The Annals of Statistics,
1983, Vol. 11, No. 4, 1142–1155

OPTIMAL RATES OF CONVERGENCE TO BAYES RISK IN NONPARAMETRIC DISCRIMINATION

BY JAMES STEPHEN MARRON

University of North Carolina

Consider the multiclassification (discrimination) problem with known prior probabilities and a multi-dimensional vector of observations. Assume the underlying densities corresponding to the various classes are unknown but a training sample of size N is available from each class. Rates of convergence to Bayes risk are investigated under smoothness conditions on the underlying densities of the type often seen in nonparametric density estimation. These rates can be drastically affected by a small change in the prior probabilities, so the error criterion used here is Bayes risk averaged (uniformly) over all prior probabilities. Then it is shown that a certain rate, N^{-r}, is optimal in the sense that no rule can do better (uniformly over the class of smooth densities) and a rule is exhibited which does that well. The optimal value of r depends on the smoothness of the distributions and the dimensionality of the observations in the same way as for nonparametric density estimation with integrated square error loss.

1. Introduction. The classification or discrimination problem arises whenever one wants to assign an object to one of a finite number of classes based on a vector of d measurements. More precisely, let $f_1, \cdots, f_K$ be probability densities (with respect to Lebesgue measure) on $\mathbb{R}^d$. Select one of these at random, where prior probability π_k is put on f_k, $k = 1, \cdots, K$. Define the random variable θ to be the index of the chosen density. The classification (or discrimination) problem is to guess the value of θ, using an observation $\mathbf{X}$ from f_θ.

For notational convenience, let $\mathbf{f} = (f_1, \cdots, f_K)$ and let $\boldsymbol{\pi} = (\pi_1, \cdots, \pi_K)$. The entries of $\boldsymbol{\pi}$ are nonnegative and sum to 1, so $\boldsymbol{\pi}$ is an element of $\mathscr{S}_K$, the unit simplex in $\mathbb{R}^K$.

If both $\boldsymbol{\pi}$ and $\mathbf{f}$ are known, then it is simple to compute the best, or in other words, Bayes classification rule (see (1.5)). In this paper it is assumed that $\boldsymbol{\pi}$ is known and $\mathbf{f}$ is unknown. This assumption is not at all restrictive, because in the case of unknown $\boldsymbol{\pi}$, the usual estimates of $\boldsymbol{\pi}$ converge much faster than estimates of $\mathbf{f}$. The reason for this assumption is technical convenience in formulating the theorems.

It is assumed that for each $N \in \mathbb{Z}^+$, there is a "training sample," Z^N, which consists of a sample of size N from each of $f_1, \cdots, f_K$. For $k = 1, \cdots, K$, let $\mathbf{X}^{k1}, \cdots, \mathbf{X}^{kN}$ denote the sample from f_k. Further assume that these K samples are independent of each other, and that Z^N is independent of $\mathbf{X}$ and θ.

In this setting, any classification rule may depend on the observed value of $\mathbf{X}$, on the prior probability vector $\boldsymbol{\pi}$, and on Z^N. Hence it may be thought of as a measurable function

$$\hat{\theta}_N: \mathbb{R}^d \times \mathscr{S}_k \times (\mathbb{R}^d)^{NK} \to \{1, \cdots, K\}.$$

It is useful to consider the problem from a decision theoretic viewpoint. Both the action space and the parameter space are the set $\{1, \cdots, K\}$. An arbitrary loss function L is a real valued function on $\{1, \cdots, K\} \times \{1, \cdots, K\}$; $L(i, j)$ is the loss when one guesses i, but $\theta = j$. The $L(i, j)$ are allowed to be different to quantify any feelings the experimenter may have about one type of mistake being worse than another. For example, in the diagnosis of disease, it can be worse to classify a sick person as healthy, than to make the other error.

Received October 1982; revised May 1983.
AMS 1980 *subject classifications*. 62H30, 62G20.
Key words and phrases. Nonparametric classification, discrimination, optimal rates.

