha|\theta) = |\phi^{-1}(\alpha)| W^*(x|\theta) = \sum_{x \in \phi^{-1}(\alpha)} W^*(x|\theta). \quad (20)$$

For this $W^*(x|\theta)$, we have from (17) and (20) that

$$\begin{aligned} &\sum_{\theta \in \Theta} \sum_{x \in \mathscr{X}} P(\theta) W^*(x|\theta) \rho(\theta, \phi(x)) \\ &\quad = \sum_{\theta \in \Theta} \sum_{\alpha \in \mathscr{A}} \sum_{x \in \phi^{-1}(\alpha)} P(\theta) W^*(x|\theta) \rho(\theta, \phi(x)) \\ &\quad = \sum_{\theta \in \Theta} \sum_{\alpha \in \mathscr{A}} P(\theta) W^o(\alpha|\theta) \rho(\theta,\alpha) \\ &\quad \leq D. \end{aligned} \quad (21)$$

On the other hand, we have from (18), (20), and (21) that

$$\begin{aligned} &R(P,\rho,D) \\ &\quad = \sum_{\theta \in \Theta} P(\theta) \sum_{\alpha \in \mathscr{A}} \sum_{x \in \phi^{-1}(\alpha)} W^*(x|\theta) \log \frac{|\phi^{-1}(\alpha)| W^*(x|\theta)}{|\phi^{-1}(\alpha)| PW^*(x)} \\ &\quad = \sum_{\theta \in \Theta} P(\theta) \sum_{\alpha \in \mathscr{A}} \sum_{x \in \phi^{-1}(\alpha)} W^*(x|\theta) \log \frac{W^*(x|\theta)}{PW^*(x)} \\ &\quad = \sum_{\theta \in \Theta} \sum_{x \in \mathscr{X}} P(\theta) W^*(x|\theta) \log \frac{W^*(x|\theta)}{PW^*(x)} \\ &\quad \geq R(P,\phi,D). \end{aligned} \quad (22)$$

Thus, by (22) and (15) we obtain (16). Finally it follows from (15), (16), and Lemma 1 that

$$R(P,L) = R(P,\phi,L) = R(P,\rho,L), \qquad \forall L > 0, \quad \forall P \text{ on } \Theta. \quad (23)$$

This completes the proof of Theorem 1.

By virtue of this theorem, it becomes possible to reduce the computation of the basic function $R(P,L)$ to that of the well-defined rate-distortion function in information theory, which circumvents the cumbersome solution of the Bayes risk. Convexity, continuity and strictly decreasing property of the rate-distortion function is guaranteed also for the basic function $R(P,L)$ by the last theorem.

As a final consequence of this section, we give an important property of the stochastic matrix that attains a point on the $R(P,L)$ curve. It is evident from the argument previously made on the existence of the basic function that for all $L \geq 0$ there exists at least one stochastic matrix that achieves a point $(L, R(P,L))$ on the $R(P,L)$ curve. However, it is to be noted that the aforementioned optimum stochastic matrix is not automatically computable by way of the extremum problem (6) that defines the basic function $R(P,L)$, because the cardinality of the observation data set $\mathscr{X}$ is not specified and hence it may be an arbitrary finite integer. Now, by means of the next theorem, which can be easily obtained from the proof of Theorem 1, it is assured that we can find at least one optimum stochastic matrix achieving the point $(L, R(P,L))$ within the L-admissible set comprising only those matrices of a fixed size of $|\Theta| \times |\mathscr{A}|$.

Theorem 2: To every stochastic matrix W: $\Theta \to \mathscr{X}$ that achieves a point $(L, R(P,L))$ on the curve of $R(P,L)$ under the condition that $|\mathscr{X}| > |\mathscr{A}|$, there exists a corresponding stochastic matrix V: $\Theta \to \mathscr{Y}$ such that V also achieves the same point and $|\mathscr{Y}| = |\mathscr{A}|$ holds.

