LA-UR--93-919

DE93 010705

**TITLE: Information-Theoretic Performance Bounding of
Bayesian Identifiers**

AUTHOR(S): Scott D. Briles

**SUBMITTED TO: Proceedings for the International Society of Optical Engineers
(SPIE) Symposium on Aerospace and Remote Sensing 1993 in
Conference 1960 Automatic Object Recognition III.**

DISCLAIMER

This report was prepared as an account of work sponsored by an agency of the United States Government. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply its endorsement, recommendation, or favoring by the United States Government or any agency thereof. The views and opinions of authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof.

By acceptance of this article, the publisher recognizes that the U.S. Government retains a nonexclusive, royalty-free license to publish or reproduce the published form of this contribution or to allow others to do so, for U.S. Government purposes.

The Los Alamos National Laboratory requests that the publisher identify this article as work performed under the auspices of the U.S. Department of Energy.

Los Alamos Los Alamos National Laboratory
Los Alamos, New Mexico 87545

MASTER

Information-theoretic performance bounding of Bayesian identifiers

Scott D. Briles

Los Alamos National Laboratory*

Mechanical and Electronic Engineering Division

Mail Stop J580, Los Alamos, New Mexico 87545, USA

ABSTRACT

Statistical target recognition techniques perform well when the "true" target-signature data are deterministic. If the deterministic nature of the data exists and the probabilistic values of a given problem are known, then identifiers, based on Bayesian estimation theory, have great potential for solving the target identification problem. How well an identifier will perform is usually answered by Monte Carlo simulations or implementation experiments. An alternative to these performance analysis techniques is the use of information theory.

Information theory has long been applied to the investigation of data compression, which deals with average distortional measures. The association of Bayes risk with distortion allows for information-theoretic tools to be applied to the statistical target identification problem. The rate-distortion function of data compression can be extended to the Bayes rate-distortion function. Derived from designer-specified risk and identifier structure, the theoretical Bayes rate-distortion function relates mutual information to the identifier performance. Mutual information is determined from the target-signature data used by the identifier and by the nature of the noise imposed upon the data. Computational efforts in determining mutual information values allow for identifier performance bounds to be extracted from the Bayes rate-distortion function.

1. INTRODUCTION

The desire to know something, beyond its mere presence, about a target from whom data is being collected has been the motivation into research involving target identification. Efforts involved in determining the performance of various identifiers usually rely on Monte Carlo simulation or repeated implementation experiments. Both approaches require a considerable allocation of time and resources. However, the goal of gaining some estimations as to the performance of an identifier must be achieved if refinements in the identifier system are to be made. Variation in the data collected on a target, such as interrogation bandwidth and signal-to-noise ratio (SNR), are routinely being examined to evaluate diverse engineering tradeoffs. Assessing how the variation in collected data affects the performance of the identifier analytically is presented in this paper.

A type of identifier, known as a Bayesian identifier, is investigated throughout this paper. Properties of the Bayesian identifier allow for the use of information theory to lower bound the identifier's performance. Using the concept of mutual information, a relationship between mutual information and probability of identification can be proven. This relationship is referred to as the Bayesian rate distortion function, which determines the minimum value of mutual information need to obtain a given performance. The mutual information is expressed for the different targets under a certain data collection variation.

2. BAYESIAN IDENTIFICATION

Bayes decision theory is a fundamental statistical approach to target identification and is used to determine the optimum identifier. Figure 1 shows the Bayes identifier model. This approach is based on the assumption that all of the relevant probability functions are known and that an ideal sensor will produce unique reproducible responses. With this knowledge, when a measurement of a random variable(s) is observed, the best estimate of the state of the source (i.e., the identity of the target) can be made so that the risk is minimized.

*Los Alamos National Laboratory is operated by the University of California for the U.S. Department of Energy under contract W-7405-ENG-36.

Figure 1: Bayes identifier model.

If the state of the source is represented by θ , which is taken from a finite set of source states Θ with M elements, the given source state θ_i has an *a priori* probability $P(\theta_i)$ of occurring. The finite set Θ is distributed in accordance with the probability density function (pdf) $P(\theta)$. We do not observe the state of the source directly; rather the measurement x or a measurement vector $\vec{x}$ is observed. The measurement $\vec{x}$ is considered to be a continuous random vector whose distribution depends on the state of the source. It is contained in the infinite set $\vec{X}$, which defines the observation space. The relationship between the source state and the measurement vector is given by $W(\vec{x}|\theta_i)$, the state-conditional probability density function of $\vec{x}$.

