Any Discrimination Rule Can Have an Arbitrarily Bad Probability of Error for Finite Sample Size

LUC DEVROYE

Abstract—Consider the basic discrimination problem based on a sample of size \( n \) drawn from the distribution of \((X, Y)\) on the Borel sets of \( R^d \times \{0, 1\} \). If \( 0 < R^* < \frac{1}{2} \) is a given number, and \( \phi_n \to 0 \) is an arbitrary positive sequence, then for any discrimination rule one can find a distribution for \((X, Y)\), not depending upon \( n \), with Bayes probability of error \( R^* \) such that the probability of error (\( R_n \)) of the discrimination rule is larger than \( R^* + \phi_n \) for infinitely many \( n \). We give a formal proof of this result, which is a generalization of a result by Cover [1]. Furthermore,

\[
\sup_{\text{all distributions of }} (X, Y) \text{ with } R^* = 0 \quad R_n \geq \frac{1}{2}
\]

Thus, any attempt to find a nontrivial distribution-free upper bound for \( R_n \) will fail, and any results on the rate of convergence of \( R_n \) to \( R^* \) must use assumptions about the distribution of \((X, Y)\).

Index Terms—Bayes risk, consistency, discrimination rule, distribution-free inequalities, probability of error.

I. INTRODUCTION

In this paper we will try to clarify the statement that, without assumptions on the distribution of the data, no rate of convergence (to the Bayes probability of error) can be proved for any discrimination rule.

Let \((X, Y), (X_1, Y_1), \ldots, (X_n, Y_n)\) be independent identically distributed \( R^d \times \{0, 1\} \)-valued random vectors. The discrimination problem is concerned with the estimation of \( Y \) from \( X \) and the data \((X_1, Y_1), \ldots, (X_n, Y_n)\). A discrimination rule is a sequence of Borel measurable mappings \( g_1, g_2, \ldots \), where \( g_n \) maps \( R^d \times (R^d \times \{0, 1\})^n \) to \( \{0, 1\} \). The estimate of \( Y \) is

\[
\hat{Y} = g_n(X, X_1, Y_1, \ldots, X_n, Y_n)
\]

and the associated conditional and unconditional probabilities of error are

\[
L_n = P(\hat{Y} \neq Y | X_1, Y_1, \ldots, X_n, Y_n)
\]

and

\[
R_n = E(L_n).
\]

In any case we have \( L_n \geq R^* \), where \( R^* \) is the Bayes risk (or Bayes probability of error)

\[
R^* = \inf_{g : R^d \to \{0, 1\}} P(g(X) \neq Y).
\]

A discrimination rule is said to be Bayes risk consistent if \( \lim_{n \to \infty} R_n = R^* \). Some discrimination rules are Bayes risk consistent for some distributions of \((X, Y)\) only (e.g., the linear discrimination rules). It is now known that most nonparametric discrimination rules are Bayes risk consistent for all distributions of \((X, Y)\), e.g., the \( k_n \) nearest neighbor and related rules [9], the \( k_n \) nearest neighbor rule based only on the ranks of all the projections of the \( X_j \)'s [2], [7], recursive versions of the \( k_n \) nearest neighbor rule [4], the kernel rule (or potential function method) [3], [8], recursive versions of the kernel rule [6], and partitioning rules [5].

The next question one might ask is whether there are discrimination rules for which \( R_n \) converges to \( R^* \) with a certain speed (e.g., \( R_n - R^* = O(1/n) \)) regardless of the distribution of \((X, Y)\). The answer is negative in view of Theorems 1 and 2 proved in this paper.

Theorem 1: For any discrimination rule

\[
\sup_{\text{all distributions of }} (X, Y) \text{ with } R^* = 0 \quad R_n \geq \frac{1}{2}
\]

Theorem 2: Let \( \phi_n \to 0 \) be an arbitrary positive sequence, and let \( 0 \leq R^* \leq \frac{1}{2} \) be a given number. For any discrimination rule there exists a distribution of \((X, Y)\) with Bayes risk \( R^* \) such that

\[
R_n \geq \min(R^* + \phi_n, \frac{1}{2})
\]

for infinitely many \( n \).