By combining this theorem with Theorem 1, we are led to the conclusion that every stochastic matrix achieving a point on the rate-distortion $R(P,\rho,L)$ curve is a representative of the set of all solution matrices to the extremum problem (6) that defines the basic function $R(P,L)$. It should be noticed that the random variable Y that takes values in $\mathscr{Y}$ and is specified by the joint distribution $\Pr[S = \theta, Y = y] = P(\theta)V(y|\theta)$ is actually a sufficient statistic of the corresponding observed random variable X for the unknown parameter S (see, e.g., [4, pp. 303–306]).

III. Main Theorems

The basic function specifies by definition the minimum possible amount of information that must be given by the observations about the unknown parameter for one to achieve the Bayes risk not greater than the prescribed value L. The purpose of the subsequent argument is to demonstrate the practical implication of the basic function $R(P,L)$ through proving main theorems that relate it to the asymptotic behavior of an actual decisionmaking procedure.

Now consider an actual decisionmaking context in which a component statistical decision problem having identical generic structure is repeated n times. We assume that there is no relationship among the n repetitions. As a practical matter, the question of major interest to the decisionmaker is the actual (random) averaging loss that accrues from the real decisionmaking procedure. Suppose that a compound decision function $\mathbf{\Psi}^n$: $\mathscr{X}^n \to \mathscr{A}^n$ is used. Then the actual average loss having resulted from the n decisions is given by

$$\rho^n(\boldsymbol{\theta}, \mathbf{\Psi}^n(\boldsymbol{x})) = \frac{1}{n} \sum_{i=1}^{n} \rho(\theta_i, \Psi_i(\boldsymbol{x})), \quad (24)$$

where an n-length sequence of observed data is denoted by a vector $\boldsymbol{x} = x_1 x_2 \cdots x_n \in \mathscr{X}^n$ and an underlying sequence of true parameters is denoted by a vector $\boldsymbol{\theta} = \theta_1 \theta_2 \cdots \theta_n \in \Theta^n$. We also use for a compound decision function notation $\mathbf{\Psi}^n(\boldsymbol{x}) = \Psi_1(\boldsymbol{x}) \Psi_2(\boldsymbol{x}) \cdots \Psi_n(\boldsymbol{x}) \in \mathscr{A}^n$. Henceforth, we denote by $E_{S^n, X^n}[\cdot]$ the mathematical expectation of a random variable with respect to the joint distribution $\Pr[S^n = \boldsymbol{\theta}, X^n = \boldsymbol{x}] = P^n(\boldsymbol{\theta}) W^n(\boldsymbol{x}|\boldsymbol{\theta})$ of random variables S^n and X^n.

Prior to stating the main theorems, we give two preliminary lemmas needed for their proof.

Lemma 2: To every R satisfying the condition that $R(P,L) \leq R \leq \log|\Theta|$, there exist a stochastic matrix W: $\Theta \to \mathscr{X}$ satisfying $I(P,W) = R$ and, for every positive integer n, a compound

decision function Ψ^n: $\mathscr{X}^n \to \mathscr{A}^n$ such that

$$E_{S^n, X^n}[\rho^n(\theta, \Psi^n(x))] \le L.$$

Proof: Assume for the compound decision function Ψ^n a simple symmetric rule $\phi^n(x) = \phi(x_1)\phi(x_2)\cdots\phi(x_n)$ with each component ϕ being a Bayes rule against P that is determined by (3). Then the lemma follows immediately from the definition of the basic function $R(P, L)$ and the hypothesis that an identical statistical decision problem is repeated independently.