If the *a priori* and state-conditional probability density functions are known, one can determine the probability that the source is in a given state by using Bayes rule:

$$P(\theta_i|\vec{x}) = \frac{W(\vec{x}|\theta_i)P(\theta_i)}{p(\vec{x})}, \quad (1)$$

where

$$p(\vec{x}) = \sum_{i=1}^M W(\vec{x}|\theta_i)P(\theta_i). \quad (2)$$

Once a measurement $\vec{x}$ is observed, Bayes rules allows the *a posteriori* probability $P(\theta_i|\vec{x})$ to be determined. In accordance with the observed measurement $\vec{x}$, a decision α_j is made. The decision α_j is a member of a finite set A , composed of M' elements, which defines the destination space. A loss, or penalty, is assigned to each decision α that is associated with all the source states θ .

Thus, the loss associated with making decision α_j when the source is in state θ_i , is $\rho(\theta_i, \alpha_j)$. The function $\rho(\theta, \alpha)$ is the loss function in which the conventional restrictions,

$$\min_{\alpha \in A} \rho(\theta, \alpha) = 0, \quad \forall \theta \in \Theta, \quad (3)$$

and $\rho(\theta, \alpha)$ can have no infinite value, are not violated.

The true state of the source is not observed; only the measurement vector $\vec{x}$ can be seen. Thus, if the vector $\vec{x}$ is observed, the expected loss of making the decision α_j can be determined by using the *a posteriori* probabilities. The expected loss of taking action α_j after observing the measurement vector $\vec{x}$ is

$$r(\alpha_j|\vec{x}) = \sum_{i=1}^M \rho(\theta_i, \alpha_j)P(\theta_i|\vec{x}) \quad (4)$$

Therefore, this is the conditional risk of making decision α_j . Whenever a particular vector $\vec{x}$ is observed, the expected loss can be minimized by selecting the decision α that minimizes the conditional risk.

The selection of a particular decision for a given observation vector is determined by the decision function, ψ . The decision function ψ performs a mapping from the observation space $\bar{X}$ to the destination space A (i.e., $\psi: \bar{X} \rightarrow A$). For a given decision function, ψ the average risk can be determined by

$$r(\psi, P(\theta), W) = \sum_{i=1}^M P(\theta_i) \int_{\bar{X}} W(\bar{x}|\theta_i) \rho(\theta_i, \psi(\bar{x})) d\bar{x}, \quad (5)$$

where $W = \{W(\bar{x}|\theta_i), \forall i\}$, is a set of conditional pdfs.

The decision function ψ can be considered as being from an infinite set of decision functions Ψ . The decision function that yields the minimum overall risk values is regarded as the Bayes decision function. The risk associated with the Bayes decision function is then

$$r(P(\theta), W) = \min_{\psi \in \Psi} r(\psi, P(\theta), W), \quad (6)$$

the Bayes risk. The Bayes risk is the best or optimum performance, on average, that can be achieved for the identifier.

2.1 Probability of error risk

A loss function $\rho(\theta, \alpha)$ must first be specified to determine the Bayes decision function. If both Θ and A have an equal number of elements, a one-to-one mapping can be done between the sets. From this relationship, it is interpreted that by making decision α_i , the estimated state of the source $\hat{\theta}$ is θ_i . If the decision α_j is made, and the true state of the source is θ_i , then the decision is correct if $i = j$ and in error if $i \neq j$. A decision function that minimizes the average probability of error, $P[\mathcal{E}]$, should be used to avoid errors. The Hamming-distance loss function minimizes the $P[\mathcal{E}]$. This function is defined as

$$\rho(\theta_i, \alpha_j) = \begin{cases} 0 & \text{if } i = j \\ 1 & \text{if } i \neq j \end{cases}, \quad (7)$$

where the number of elements of Θ and A are equal to M .

The average risk for the Hamming-distance loss function is the average probability of error. The observation space is divided into regions, with each region associated with a certain decision. Thus, the region $\{I_i\}$ is the region of $\bar{X}$ that would result in the decision α_i being made if $\bar{x}$ was located in $\{I_i\}$. For this reason, the observation space is also referred to as the decision space. The conditional probability $W(\alpha_j|\theta_i)$ is determined by

$$W(\alpha_j|\theta_i) = \int_{I_i \cap \{I_j\}} W(\bar{x}|\theta_i) d\bar{x}, \quad (8)$$

where the integration is taken over the region of $\bar{X}$ corresponding to an α_j decision. Averaging over the source and the destination sets yields the average risk

$$r(P, W) = P[\mathcal{E}], \quad (9)$$

which is the average probability of error.

The conditional risk for the loss function is

$$r(\alpha_j|\bar{x}) = 1 - P(\theta_j|\bar{x}), \quad (10)$$

the conditional probability that decision α_j is correct is $P(\theta_j|\bar{x})$. The Bayes decision function seeks to minimize the overall risk. Thus for every value of $\bar{x}$, the decision α_i would be made if

$$P(\theta_i|\bar{x}) > P(\theta_j|\bar{x}), \forall j \neq i \quad (11)$$

In the target identification problem, the identity of the target would be estimated to be $\theta = \theta_i$. Thus, to minimize the average probability of error, the target that maximizes the *a posteriori* probability is selected as the target's identity.