Theorem 1 states that no nontrivial distribution-free upper bounds for \( R_n \) exist. In particular, if \( \phi_n \) is a given positive number sequence tending to 0, and if \( R_n \leq R^* + c \phi_n \), then the constant \( c \) must necessarily depend upon the distribution of \((X, Y)\).

Theorem 2 is a bit different; it states that no inequality of the form \( R_n \leq R^* + c \phi_n \) can be obtained even if \( c \) depends upon the distribution of \((X, Y)\)!

The reason is that we can always find some distribution of \((X, Y)\) for which \( R_n \geq R^* + \sqrt{\phi_n} \) infinitely often. Theorem 2 refines a result by Cover [1]. Notice finally that Theorem 2 also applies to the \( k \)-nearest neighbor rules \((k \text{ fixed})\) for the special case that \( R^* = 0 \). Thus, for the case \( R^* = 0 \), Theorem 2 contains no information because the \( k \)-nearest neighbor rules are not Bayes risk consistent.

By carefully analyzing the proofs, we see that Theorems 1 and 2 remain valid when \( X \) has a fixed (but otherwise arbitrary) density. Thus, putting restrictions on the distribution of \( X \) alone will not suffice for the study of the rate of convergence of discrimination rules. One needs at least conditions on the regression functions \( \tau_i(x) = P(Y = i | X = x) \).

Manuscript received December 29, 1980; revised September 16, 1981. The author is with the School of Computer Science, McGill University, Montreal, P.Q., Canada.
II. PROOFS

We will prove both theorems by constructing an example in $R^1$, in which $X$ has a density with infinite support. One can easily construct similar examples in $R^2$, or examples in which $X$ is purely atomic, or examples in which $X$ has a uniform distribution on $[0, 1]$. The density of $X$ is

$$f(x) = \sum_{i=0}^{\infty} p_i I_{[i+1, i+1)}(x)$$

where $I$ is the indicator function and $(p_0, p_1, \cdots)$ is a probability vector. In other words, the density of $X$ is histogram-shaped with height $p_i$ on the interval $[i, i+1)$.

In the proof we will need a family of distributions of $(X, Y)$. This family will have a parameter $z \in [0, 1)$, and the distribution $D(z)$ is defined as follows. Let $z$ have binary expansion $0.z_0z_1z_2\cdots$, then $Y = z_X$ where $X = \text{integer part of } X$, and $X$ has density $f$. For all $D(z)$, it is clear that $R^* = 0$ because $Y$ depends upon $X$. Let us call the probability of error for fixed $z R_n(z)$. The main steps in our proof is based upon a randomization argument. Let us introduce an independent uniform $[0, 1]$ random variable $Z$ with binary expansion $0.Z_0Z_1Z_2\cdots$. Then we have the following chain of inequalities:

$$\sup_{z \in [0, 1)} R_n(z) \geq E(R_n(Z)) \ (\text{the supremum always exceeds the mean})$$

$$= P(g_n(X, X_1, Z_X, \cdots, X_n, Z_X) = Z_X)$$

$$= E(P(g_n(X, X_1, Z_X, \cdots, X_n, Z_X)

$$\geq [X \neq X_i \text{ for all } i, \text{then } Z_X \text{ is independent of } g_n(X, X_1, \cdots, Z_X) \text{ and takes the values 0 and 1 with equal probability})$$

$$= \frac{1}{2} \sum_{i=0}^{\infty} p_i (1 - p_i)^n$$

$$\geq \frac{1}{2} \left(1 - \frac{1}{2n}\right)^n \sum_{p_i < 1/(2n)} p_i$$

$$\geq \frac{1}{4} \sum_{p_i < 1/(2n)} p_i.$$

Proof of Theorem 1: Let $p_i = 1/K$, $1 \leq i < K$, and $p_i = 0$ elsewhere. Then by (1)

$$\sup_{z \in [0, 1)} R_n(z) \geq \frac{1}{2} \sum_{i=1}^{K} p_i (1 - p_i)^n \geq \frac{1}{2} \left(1 - \frac{1}{K}\right)^n$$

which is arbitrarily close to $\frac{1}{2}$ by the choice of $K$. Theorem 1 follows since $R^* = 0$ for all distributions $D(z)$, $z \in [0, 1)$.