Lemma 3: Suppose that for a given positive integer n there exist a stochastic matrix W: $\Theta \to \mathscr{X}$ and a compound decision function Ψ^n: $\mathscr{X}^n \to \mathscr{A}^n$ such that

$$E_{S^n, X^n}[\rho^n(\theta, \Psi^n(x))] \le L.$$

Then we must have

$$I(P, W) \ge R(P, L).$$

Proof: First it can be shown by some calculation that for every positive integer n and every compound decision function Ψ^n: $\mathscr{X}^n \to \mathscr{A}^n$

$$E_{S^n, X^n}[\rho^n(\theta, \Psi^n(x))] = \frac{1}{n}\sum_{i=1}^{n} D_i,$$

where D_i represents the next distortion associated with a properly defined stochastic matrix W_i: $\Theta \to \mathscr{A}$

$$D_i \triangleq \sum_{\theta \in \Theta}\sum_{\alpha \in \mathscr{A}} P(\theta)W_i(\alpha|\theta)\rho(\theta, \alpha).$$

Hence, by using a data processing lemma (see, e.g., [3, pp. 55–56]) and a rate-distortion theorem (e.g., [5, Theorem 9.2.1]), we can easily obtain

$$I(P, W) \ge R(P, \rho, E_{S^n, X^n}[\rho^n(\theta, \Psi^n(x))]). \tag{25}$$

Then it follows by the strictly decreasing property of the rate-distortion function $R(P, \rho, D)$ and from the postulate $E_{S^n, X^n}[\rho^n(\theta, \Psi^n(x))] \le L$ that

$$I(P, W) \ge R(P, \rho, L). \tag{26}$$

Finally the lemma follows immediately from both (26) and Theorem 1.

Now we are ready to prove two main theorems. First we prove the decisionmaking theorem. The significance of this theorem is that it ensures that the probability of averaging loss greater than some prescribed value L goes asymptotically to zero for sufficiently large values of n, if one uses proper observations that yield the amount of information about unknown parameter not less than is required by the basic function $R(P, L)$.

Theorem 3 (Decisionmaking Theorem): To every R satisfying the condition that $R(P, L) \le R \le \log|\Theta|$, there exists a stochastic matrix W: $\Theta \to \mathscr{X}$ satisfying $I(P, W) = R$ and a sequence of n-length decisionmaking procedures $\{\Psi^n$: $\mathscr{X}^n \to \mathscr{A}^n\}$, $n = 1, 2, \cdots$ such that for every $\eta > 0$ and $\epsilon > 0$

$$\Pr[\rho^n(\theta, \Psi^n(x)) \le L + \eta] \ge 1 - \epsilon,$$

whenever $n \ge n_0(\eta, \epsilon)$.

Proof: For a pair of sequences $\theta = \theta_1\theta_2\cdots\theta_n \in \Theta^n$ and $x = x_1x_2\cdots x_n \in \mathscr{X}^n$ we denote its joint type by

$$\frac{1}{n}N(\theta, x|\theta, x) \triangleq \frac{1}{n}\sum_{i=1}^{n}\delta((\theta, x), (\theta_i, x_i)),$$

$$\forall (\theta, x) \in \Theta \times \mathscr{X}, \tag{27}$$

where $\delta(a, b)$ is defined as

$$\delta(a, b) \triangleq \begin{cases} 1, & \text{if } a = b \\ 0, & \text{otherwise} \end{cases}.$$

Then for any given prior P on Θ and stochastic matrix W: $\Theta \to \mathscr{X}$ we denote the set of all jointly typical sequences by (cf. [3])

$$T^n_{[PW]_{\delta_n}} \triangleq \left\{(\theta, x): \left|\frac{1}{n}N(\theta, x|\theta, x) - P(\theta)W(x|\theta)\right| \le \delta_n, \right.$$

$$\left. \forall (\theta, x) \in \Theta \times \mathscr{X}\right\}. \tag{28}$$

Here the additional requirement must be met such that $N(\theta, x|\theta, x) = 0$ whenever $P(\theta)W(x|\theta) = 0$ and the sequence $\{\delta_n\}$ is chosen such that

$$\delta_n \to 0, \qquad \sqrt{n}\cdot\delta_n \to \infty, \qquad \text{as } n \to \infty. \tag{29}$$