The Bayes decision rule is based on maximizing the *a posteriori* probability, which can be expressed through Bayes rule as

$$\begin{aligned} &\text{Decide } \theta_i \text{ if} \\ &W(\bar{x}|\theta_i)P(\theta_i) > W(\bar{x}|\theta_j)P(\theta_j), \forall j \neq i. \end{aligned} \quad (12)$$

If the state of the source is or is assumed to be equally likely (i.e., $P(\theta_i) = \frac{1}{M}$ for $i = 1, \dots, M$), then the decision rule can be stated as

$$\begin{aligned} &\text{Decide } \theta_i \text{ if} \\ &W(\bar{x}|\theta_i) > W(\bar{x}|\theta_j), \forall j \neq i. \end{aligned} \quad (13)$$

For equally likely *a priori* probabilities and a Hamming-distance loss function, the Bayes decision rule corresponds to the most probable source state that would cause the measurement vector $\bar{x}$ to be observed.

The random vector $\bar{x}$ is composed of N random variables, where N is a positive integer. These random variables could be formed by adding a random vector to a deterministic vector. A common resulting random vector, in real-world electronics, is the multivariate normal vector. The multivariate normal vector is composed of normally distributed or Gaussian distributed random variables. It results from the vector of zero-mean random variables $\bar{n}$ being added to the deterministic vector $\bar{s}$. Each deterministic $\bar{s}$ is the consequence of a certain target identity. Thus, the observed measurement vector can be expressed as

$$\bar{x} = \bar{s}_i + \bar{n}, \quad (14)$$

where $\bar{s}_i$ is a unique vector, resulting from the identity of the target being θ_i , and $\bar{n}$ is composed of zero mean Gaussian random variables.

If the zero-mean Gaussian random variables are distributed identically, then each variable has the same variance. Furthermore, if the random variables are uncorrelated, then Gaussian random variables are statistically independent.¹ Such a random vector results from the proper sampling of the sum of a white Gaussian noise process with a deterministic time function. In such a case, the variance of vector components would correspond to the noise power spectral density in units of watts per hertz. Under these conditions, the measurement vector $\bar{x}$ is equivalent to the sum of an independently, identically distributed (iid) Gaussian random vector with zero mean and a deterministic vector associated with a given target. The mean vector of the measurement vector is the deterministic, target associated vector, $\bar{s}_i$.

For the iid additive Gaussian noise case, the state-conditional probability density function can be expressed as

$$\begin{aligned} &\text{Decide } \theta_i \text{ if} \\ &\bar{x} \cdot \bar{s}_i + c_i > \bar{x} \cdot \bar{s}_j + c_j, \forall j \neq i, \end{aligned} \quad (15)$$

where the constants c_i , $i = 1, \dots, M$, are defined as

$$c_i = \frac{|\bar{s}_i|^2}{2}, \forall i. \quad (16)$$

The above decision criterion is based on the square of the Euclidean distance between two vectors. The optimum identifier is a *minimum distance* identifier.

The decision rule of Equation 15 is equivalent to the decision rule implemented by a correlation identifier. Thus, the correlation identifier is an optimal identifier, in the Bayes sense when the loss function is the Hamming distance loss function, the measurement vector is composed of iid Gaussian random variables, and the source states are equally likely.

3. BAYES RATE-DISTORTION FUNCTION

Information theory is usually used in the design of communication systems. Information theory provides a mathematical basis for dealing with the transmission of symbols between a source and a destination. With proper

interpretation, information theory can be applied to target identification and aid in the design of such a system. Information is defined in a way that is inversely proportional to the likelihood of that event occurring. Expressed mathematically, the information value of target θ_i being present is

$$I(\theta_i) = -\log P(\theta_i) , \quad (17)$$

where the base is assumed to be 2 and the units of information are in bits.

Averaging the information over the set of target identities determines the entropy of the source, which is given by

$$H(P(\theta)) = \sum_{i=1}^M P(\theta_i) I(\theta_i) . \quad (18)$$

Entropy is the average amount of information produced per target presents which has the units bits per symbol. The source is considered a *memoryless* source when the outputs of the source are statistically independent. The maximum entropy of a memoryless source, with M symbols, can be proven² to occur when the source symbols are equally probable. Thus, for a source pdf of $P(\theta_i) = \frac{1}{M}$, $\forall \theta_i$, the entropy of the source is

$$H(P(\theta)) = \log M . \quad (19)$$

When nothing is known about the source pdf, the assumption of equiprobable symbols yields the maximum entropy source.³

Suppose that source symbols are transmitted through a channel to a destination. In order to be transmitted over this channel, the symbols may, in some manner, be encoded before the channel and decoded after the channel, as shown in Figure 2.