The only assumption needed about L is

$$\max_i L(i, i) < \min_{i \neq j} L(i, j).$$

It will be convenient to define

$$\underline{L} = \min_{i \neq j} L(i, j) - \max_i L(i, i) > 0 \tag{1.1}$$

and

$$\bar{L} = \max_{i,j} | L(i, j) |. \tag{1.2}$$

The loss function appearing most often in the literature is 0-1 loss, where for $i = 1, \cdots, K$, $L(i, i) = 0$, and for $i \neq j$, $L(i, j) = 1$.

Next, for $\mathbf{x} \in \mathbb{R}^d$ and $\pi \in \mathscr{S}_K$, note that the posterior probability of the class i, $i = 1, \cdots, K$, is given by

$$P_{\mathbf{f}}[\theta = i \mid \mathbf{X} = \mathbf{x}] = \frac{\pi_i f_i(\mathbf{x})}{\sum_j \pi_j f_j(\mathbf{x})}. \tag{1.3}$$

For $k = 1, \cdots, K$, the expected value of $L(k, \theta)$ with respect to this posterior distribution is given by

$$R_{\mathbf{f}}(k, \mathbf{x}, \pi) = \sum_i L(k, i) P_{\mathbf{f}}[\theta = i \mid \mathbf{X} = \mathbf{x}]. \tag{1.4}$$

Throughout this paper $R_{\mathbf{f}}(\hat{\theta}_N(\mathbf{x}, \pi, Z^N), \mathbf{x}, \pi)$ will be denoted $R_{\mathbf{f}}(\hat{\theta}, \mathbf{x}, \pi)$. The risk function, $R_{\mathbf{f}}$, can now be interpreted as: expected loss where expectation is taken conditioned on Z^N and on the event $\mathbf{X} = \mathbf{x}$. Thus $R_{\mathbf{f}}(\hat{\theta}_N, \mathbf{x}, \pi)$ is a random variable which gets its randomness from the dependence of $\hat{\theta}_N$ on Z^N.

The form of the rules which are Bayes with respect to $R_{\mathbf{f}}$ will now be given. For each $\mathbf{x} \in \mathbb{R}^d$ and each $\pi \in \mathscr{S}_K$, pick $\hat{\theta}_B \in \{1, \cdots, K\}$ so that:

$$R_{\mathbf{f}}(\hat{\theta}_B, \mathbf{x}, \pi) = \min_{k=1,\cdots,K} R_{\mathbf{f}}(k, \mathbf{x}, \pi). \tag{1.5}$$

When the minimum is not unique, the manner in which ties are broken is irrelevant, but for definiteness take $\hat{\theta}_B$ as small as possible. This defines a classification rule, $\hat{\theta}_B(\mathbf{x}, \pi)$ which is independent of Z^N, but unfortunately depends on the unknown $\mathbf{f}$.

In the particular case of 0-1 loss, it is easy to compute a Bayes rule, since $R_{\mathbf{f}}(k, \mathbf{x}, \pi)$ is a linear combination of the posterior probabilities, where the kth coefficient is 0 and all the rest are 1. Thus the Bayes rule is to choose that k which maximizes the posterior probability. Note that this is the "intuitive solution" to the classification problem.

The first thing one might hope to find in this setting, is a $\hat{\theta}_N$ that behaves, at least asymptotically as $N \to \infty$, like $\hat{\theta}_B$. In the literature, there are several papers which propose classification rules, $\hat{\theta}_N$, which are "Bayes Risk Consistent," in the sense that, under mild conditions on $\mathbf{f}$, in some mode of convergence,

$$\lim_{N \to \infty} R_{\mathbf{f}}(\hat{\theta}_N, \mathbf{x}, \pi) = R_{\mathbf{f}}(\hat{\theta}_B, \mathbf{x}, \pi).$$

Among these are: Fix and Hodges (1951), Das Gupta (1964), Quesenberry and Loftsgaarden (1965), Van Ryzin (1966), Glick (1972), Devroye and Wagner (1977), and Greblicki (1978). For results with essentially no assumptions on the underlying distributions (including the existence of densities), see Stone (1977), Devroye and Wagner (1980), and Gordon and Olshen (1978).