Now, according to Lemma 2, we can choose W: $\Theta \to \mathscr{X}$ and Ψ^n: $\mathscr{X}^n \to \mathscr{A}^n$ such that $I(P, W) = R$ and $\Psi^n(x) = \phi(x_1)\phi(x_2)\cdots\phi(x_n)$ with components ϕ satisfying $\Sigma_{\theta \in \Theta}\Sigma_{x \in \mathscr{X}} P(\theta)W(x|\theta)\rho(\theta, \phi(x)) \le L$. Then it follows from (27) and (28) that for every $(\theta, x) \in T^n_{[PW]_{\delta_n}}$

$$\rho^n(\theta, \Psi^n(x)) = \frac{1}{n}\sum_{\theta \in \Theta}\sum_{x \in \mathscr{X}} N(\theta, x|\theta, x)\rho(\theta, \phi(x))$$

$$\le L + \delta_n\rho_M|\Theta||\mathscr{X}|, \tag{30}$$

where we put

$$\rho_M \triangleq \max_{\theta \in \Theta, \alpha \in \mathscr{A}} \rho(\theta, \alpha).$$

On the other hand, by using Chebyshev's inequality, it is easily shown that

$$P^nW^n\left(T^n_{[PW]_{\delta_n}}\right) \ge 1 - \frac{1}{4n\delta_n^2}|\Theta||\mathscr{X}|. \tag{31}$$

Hence, considering (29), the theorem follows immediately from (30) and (31).

Due to the general feeling that in the practical statistical decisionmaking case the stochastic matrix W cannot be selected arbitrarily as opposed to the test channel in the rate distortion case, this theorem would not be considered of so much practical significance. However, it should be noticed that in the context of the practical pattern recognition, the problem of feature selection has long been recognized and intensely studied. (For an enlightening perspective of this field, see e.g., [6].) It could be considered from the statistical decision theoretic viewpoint that the essence of this problem virtually consists in the selection of the optimal stochastic matrix W within the admissible set that is determined by some practical constraints. Thus, for instance, it would be possible for this theorem to help prompt a new idea of optimal feature selection.

Next we prove the converse to the decisionmaking theorem.

Theorem 4 (Converse Theorem): For every stochastic matrix W, $\Theta \to \mathscr{X}$ subject to the constraint that $I(P, W) < R(P, L)$, there exists a positive number $\beta(P, W, L)$ such that

$$\Pr[\rho^n(\theta, \Psi^n(x)) > L] \ge \beta(P, W, L),$$

for every positive integer n and every n-length decisionmaking procedure Ψ^n: $\mathscr{X}^n \to \mathscr{A}^n$.

Proof: Suppose we are given an arbitrary compound decision function Ψ^n: $\mathscr{X}^n \to \mathscr{A}^n$. Then we have

$$\begin{aligned} E_{S^n,X^n}[\rho^n(\theta,\Psi^n(x))] &\triangleq \sum_{(\theta,x)} P^n(\theta)W^n(x|\theta)\rho^n(\theta,\Psi^n(x)) \\ &\le \left\{1 - \sum_{\rho^n(\theta,\Psi^n(x))>L} P^n(\theta)W^n(x|\theta)\right\}L \\ &\quad + \rho_M \sum_{\rho^n(\theta,\Psi^n(x))>L} P^n(\theta)W^n(x|\theta) \\ &= L + (\rho_M - L) \sum_{\rho^n(\theta,\Psi^n(x))>L} P^n(\theta)W^n(x|\theta). \end{aligned}$$

Hence, for every $L < \rho_M$,

$$\begin{aligned} \Pr[\rho^n(\theta,\Psi^n(x)) > L] &\triangleq \sum_{\rho^n(\theta,\Psi^n(x))>L} P^n(\theta)W^n(x|\theta) \\ &\ge \frac{E_{S^n,X^n}[\rho^n(\theta,\Psi^n(x))] - L}{\rho_M - L}. \end{aligned} \tag{32}$$