Figure 2. Target identifier as a communication system

Encoding and decoding of the source symbols, before and after the channel, are performed to decrease the effects of noise introduced by the channel or the particular system may only be capable of transmitting the encoded symbols and not the source symbols themselves. In the target identifier paradigm, the source output symbols are the target identities. The noiseless sensor measurement of the target can be regarded as the encoded symbols of the source. The channel is considered here to corrupt the source information by adding or introducing some form of noise. The uncertainty about what source symbol was sent is the result of noise.

An information measure also specified the relationship between to locations of the communications system. Averaging mutual information, $I(P, W)$, where W refers to the conditional pdf that describes the channel and P refers to the a priori pdf $P(\theta)$, can be determined to be

$$I(P, W) = \sum_{i=1}^M \sum_{k=1}^N P(\theta_i) W(b_k | \theta_i) \log \left\{ \frac{W(b_k | \theta_i)}{\sum_{i=1}^M P(\theta_i) W(b_k | \theta_i)} \right\} \quad (20)$$

The average mutual information may be considered to be the average *a priori* uncertainty about the true target identity, minus the average uncertainty that remains after the identifier output is specified. Thus, if the mutual information has a value of zero, then the knowledge of the destination symbols gives no clue as to what source symbol was sent. This zero value of mutual information can also be interpreted as no information about the source was transmitted through the channel. The other extreme is when the average mutual information equals the source entropy, $I(P, W) = H(P)$. For this extreme case, knowing the symbol received at the destination makes it possible to determine, without any uncertainty, which source symbol was sent. Thus, when $I(P, W) = H(P)$, all the information about the source is transmitted through the channel.

3.1 Rate distortion theory

Rate distortion theory is information theory applied to the problem of data compression. In data compression, the goal is to determine the least number of bits of information passing through the channel, on the average, that are needed to represent a source symbol if a certain level of distortion is acceptable. Rate distortion theory can be applied to target identification. Here, the data transmission problem is an extreme case. Many signatures can be mapped to one target identity.

A source symbol (target identity) is encoded for channel transmission. In the problem presented in this paper, the encoding is performed by the sensor. The decoder (target identifier) reconstructs the corrupted encoded information into an estimate of the source symbol. An estimate of the source symbol corresponds to an estimate of the target's identity. As in the communication of data, noise may be introduced into the encoded symbols that are being communicated across the channel. Noise refers to a random process that is introduced in the channel. In the model of the target identifier, the noise is introduced after the encoding of the source symbol. Noise can be introduced by the sensor, its electronics, and the medium.

The problem to be addressed is that of reconstructing the output of the source within a certain accuracy at the receiving end of the channel. To constrain the problem, the source is considered a discrete memoryless source, that is, the successive symbols generated by the source are from a finite set and are iid, (i.e., $P(a_i)$ does not depend on time). Next, it is assumed that the channel is a memoryless channel, that is, the channel produces a source symbol representation independently of the source symbol representations that have previously passed through the channel.

Discrete and continuous channels exist. A discrete channel may contain a continuous channel and some form of quantization to yield a discrete alphabet at the output. A conditional probability matrix $W(b_k|\theta_i)$ characterizes a discrete channel. A continuous channel is characterized by a set of conditional pdfs $W(b|\theta_i)$. Let's denote the set by W , in which

$$W = \{W(b|\theta_i) | i = 1, 2, \dots, M\} . \quad (21)$$

It will be assumed that a discrete channel is being used in the following equation. For that continuous channel, an integration will have to replace the summation over the destination alphabet.

To determine whether or not the required accuracy has been achieved, a quantitative measure of the distortion is needed. The distortion is that which exists between the source symbol and a representation of the source that occurs somewhere after the channel. A loss function $\rho(\theta, b)$ that places a penalty on each source-destination pair (θ, b) can be specified. This loss function assigns a nonnegative real number to all source-destination pairs.

The expected value of the loss is referred to as distortion and can be expressed as

$$d(W) = \sum_{i=1}^M \sum_{k=1}^N P(\theta_i) W(b_k|\theta_i) \rho(\theta_i, b_k) \quad (22)$$

Distortion depends on the channel according to the probabilities of $W(b_k|\theta_i)$ for a given source pdf and loss function. Thus, $d(W)$ is the distortion associated with the channel. A channel is D admissible if $d(W) \leq D$ for a given choice of source pdf $P(\theta)$ and loss function $\rho(\theta, b)$. The set of all D admissible conditional pdfs, which define different channels, is denoted by

$$\mathcal{W}_D = \{W | d(W) \leq D\} \quad (23)$$

For each conditional probability matrix defining a channel, an associated distortion value $d(W)$ exists. In addition, an average mutual information value can be associated with a channel, for a given source pdf $P(a)$. The rate-distortion function $R(P, \rho, D)$ is defined as

$$R(P, \rho, D) = \min_{W \in \mathcal{W}_D} I(P, W) \quad (24)$$

for the distortion measure $\rho(\theta, b)$. The rate-distortion function $R(P, \rho, D)$ is measured in bits per source symbol. The rate-distortion function seeks to minimize the average mutual information with respect to the channel's conditional probabilities for a given source and is always a monotonically decreasing function.⁴

Average mutual information, $I(P, W)$, is the rate of information transfer from the source to a destination. Thus, $R(P, \rho, D)$ is the minimum rate at which information about a source can be supplied to a destination in order that a destination may reproduce that information within a prescribed fidelity criteria. On the average, then, the rate-distortion function specifies the least number of information bits needed to reproduce the source symbol with a distortion of D or less.