Now with several such rules, the next thing to look for is some means of comparing them. This paper takes a step in this direction by considering rates of convergence in a manner similar to that found in nonparametric density estimation. In that field one can find two types of convergence rate results.

The first is the "achievability" type of result, in which an estimate is proposed, and it is shown that in some norm, the error goes to 0 at the rate N^{-r} (or sometimes $(N^{-1} \log N)^r$), for some $r > 0$, which depends on the "smoothness" of the true density and

the dimension of the sample space. Results of this type are too numerous to list here, but surveys can be found in Wegman (1972a, b), Tartar and Kronmal (1976), and Wertz (1978). An elegant result, wherein achievability is shown for many "different" density estimators in a single theorem, is in Walter and Blum (1979).

The second is the "bound" type of result, which shows that, uniformly over the class of "smooth" densities, the norm of the error can go down no faster than N^{-r} (or $(N^{-1} \log N)^r$), regardless of the estimator. This type of result can be found in: Farrell (1972), Wahba (1975), Khasminskii (1978), Bretagnolle and Huber (1979), Müller and Gasser (1979), and Stone (1980).

When the achievable rate is the same as the bound rate, then that rate is called "optimal," and any estimator that achieves it is, in this sense, optimal.

This paper presents both achievability (see Theorem 1) and bound (see Theorem 2) results for convergence to Bayes risk in the classification problem. The optimal rate turns out to be the same as that for density estimation with mean square error. The optimal classification rule is that which has been studied, in different forms, by many previous authors. The basic idea is to use a good density estimator to estimate the posterior probabilities, and then form an "estimated Bayes rule" based on these.

To implement this, one needs a density estimate which achieves the optimal rate for density estimators. Unfortunately the literature does not contain a result of quite the generality required here. Hence, the needed result is included in this paper (see Theorem 3).

To define the mode of convergence used in this paper, first fix a compact set $\mathscr{C} \subset \mathbb{R}^d$, which has nonempty interior. Then the mode is convegence in probability of

$$\int_{\mathscr{S}_K} \int_{\mathscr{C}} [R_f(\hat{\theta}_N, \mathbf{x}, \pi) - R_f(\hat{\theta}_B, \mathbf{x}, \pi)]\, d\mathbf{x}\, d\pi.$$

Absolute values are not required because, for $k = 1, \cdots, K$, and for each $\mathbf{x} \in \mathbb{R}^d$ and each $\pi \in \mathscr{S}_K$,

$$R_f(k, \mathbf{x}, \pi) \geq R_f(\hat{\theta}_B, \mathbf{x}, \pi).$$

The reason for integrating, with respect to $\mathbf{x}$, only over the compact set $\mathscr{C}$, instead of over all of $\mathbb{R}^d$, will be given in Section 2. The reason for the integration with respect to π will be discussed in detail in Section 3. Basically, it is that the rate of convergence of the integrand can be much slower than the "natural rate" for a very small set of π, so this effect is "averaged out" by integration.

In order to define what is meant by "smoothness," more notation is needed. Let $\alpha = (\alpha_1, \cdots, \alpha_d)$ where each α_i is a nonnegative integer. Also, let $|\alpha| = \sum_{i=1}^d \alpha_i$, and define the partial derivative operator:

$$\mathrm{D}_\alpha = \frac{\partial^{|\alpha|}}{\partial x_1^{\alpha_1} \cdots \partial x_d^{\alpha_d}}. \tag{1.6}$$

Given $\mathbf{x} = (x_1, \cdots, x_d)$ define the usual Euclidean norm,

$$\|\mathbf{x}\| = (x_1^2 + \cdots + x_d^2)^{1/2}.$$

Next fix a constant $M > 1$, a nonnegative integer m and a constant $\beta \in (0, 1]$; also set $p = m + \beta$. For $p \notin \mathbb{Z}$, m is the greatest integer in p and β is the fractional part.

Now let $\mathscr{F}_1$ denote the class of probability densities, f, on $\mathbb{R}^d$, such that:

i) $f \leq M$ on $\mathbb{R}^d$;
ii) $f \geq 1/M$ on $\mathscr{C}$;
iii) for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$, and all $|\alpha| = m$

$$|\mathrm{D}_\alpha f(\mathbf{x}) - \mathrm{D}_\alpha f(\mathbf{y})| \leq M \|\mathbf{x} - \mathbf{y}\|^\beta.$$

Condition iii) is the "smoothness" condition. In the case $p = 2$, this condition is slightly more general than the bounded second derivative used by Rosenblatt (1956) and many others.