On the other hand, it follows from Lemma 3 that for every positive integer n, we must have

$$I(P,W) \ge R(P, E_{S^n,X^n}[\rho^n(\theta,\Psi^n(x))]). \tag{33}$$

Since by postulate of this theorem

$$R(P,L) > I(P,W), \tag{34}$$

it follows from the last two inequalities that

$$R(P,L) > I(P,W) \ge R(P, E_{S^n,X^n}[\rho^n(\theta,\Psi^n(x))]). \tag{35}$$

Let $\Lambda(P,R)$ be the inverse of the basic function $R(P,L)$. Then by using the strictly decreasing property of $R(P,L)$, it follows from the last inequality that for every n and Ψ^n: $\mathscr{X}^n \to \mathscr{A}^n$,

$$L < \Lambda(P, I(P,W)) \le E_{S^n,X^n}[\rho^n(\theta,\Psi^n(x))].$$

Combining this inequality with (32), we have

$$\Pr[\rho^n(\theta,\Psi^n(x)) > L] \ge \frac{\Lambda(P,I(P,W)) - L}{\rho_M - L}. \tag{36}$$

The converse theorem then follows from this inequality immediately if we put $\beta(P,W,L) = (\Lambda(P,I(P,W)) - L)/(\rho_M - L)$.

According to the converse theorem one can not always be guaranteed of achieving the average loss of the prescribed value L as the length n goes to infinity if one utilizes for decisionmaking the observations that provide less amount of information about the unknown parameter than is dictated by the basic function $R(P,L)$.

Consequently, through proving these two theorems it is possible to verify the practical implication that the mutual information has in the context of actual decisionmaking procedure.

Finally we give one theorem that concerns the speed of convergence in probability. We henceforth denote by $D(P\|Q)$ informational divergence between two distributions P and Q on a finite set $\mathscr{X}$, i.e.,

$$D(P\|Q) \triangleq \sum_{x\in\mathscr{X}} P(x) \log \frac{P(x)}{Q(x)},$$

and also denote by $D(V\|W|P)$ conditional informational divergence between two stochastic matrices V: $\mathscr{X} \to \mathscr{Y}$ and W: $\mathscr{X} \to \mathscr{Y}$, i.e.,

$$D(V\|W|P) \triangleq \sum_{x\in\mathscr{X}} P(x) D(V(\cdot|x)\|W(\cdot|x)).$$

Here notations $V(\cdot|x)$ and $W(\cdot|x)$ have been used for row vectors of respective stochastic matrices.

Theorem 5: Let L be a positive number less than ρ_M, and W: $\Theta \to \mathscr{X}$ an arbitrary stochastic matrix satisfying $I(P,W) \ge R(P, L-\eta)$ for some arbitrarily small $\eta > 0$. Then for every compound decision function Ψ^n: $\mathscr{X}^n \to \mathscr{A}^n$,

$$\begin{aligned} \varlimsup_{n\to\infty} \frac{1}{n} \log \Pr[\rho^n(\theta,\Psi^n(x)) > L] \\ = - \inf_{V:\, I(P,V) < R(P,L)} D(V\|W|P). \end{aligned}$$

Proof: Let V: $\Theta \to \mathscr{X}$ be an arbitrary stochastic matrix such that $I(P,V) < R(P,L)$. According to Theorem 4 we have for any given Ψ^n: $\mathscr{X}^n \to \mathscr{A}^n$,

$$P^nV^n(\{(\theta,x)\colon \rho^n(\theta,\Psi^n(x)) > L\}) > \beta(P,V,L) > 0, \quad \text{for } n = 1,2,\cdots. \tag{37}$$