3.2 Bayes rate distortion for target identification

The concept of applying rate-distortion theory to target identification, or pattern classification, has been proposed before. Berger⁴ observes that data compression and pattern recognition are closely related. He points out that rate-distortion theory shows how the probability of false identification varies with the number and quality of the observations.

Examining the target identification problem from an Bayesian perspective yields the model shown in Figure 1. The source is the scenario in which a target could be present and the source output or state is the identity of the target. In the model, $\Theta = \{\theta\}$ is the finite set of source states (i.e., the various target identities) that occur with *a priori* probability, $P(\theta)$, where $\theta \in \Theta$. The source state can not be known directly; instead, only a measurement vector $\vec{x}$ can be observed.

This observation vector, $\vec{x}$, ($\vec{x} \in \vec{X}$), is the continuous target signature that has been sampled after passing through the channel. The observation vector is assumed to be formed by an additive noise process. Thus, $\vec{x}$ is a vector composed of continuous random variables and is conditionally related to the source state. A conditional probability density function $W(\vec{x}|\theta)$ specifies the likelihood of receiving a certain target signature for a given target identity, θ . A decision as to the identity of the target can be made from the observation vector.

The destination alphabet set, $A = \{\alpha\}$, is composed of all possible estimated target identities, $\alpha \in A$. The set A could be identical to Θ or could have an extra "unknown identity" element in which no decision is made concerning the target identity. This paper will address only the case of $A = \Theta$. The many to one mapping of the observation space to the destination space is performed by some decision function ψ , which is taken from the set of all decision functions $\Psi = \{\psi\}$.

A loss or penalty, $\rho(\theta, \alpha)$, can be associated with each source destination pair, (θ, α) . The loss function $\rho(\theta, \alpha)$ is a mapping of the 2 space $\Theta \times A$ to the nonnegative real numbers. The Hamming loss function was defined in which the loss is one if $\theta \neq \alpha$ and zero if $\theta = \alpha$.

With a loss function specified, the average loss, or risk, of a given decision function ψ can be determined by

$$r(\psi(\vec{x}), P(\theta), W(\vec{x}|\theta)) = \sum_{\theta \in \Theta} \int_{\vec{x} \in \vec{X}} P(\theta) W(\vec{x}|\theta) \rho(\theta, \psi(\vec{x})) d\vec{x} \quad (25)$$

The Bayes decision function minimizes the risk over all decision functions. It yields the Bayes risk, which is given by

$$r(P, W) = \min_{\psi \in \Psi} r(\psi, P, W) \quad (26)$$

The Bayes risk depends on the conditional probability density function set, $W = \{W(\vec{x}|\theta)\}$ and the *a priori* probabilities $P(\theta)$. These two probability density function entities are used to calculate mutual information between the observation space and the source space. With this knowledge, the Bayes rate distortion function can be defined as

$$R(P, I) \triangleq \min_{W \in \mathcal{W}_I} I(P, W), \quad \forall I > 0. \quad (27)$$

where $\mathcal{W}_L$ is the set of all conditional probability density function sets that yield a Bayes risk less than or equal to some nonnegative real number L . This can be expressed as

$$\mathcal{W}_L \triangleq \{W : r(P, W) \leq L\}, \forall L \geq 0. \quad (28)$$

In terms of a rate-distortion function, the distortion, which is the Bayes risk, is the solution of a separate optimization problem. The Bayes risk-optimization problem is the optimization of a decisionmaking procedure. For the classical rate-distortion function, the distortion is a simple average. Thus, instead of minimizing the mutual information over a set of conditional pdf sets that are determined by an average distortion, the Bayes rate-distortion function minimizes the mutual information over a set of conditional pdf sets that are determined by an optimization problem. Clearly, the Bayes rate-distortion function is more difficult to calculate than the classical rate-distortion function.

It can be shown for the discrete-channel case⁵ and the continuous-channel case⁶ that the Bayes rate-distortion function is equal to the classical rate-distortion function between the source and destination alphabets. This is stated formally in the following.

Theorem 1 *Let $R(P, \rho, D)$ be the rate-distortion function that regards Θ and A as a hypothetical source alphabet and a destination alphabet, respectively, with the loss function $\rho(\theta, \alpha)$ adopted as the hypothetical distortion measure between the two alphabets. Then the following equality holds between the previously mentioned rate-distortion function and the Bayes rate-distortion function:*

$$R(P, L) = R(P, \rho(\theta, \alpha), L), \forall L > 0, \forall P \text{ on } \Theta. \quad (29)$$

This theorem guarantees that convexity, continuity, and the strictly decreasing property of the rate-distortion function apply to the Bayes rate-distortion function.