Proof of Theorem 2: Let us first note that we can assume that $\delta_n$ is strictly monotone. If we can prove Theorem 2 for all $\delta_n = 0$ (i.e., denotes monotone convergence), then we have proved it for all sequences $\delta_n = 0$. Indeed, $\psi_n = \sup_{n \geq 0} (\phi_i : i \geq n) + 1/n \delta_n = 0$. Thus, we may find a distribution of $(X, Y)$ with Bayes probability of error $R^*$ such that $R_n \geq \min (R^* + \psi_n, \frac{1}{n}) \geq \min (R^* + \phi_n, \frac{1}{n})$ infinitely often.

We begin by showing that we can find $N$ and $(p_0, p_1, \cdots)$ such that

$$\frac{1}{n} \sum_{i=1}^{n} p_i > \phi_n, \forall n > N.$$ \hspace{1cm} (1)

The construction is rather straightforward: let

$$N = \inf (n : n \geq 1, \phi_n < \frac{1}{n}),$$

$$k_0 = 0 = k_1 = \cdots = k_{N-1} = 1.$$ \hspace{1cm} (1)

Assume that $n > N$. Given $k_{n-1}$, find $k_n > k_{n-1}$ and $p_l$, $k_{n-1} < l < k_n$ as follows:

1. all $p_l$'s are positive;
2. the $p_l$ sequence is nonincreasing (starting with $p_1$);
3. $p_l \leq 1/(2n)$;
4. $\sum_{i=k_{n-1}+1}^{k_n} p_i = 4(\phi_n - \phi_{n+1}).$

Clearly

$$\frac{1}{n} \sum_{i=1}^{n} p_i \geq \frac{1}{n} \sum_{i=k_{n-1}+1}^{k_n} p_i \geq \frac{1}{n} \sum_{i=k_{n-1}+1}^{k_{n-1}} p_i = \frac{1}{n} \cdot 4(\phi_n - \phi_{n+1}) = \phi_n, \forall n > N.$$

To make $(p_0, p_1, \cdots)$ a probability vector, note that

$$\sum_{i=1}^{\infty} p_i = 4\phi_N \leq \frac{1}{2}.$$ \hspace{1cm} (1)

Choose $p_l$ such that the sum of the $p_l$'s is one.

Consider first the case $R^* = 0$. Let

$$\bar{R}_n(z) = \sup_{m > n} R_m(z)/\phi_m.$$ \hspace{1cm} (1)

We assume first that $\bar{R}_n(z) \rightarrow 0$ for all $z \in [0, 1)$. We will reach a contradiction, and must conclude that

$$\sup_{z \in [0, 1)} \lim_{n \rightarrow \infty} R_n(z)/\phi_n > 0.$$ \hspace{1cm} (1)

Let $D_n = \{ z : z \in [0, 1), \bar{R}_n(z) \leq 1, \bar{R}_n(z) \leq 1 \}, D_n^c$-complement of $D_n$. Clearly, $P(Z \in D_n^c) \rightarrow 0$ as $n \rightarrow \infty$ by the Lebesgue dominated convergence theorem. Let $(p_1, p_2, \cdots)$ be a probability vector chosen as in the previous construction. By Fatou's lemma

$$\sup_{z \in [0, 1)} \lim_{n \rightarrow \infty} R_n(z)/\phi_n$$

which is arbitrarily close to $\frac{1}{2}$ by the choice of $K$. Theorem 1 follows since $R^* = 0$ for all distributions $D(z), z \in [0, 1)$. 