Therefore, it follows immediately from Stein's Lemma (see., e.g., [3, p. 28]) that for every $\delta > 0$ and sufficiently large n

$$\begin{aligned} \frac{1}{n} \log P^nW^n(\{(\theta,x)\colon \rho^n(\theta,\Psi^n(x)) > L\}) \\ \ge - D(P\times V\|P\times W) - \delta. \end{aligned} \tag{38}$$

Here $P\times V$ and $P\times W$ designate the joint distributions $P(\theta)V(x|\theta)$ and $P(\theta)W(x|\theta)$, respectively. Thus, considering the fact that by definition

$$\Pr[\rho^n(\theta,\Psi^n(x)) > L] = P^nW^n(\{(\theta,x)\colon \rho^n(\theta,\Psi^n(x)) > L\}),$$

we obtain for every Ψ^n: $\mathscr{X}^n \to \mathscr{A}^n$ and sufficiently large n

$$\frac{1}{n} \log \Pr[\rho^n(\theta,\Psi^n(x)) > L] \ge - D(V\|W|P) - \delta, \qquad \forall \delta > 0. \tag{39}$$

Here we have used the identity $D(P\times V\|P\times W) = D(V\|W|P)$. Since V was arbitrary subject to the constraint that $I(P,V) < R(P,L)$ and $\delta > 0$ can be made arbitrarily small, proof follows immediately from (39).

It should be noted that by nonnegativity of informational divergence and the postulate $I(P,W) \ge R(P,L-\eta)$ for some $\eta > 0$, $\inf_V D(V\|W|P)$ is guaranteed to be strictly positive in the last theorem.

IV. Conclusion

We have presented an attempt to look at the two different disciplines of Shannon theory and Bayes statistical decision theory from a unifying perspective. We have viewed as key notions the mutual information in the former and the Bayes risk in the latter. First we have defined a basic mathematical function that conjoins them. Then, through mathematically rigorous arguments based on major properties of this basic function, we have proved a decisionmaking theorem and its converse which enables us to verify that even for actual statistical decisionmaking procedures, the mutual information has important pragmatic implications, as has long been recognized in conjunction with source and channel coding procedures.

Acknowledgment

The authors are grateful to the anonymous referees for their valuable suggestions and comments that improved the presentation of the results in this correspondence.

References

[1] T. Berger, *Rate Distortion Theory*. Englewood Cliffs, NJ: Prentice-Hall, 1971.
[2] J. Stjernvall, "Dominance—A relation between distortion measures," *IEEE Trans. Inform. Theory*, vol. IT-29, no. 6, pp. 798–807, Nov. 1983.
[3] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic Press, 1981.
[4] S. Guiaşu, *Information Theory with Application*. London: McGraw-Hill, 1977.
[5] R. G. Gallager, *Information Theory and Reliable Communication*. New York: John Wiley, 1968.
[6] S. D. Morgera and L. Datta, "Toward a fundamental theory of optimal feature selection: Part I," *IEEE Trans. Pattern. Anal. Machine Intell.*, vol. PAMI-6, pp. 601–616, Sept. 1984.

On the Wavelet Transform of Fractional Brownian Motion

J. Ramanathan and O. Zeitouni

***Abstract*—A theorem characterizing fractional Brownian motion by the covariance structure of its wavelet transform is established.**

***Index Terms*—Wavelet transform, fractional Brownian motion.**

I. Introduction

The wavelet transform of a function $f(t)$ is defined by the formula

$$\mathcal{W}f(t,a) = \mathcal{W}_a f(t) = \frac{1}{\sqrt{a}} \int f(s)\, g\left(\frac{t-s}{a}\right) ds,$$

where $g(t)$ is a fixed function, $t \in \mathbb{R}$, and $a \in \mathbb{R}^+$. This transform yields a joint time-scale representation of the original input function that has been of great recent interest. (See, e.g., [1], [2], and [4]).