By Theorem 1, the computation of the Bayes rate-distortion function is reduced to that of the well defined rate-distortion function of information theory. This circumvents the need to determine the Bayes risk that determines the Bayes decision function. The loss function $\rho(\theta, \psi(\bar{F}))$, and the loss function $\rho(\theta, \alpha)$, are the same for the Bayes rate-distortion function, which evaluates mutual information between the source and observation spaces, and the rate-distortion function, which evaluates information between the source and destination spaces. Figure 3 shows the communications system block diagram interpretation of the target-identification system and the corresponding placements of the Bayes rate distortion function and its equivalent rate-distortion function.

Figure 3. Rate distortion functions placement

If the loss function is the Hamming loss function, then the distortion is the average probability of error in determining the target identity. Thus, the performance of the identifier is mathematically related to the information rate of the identification system. If the source states are (or assumed to be) equiprobable, then the rate distortion function can be expressed mathematically.⁴ Thus, for a Hamming distance distortion function and equiprobable source states, the Bayes rate distortion function can be expressed as

$$R(P, L) = R(P, \rho_H(\theta, \alpha), L) = \log(M) + L \log(L) + (1 - L) \log(1 - L) - L \log(M - 1), \quad (30)$$

for

$$0 \leq L \leq \frac{M-1}{M}.$$

The distortion, or probability of error, of Equation 30 has an upper limit of $L = \frac{M-1}{M}$. This value is the probability of error that would occur because of random selection. This distortion value corresponds to an information rate of zero. Thus, if nothing is known about the source state, a random selection is the best decision rule.

The Bayes rate-distortion function conveys the needed minimal amount of mutual information between the source and observation spaces to obtain a given identifier performance. The Bayes rate-distortion function is also used to determine whether a given system is capable of obtaining the average loss (i.e., probability of error) predicted by the Bayes rate-distortion function.

In some identification procedures, the target is interrogated more than once. Identical interrogation procedures can be repeated n times, where n is a positive integer. If a compound decision function $\tilde{\psi}^n : \tilde{\mathbf{X}}^n \rightarrow A^n$ is used to determine the identity of the target, then the sequence-average loss for the n decisions would be

$$\rho^n(\tilde{\theta}, \tilde{\psi}^n(\tilde{\mathbf{x}})) = \frac{1}{n} \sum_{i=1}^n \rho(\theta_i, \psi_i(\tilde{\mathbf{x}})), \quad (31)$$

where the n -sequence of the target signatures that were received is denoted by a matrix $\tilde{\mathbf{x}} = \tilde{x}_1 \tilde{x}_2 \cdots \tilde{x}_n \in \tilde{\mathbf{X}}^n$, and the sequences of true identities is denoted by a vector $\tilde{\theta} = \theta_1 \theta_2 \cdots \theta_n \in \Theta^n$. For target identification problems, the components of $\tilde{\theta}$ would be identical. The compound decision function is denoted by $\tilde{\psi}^n(\tilde{\mathbf{x}}) = \psi_1(\tilde{\mathbf{x}}) \psi_2(\tilde{\mathbf{x}}) \cdots \psi_n(\tilde{\mathbf{x}}) \in A^n$. Each decision of the compound decision is dependent on the sequence of target signatures.

Using the above notation, the converse decisionmaking theorem⁵ is introduced and has been proven for the continuous channel case.⁶

Theorem 2 (Converse Decisionmaking Theorem) *For every conditional probability density function set, $W : \Theta \rightarrow \tilde{\mathbf{X}}$ that is subject to the constraint $I(P, W) < R(P, L)$, there exists a positive number $\beta(P, W, L)$ such that*

$$\Pr[\rho^n(\tilde{\theta}, \tilde{\psi}^n(\tilde{\mathbf{x}})) > L] \geq \beta(P, W, L) \quad (32)$$

for every positive integer n and every n -length decisionmaking procedure $\tilde{\psi}^n : \tilde{\mathbf{X}}^n \rightarrow A^n$.

This theorem states that if the mutual information between the source and the observation spaces, for the channel in use, is smaller than the rate value of the Bayes rate-distortion curve for some distortion value, then there is always a chance that the sequence-average loss will be greater than the distortion of the point on the Bayes rate-distortion curve. This applies no matter how large n is or what compound decision function is used. A simple way of considering the theorem is that for $I(P, W) = R(P, L) - \epsilon$, where ϵ is some infinitesimal amount, then the probability of error for the identifier sequence maybe greater than L , (i.e., $P[\mathcal{E}] > L$).

The above two theorems clearly show that the performance of the Bayes target identifier is dependent on the source-observation mutual information. In fact, a knowledge of the source observation mutual information can be used to determine how well a correlation identifier will perform. The correlation identifier is a Bayes identifier for additive Gaussian noise and equally likely source probabilities. Empirically, it is anticipated that the source observation mutual information will depend on the SNR, the bandwidth, and the constellations of the different deterministic target signatures.