Proof of Theorem 2: Let us first note that we can assume that $\delta_n$ is strictly monotone. If we can prove Theorem 2 for all $\delta_n = 0$ (i.e., denotes monotone convergence), then we have
\[ g_n(X, X_1, Z_{X_1}, \cdots, X_n, Z_{X_n}) \neq Z_X \]
\[ = \limsup_n \phi_n^{-1} E(P(Z \in D_n, g_n(X, X_1, \cdots, Z_{X_n}) \neq Z_X \mid X, X_1, \cdots, X_n)) \]
\[ = \limsup_n \phi_n^{-1} E \left( I_{\bigcap_{j=1}^n \{ X \neq X_j \}} \sum_{i=0}^n P(Z \mid X, X_1, \cdots, X_n) \right) \]
\[ = I_{\{ g_n(X, X_1, \cdots, Z_{X_n}) \neq j \}} Z \text{ with } j \text{ forced in the } X^\text{th position} \in D_n \]
\[ = \limsup_n \phi_n^{-1} E \left( I_{\bigcap_{j=1}^n \{ X \neq X_j \}} \right) \]
\[ P(Z \text{ with } 0 \text{ forced in } X^\text{th position} \in D_n, Z \text{ with } 1 \text{ forced in } X^\text{th position} \in D_n \mid X) \]
\[ = \limsup_n (2\phi_n)^{-1} P \left( \bigcap_{j=1}^n \{ X \neq X_j \}, \mid \sum_{i=0}^n p_i (1 - p_i)^n \right) \]
\[ = \limsup_n (2\phi_n)^{-1} \left( \sum_{i=0}^\infty p_i (1 - p_i)^n \right) \left( 1 - 2P(Z \in D_n) \right) \]
\[ \geq \frac{1}{2}. \]  

By assumption, the left-hand side of (2) is 0, so we have a contradiction. Thus, there exists a \( z \in [0, 1] \) and a constant \( c > 0 \) such that \( R_n(z) \geq c \phi_n \) infinitely often. But \( \phi_n \) is arbitrary and can be replaced by \( \sqrt{\phi_n} \) and thus there exists a \( z \in [0, 1] \) and \( c > 0 \) such that \( R_n(z) \geq c \sqrt{\phi_n} \phi_n \) infinitely often.

In the case \( R^* = \frac{1}{2} \), Theorem 2 becomes trivial. We may thus assume that \( R^* \in (0, \frac{1}{2}) \). We will once again use a family of distributions \( D(z), z \in [0, 1] \). As before, \( (p_0, p_1, \cdots) \) is a probability vector, but now we set \( p_0 = 2R^* \). Let \( W \) and \( X \) be independent random variables where \( X \) has density \( f \) and \( W \) is Bernoulli, taking the values 0 and 1 with equal probability. Let

\[ Y = \begin{cases} X, & X > 0, \\ W, & X = 0. \end{cases} \]

For each \( z \) we have \( R^* = \frac{1}{2} P(X = 0) = \frac{1}{2} p_0 \), which explains our choice for \( p_0 \). Note that the sequence \( (X, Y), (X_1, Y_1), \cdots, (X_n, Y_n) \) is known if \( (X, W), (X_1, W_1), \cdots, (X_n, W_n) \) is known. Here the couples \( (X_i, W_i) \) are independent and distributed as \( (X, W) \). The randomization is performed as before, i.e., by introducing a uniform \([0, 1]\) random variable \( Z \), independent of \( (X, Y), (X_1, Y_1), \cdots, (X_n, Y_n) \). Arguing as in (1), we have

\[ \sup_{z \in [0, 1]} R_n(z) = E(R_n(Z)) \]
\[ = E(P(g_n(X, X_1, Y_1, \cdots, X_n, Y_n) \neq Y \mid X, X_1, \cdots, X_n)) \]
\[ \geq E(\ell_{\{X = 0\}} \frac{1}{2} + \ell_{\{X \neq 0\}} \frac{1}{2} \mid \bigcap_{j=1}^n \{ X \neq X_j \} \frac{1}{2}) \]

(3)

The remainder of the proof can be mimicked from the case \( R^* = 0 \). Define \( N = \inf(n: n \geq 1, \phi_n \leq (1 - 2R^*)/8) \); \( k_n = k_1 = \cdots = k_{N-1} = 1 \). Then construct \((p_2, p_3, \cdots)\) as before (nondecreasing from \( p_2 \) onwards) so that

\[ \sup_z R_n(z) > R^* + \frac{1}{4} \sum_{i=1}^\infty p_i (1 - p_i)^n. \]

Define \( p_1 = 1 - 2R^* - 4\phi_N \), and note that \( p_1 > 0 \), and that \( p_0 + p_1 + p_2 + \cdots = 2R^* + (1 - 2R^* - 4\phi_N) + 4\phi_N = 1 \).