In a recent correspondence, Flandrin [3] proposed the use of the wavelet transform to analyze the behavior of fractional Brownian motion, a highly nonstationary random process. (For a background on fractional Brownian motion and some of its applications, see [5] and [6]). The wavelet transform of a stochastic process $X(t)$ is a random field $\mathcal{W}X(t,a)$ on the upper half plane. The process $t \mapsto \mathcal{W}_a X(t)$ can be thought of as the component of the original process at scale a. A consequence of Flandrin's computation is that fractional Brownian motion is stationary at each fixed scale. In particular, when $X(t)$ is a fractional Brownian motion, the covariance of the process $t \mapsto \mathcal{W}_a X(t)$ is of the form

$$E[\mathcal{W}_a X(t)\, \mathcal{W}_a X(s)] = a^{\lambda} \rho\left(\frac{t-s}{a}\right), \tag{1}$$

where ρ is a positive definite function determined in an explicit manner by the order or the fractional Brownian motion and the defining function $g(t)$ of the wavelet transform. This fact is used by Flandrin to make rigorous sense of the spectral content of fractional Brownian motion.

It is natural to ask whether there are other Gaussian processes whose wavelet transforms have this natural covariance structure. In addition, are there any Gaussian processes whose wavelet transform is stationary with respect to the affine group (i.e., the statistics of the wavelet transform do not depend on translations and dilations of the process)? The purpose of this paper is to point out that the answers to both these questions are negative. In particular, fractional Brownian motion is characterized by the property that its wavelet transform has the form shown above in (1). A consequence is that there are no nontrivial Gaussian processes on the real line whose wavelet transform produces a random field that is stationary with respect to the affine group. It should be remarked that the results presented assume some fairly nonrestrictive growth conditions on the covariance of the process $X(t)$ and the kernel $g(t)$ used to define the wavelet transform (see Remark 2 at the end of the correspondence).

II. A Characterization of Fractional Brownian Motion

Let $X(t)$ be a Gaussian random process defined for $t \in \mathbb{R}$ such that the covariance $R(s,t) = E[X(s)X(t)]$ is continuous and satisfies

$$R(0,0) = 0$$

$$|R(s,t)| \le C\left(1 + |s|^2 + |t|^2\right)^{N/2}, \tag{2}$$

where $N \in Z$ is fixed. In addition, we also impose the following conditions on the analyzing function $g(t)$:

a) $\int_{-\infty}^{\infty} |g(t)|(1+|t|^2)^{N/2}\, dt < \infty$,
b) $\int_{-\infty}^{\infty} |\hat{g}(\xi)|^2 / |\xi|\, d\xi < \infty$, and
c) $\hat{g}$ is smooth and has a simple zero at the origin.

(The Fourier transform of g is denoted by $\hat{g}$.) The growth conditions for the process $X(t)$ and the analyzing wavelet $g(t)$ insure that the wavelet transform produces a random field with finite covariance.

Theorem: Suppose the wavelet transform $\mathcal{W}_a X$ satisfies

$$E[\mathcal{W}_a X(t)\, \mathcal{W}_a X(s)] = a^{\lambda} \rho\left(\frac{t-s}{a}\right),$$

where $\lambda \in \mathbb{R}$ and ρ is a positive semidefinite function on the real line. Then $X(t)$ is fractional Brownian motion of order $H = (\lambda - 1)/2$. In particular, $\lambda \in (1,3)$.

Manuscript received December 19, 1989; revised January 6, 1991. J. Ramanathan was supported by the MITRE sponsored Research Program. O. Zeitouni was supported in part by the Army Research Office under Contract DAAL03-86-K-0171. This work was performed while O. Zeitouni was visiting the Laboratory for Information and Decision Systems at M.I.T.

J. Ramanathan was with the MITRE Corporation, Mail Stop E025, Burlington Road, Bedford, MA 01730. He is now with the Department of Mathematics, Eastern Michigan University, Ypsilanti, MI 48197.

O. Zeitouni is with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa 32000, Israel.

IEEE Log Number 9100446.