4. MUTUAL INFORMATION CALCULATION

For the target identifier, assumptions can be made to simplify the mutual information equation. First, the probabilities of the occurrence of target identities are equally likely when designing for the worst case condition.³ Second, white Gaussian noise is added to the deterministic sensor signature, and the signature vector can be assumed to be composed of iid random variables.

Each target identity is mapped one-to-one by deterministic means into a deterministic target signature, $\vec{s}_i$. The vectors are N dimensional, in accordance with the sampling. The finite set of deterministic target signatures, $\vec{S}$, has the same entropy as the source. The mutual information between $\vec{S}$ and $\vec{X}$ is equal to $I(P, W)$, which yields the following:

$$I(P, W) = \frac{1}{M} \sum_{i=1}^M \int_{-\infty}^{\infty} W(\vec{x}|\vec{s}_i) \log_2 \frac{MW(\vec{x}|\vec{s}_i)}{\sum_{j=1}^M W(\vec{x}|\vec{s}_j)} d\vec{x}. \quad (33)$$

The channel is considered to be an additive Gaussian white-noise (AWGN) channel. Under the constraints of sampling, each element of a vector is statistically independent of all other elements. Thus, each conditional pdf of the above mutual information equation is a product of Gaussian pdfs, with the means determined by the elements of the vector $\vec{s}$.

Algebraic manipulations can be used in reducing Equation 33 into a form that can be evaluated by numerical integration. The final form for the mutual information equation is

$$I(P, W) = \log_2 M + \frac{1}{M 2\sigma^2 \ln 2} \sum_{i=1}^M |\vec{s}_i|^2 - \frac{1}{M} \int_{\{\vec{\theta}\}} W(\vec{\theta}) \int_0^{\infty} W(\chi) \sum_{i=1}^M \log_2 \left[\sum_{j=1}^M \exp \left(\frac{\sqrt{\chi} |\vec{s}_j| \cos \phi_j}{\sigma} \right) \exp \left(\frac{2\vec{s}_i^T \vec{s}_j - |\vec{s}_i|^2}{2\sigma^2} \right) \right] d\chi d\vec{\theta}, \quad (34)$$

where $|\cdot|$ is the magnitude of a given vector, $W(\chi)$ is the Chi-square distribution with N degrees of freedom, and ϕ_j is a function of $\vec{s}_j$ and the angular vector $\vec{\theta}$ which has the pdf $W(\vec{\theta})$. Equation 34 requires the use of numerical integration to determine the value of the third term. The inside integration over the Chi-squared distribution can be performed by a standard numerical method for one-dimensional integration. For the calculations of this paper, the *alternative extended Simpson's rule* is used with the limits of integration encompassing over 99.999% of the Chi-squared distribution. The error in evaluating the integral is proportional to the inverse of the number of evaluation points to the forth power. The outside integration requires the use of an alternative method of numerical integration.

The integration over the $(N - 1)$ -dimensional angular space requires a numerical integration technique that can deal with the multidimensional aspect of the problem. An examination of Equation 34, shows that the angular dependency lies in the determination of the angular separation of the noise vector with each of the deterministic signature vectors. By using the relationship

$$\cos(\phi_j) = \frac{\vec{n}^T \vec{s}_j}{|\vec{n}| |\vec{s}_j|}, \quad (35)$$

a cartesian-noise-vector angular relationship, with each of the deterministic signature vectors, can be determined. The magnitude of the noise vector is accounted for in the integration over the Chi-squared distribution. The cartesian noise vector can be generated with identically distributed outputs of a Gaussian random variable generator. However, the dimensionality of the problem remains to be addressed.

The pdf of the angular vector $\vec{\theta}$ is uniform in the $(N - 1)$ -dimensional angular space.⁷ Thus the selection of one angular orientation is just as likely as another. The technique of Monte Carlo integration, can be used to solve the multidimensional integration considering the outside integration problem (i.e., the integration over $\{\vec{\theta}\}$) as the evaluation of the expected value of a function, if that function is the inside integral. Monte Carlo integration relies on the Weak Law of Large Numbers. By randomly selecting values of evaluation in accordance with the pdf of the values, the average of the evaluations of the function gives an estimate of the integration. Thus, by randomly generating noise vectors and then evaluating the inside integral of Equation 34, the average for those noise vectors will yield an estimate of the outside integral.

Monte Carlo integration introduces the difficulty of error proportional to the inverse of the square root of the number of evaluations. While the integration accuracy over the Chi-squared distribution is related to the inverse of the forth power of the number of evaluations, consequently the Monte Carlo integration is substantially less accurate for the same number of evaluations. Thus, the greatest error in determining mutual information will arise from the use of the Monte Carlo integration. This error can be confined to a confidence interval. This paper uses a confidence interval of 95%.