We continue the proof as for the case \( R^* = 0 \). Let \( R_n(z) = \sup_{m \geq n} (R_m(z) - R^\ast) \omega_m \), and let \( (p_0, p_1, \cdots) \) be a probability vector chosen as indicated above. Assume first that \( R_n(z) \to 0 \) for all \( z \in [0, 1] \); we will once again reach a contradiction, and must then conclude that for some \( z \in [0, 1] \), \( \sup_n \phi_n (R_n(z) - R^\ast) \phi_n > 0 \). This would conclude the proof of the theorem.

Under the said assumption, we have

\[ \sup_{z \in [0, 1]} \limsup_n \sup \left( R_n(z) - R^\ast \right) \phi_n \]
\[ \geq \frac{1}{2}. \]  

REFERENCES

A Model for Radar Images and Its Application to Adaptive Digital Filtering of Multiplicative Noise

VICTOR S. FROST, STUDENT MEMBER, IEEE, JOSEPHINE ABBOTT STILES, STUDENT MEMBER, IEEE, K. S. SHANMUGAN, SENIOR MEMBER, IEEE, AND JULIAN C. HOLTZMAN, MEMBER, IEEE

Abstract—Standard image processing techniques which are used to enhance noncoherent, optically produced images are not applicable to radar images due to the coherent nature of the radar imaging process. A model for the radar imaging process is derived in this paper and a method for smoothing noisy radial images is also presented.

The imaging model shows that the radar image is corrupted by multiplicative noise. The model leads to the functional form of an optimum (minimum MSE) filter for smoothing radial images. By using locally estimated parameter values the filter is made adaptive so that it provides minimum MSE estimates inside homogeneous areas of an image while preserving the edge structure. It is shown that the filter can be easily implemented in the spatial domain and is computationally efficient. The performance of the adaptive filter is compared (qualitatively and quantitatively) with several standard filters using real and simulated radar images.

Index Terms—Adaptive filtering, image enhancement, minimum mean square error (MMSE), multiplicative noise, radar image modeling, radar image processing, speckle reduction, synthetic aperture radar (SAR).

I. INTRODUCTION

A LARGE number of image restoration and enhancement techniques have been proposed in recent years for removing a variety of degradations in recorded images of objects and scenes. These degradations result from the nonideal nature of practical imaging systems. The design of optimum image restoration and enhancement techniques requires a mathematical model of the imaging process. This paper presents a model for the noise in radar images and uses the model to develop an adaptive algorithm to smooth noisy nonstationary images.

Imaging radars, specifically the synthetic aperture radar (SAR), are beginning to make use of the digital techniques, and digitally correlated SAR images are now becoming available. However, optimum techniques for digitally processing radar images are not fully developed due to a lack of understanding of the properties of radar images from a digital image processing perspective. Thus, there is an important need for developing statistical models for radar noise and for using them in deriving appropriate algorithms for processing radar images.

This paper presents a model and a model-based image enhancement technique which is specifically designed for active microwave sensors utilizing coherent imaging techniques. The model portrays the observed radar image as corrupted by multiplicative-convolved noise. That is, the desired information, the terrain backscatter, is multiplied by a stationary random process which represents the effects of coherent fading [1]-[3]. The product signal is then processed (convolved) with the point spread function of the radar system to produce the observed image.

This model can be applied to the design of digital image enhancement algorithms through several approaches. The procedure used here was to develop a minimum mean square error (MMSE) filter to estimate the terrain backscatter from the