For the computational implementation of evaluating mutual information, tradeoffs in accuracy and computational time must be considered. Since the Monte Carlo integration requires an extremely higher number of evaluations to reduce error, it is assigned the majority of the computational effort.

function has a concave profile. Thus, differences between the theoretical and experimental are greatest in the center of the plane. However, the region of 0% to 20% probability of error will be of the most interest. Here the difference between theoretical and experimental is less.

6. SUMMARY

This paper has provided the basis for using information theory to aid in the design of target-identification systems. By defining the Bayes rate-distortion function and showing its relationship to the classical rate-distortion function, a theoretical bounding has been determined for a Bayesian approach to target identification. The Converse Decisionmaking theorem provides the foundation for upper bounding the performance of an identifier system under given conditions.

The Converse Decisionmaking Theorem makes it possible to bound the probability of error for a target being interrogated. Once the mutual information has been calculated for a given set of conditions and deterministic target signatures, the least possible probability of error can be determined from the Bayes rate-distortion function. A change in a condition, (e.g., sensor bandwidth, SNR), can have its effects on identifier performance determined from the Bayes rate-distortion function.

The experimental computations of this paper support the theory. Notably, the computational work was performed on platforms that could be considered rudimentary by present-day standards because the method of determining mutual information was implemented on a personal computer. Determining mutual information involves, to a degree, advanced numerical techniques; yet it is within the scope of most researchers. This tool for determining mutual information across a continuous channel should make the use of the Bayes rate-distortion function more feasible.

The tools presented in this research for applying information theory to the design of a target-identification system provide a means of determining specifications without expensive trial and error techniques. Capabilities of systems can be compared with the theoretical limits to determine the directions to pursue to increase performance.

7. REFERENCES

- [1] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*. New York, NY: McGraw-Hill, 1984.
- [2] R. W. Hamming, *Coding and Information Theory*. Englewood Cliffs, NJ: Prentice-Hall, second ed., 1986.
- [3] J. T. Tou and R. C. Gonzalez, *Pattern Recognition Principles*. Reading, MA: Addison Wesley, 1974.
- [4] T. Berger, *Rate Distortion Theory: A Mathematical Basis For Data Compression*. Englewood Cliffs, NJ: Prentice Hall, 1971.
- [5] F. Kanaya and K. Nakagawa, "On the practical implication of mutual information for statistical decisionmaking," *IEEE Trans. Information Theory*, vol. 37, pp. 1151-1156, July 1991.
- [6] S. D. Briles, *The Theory for, and Demonstration of, Information Theory Applied to Radar Target Identification*. PhD thesis, Kansas State University, June 1992.
- [7] M. G. Kendall, *The Geometry of n Dimensions*. New York, NY: Hafner, 1961.

5. RESULTS

Collected experimentally, fifteen ultra-wideband-radar signatures were obtained from three different objects at five standardized positions. The signatures were collected in such a ways as to be consider deterministic. Each signature was filtered and sampled to yield fifteen 2-GHz bandlimited measurement vectors hat are each comprised of 400 uncorrelated samples. The set of measurement vectors were simulated passing through thirteen AWGN channels of different SNRs.

Monte Carlo simulation was used to estimated the probability of error, $P[\mathcal{E}]$, for each of the channels. Using a correlation identifier, 10,000 trials were implemented for each channel. For each trial, a measurement vector was chosen randomly from a uniform distribution and had noise added to the vector. Then, the correlation identifier estimates the measurement vector's identity, which was compared to the actual measurement vector's identity. The confidence interval associated with this experiment was 95% given a maximum error of $\pm 1\%$ at $\widehat{P}[\mathcal{E}] = 50\%$.

Mutual information was computed for the measurement vector set for the thirteen channels. The Monte Carlo integration was determined to require 4000 evaluations to maintain a reasonably small confidence interval. In balancing the larger number of Monte Carlo evaluations, the Simpson's rule integration was performed with 50 evaluations. Since, Monte Carlo integration is used, the computation provides an estimated of mutual information.

By pairing the two computations, probability of error and mutual information, a point in the rate-distortion plane can be determined for each channel. The Bayes rate-distortion function of Equation 30 can also be placed in the rate-distortion plane. Figure 4 shows the rate-distortion plane with each channel point and the Bayes rate-distortion function. Dependence can be associated with either mutual information or probability of error due to the parametric

Figure 4: Rate distortion plane for 2 GHz signatures

nature of calculations involved

The experimental points in the planes are to the right of the Bayes rate distortion function for all channels, as predicted by the theory. Examining the rate distortion planes in Figure 4, it is apparent that the experimental points do not form a smooth curve in the plane. This is explained by the fact that the values of probability of error and mutual information used to determine the points are estimates. Thus, a confidence region exists around each experimental point. Visualizing such a curve would reveal a convex profile to it, whereas the Bayes rate distortion