Note, that, for condition ii) to be satisfied by any probability density, it must be assumed that M is larger than the d-dimensional volume of $\mathscr{C}$. From here on assume M is large enough so that $\mathscr{F}_1$ contains infinitely many members.

While the same M is used in i), ii), and iii) here, this is not needed for the results in this paper, but is only done for simplicity. With this in mind, condition i) is redundant, since the boundedness of f is a consequence of condition iii). Condition i) is included because the boundedness of f is required at many points in the proofs that follow.

Next recall the notation $\mathbf{f} = (f_1, \cdots, f_K)$. It will be convenient to let $\mathscr{F}$ denote the K-fold Cartesian product of $\mathscr{F}_1$.

2. Main theorems. The main result of this paper is that the optimal rate of Bayes risk convergence is N^{-r}, where $r = 2p/(2p + d)$. This is shown by the following theorems:

THEOREM 1. *There is a constant $c_1 > 0$ and a classification rule $\hat{\theta}_N(\mathbf{x}, \boldsymbol{\pi}, Z^N)$, so that,*

$$\lim_{N\to\infty}\sup_{\mathbf{f}\in\mathscr{F}} P_{\mathbf{f}}\left[\int_{\mathscr{S}_K}\int_{\mathscr{C}} [R_{\mathbf{f}}(\hat{\theta}_N, \mathbf{x}, \boldsymbol{\pi}) - R_{\mathbf{f}}(\hat{\theta}_B, \mathbf{x}, \boldsymbol{\pi})]\, d\mathbf{x}\, d\boldsymbol{\pi} > c_1 N^{-r}\right] = 0.$$

THEOREM 2. *There is a constant $c_2 > 0$, so that, for any classification rule $\hat{\theta}_N$,*

$$\lim_{N\to\infty}\sup_{\mathbf{f}\in\mathscr{F}} P_{\mathbf{f}}\left[\int_{\mathscr{S}_K}\int_{\mathscr{C}} [R_{\mathbf{f}}(\hat{\theta}_N, \mathbf{x}, \boldsymbol{\pi}) - R_{\mathbf{f}}(\hat{\theta}_B, \mathbf{x}, \boldsymbol{\pi})]\, d\mathbf{x}\, d\boldsymbol{\pi} > c_2 N^{-r}\right] = 1.$$

The proof of Theorem 1 is given in Section 4. To save space, the proof of the main lemma is given only in the special case $K = 2$ with 0-1 loss. This case contains the main ideas of the proof of the general case, which may be found in Section 6 of Marron (1982).

The proof of Theorem 2 is given in Section 5. In that section, the proof of Lemmas 5.1 and 5.3 are omitted. The proof of Lemma 5.1 is straightforward and can be found in Section 7.1 of Marron (1982). The proof of Lemma 5.3 is quite long and is omitted because similar techniques have been employed in Stone (1982), but it can also be found in Marron (1982). As above, the main ideas of the proof of Lemma 5.2 can most easily be seen in the case $K = 2$, so only that case is treated here.

REMARK 2.1. A careful inspection of the proof of Theorem 1 shows that the error criterion is bounded by a sum of quantities of the form

$$\int_{\mathscr{C}} \frac{(\hat{f}(\mathbf{x}) - f(\mathbf{x}))^2}{f(\mathbf{x})}\, d\mathbf{x},$$

where $\hat{f}(\mathbf{x})$ is an estimate of $f(\mathbf{x})$. The $f(\mathbf{x})$ in the denominator is not apparent at the end of the proof because it was replaced by $1/M$, which can only be done for $\mathbf{x} \in \mathscr{C}$. So if $\mathscr{C}$ is replaced by $\mathbb{R}^d$, then a density estimation convergence result for this error criterion is required. It is conjectured that the rate may be different from what is needed here. Of course, the compactness of $\mathscr{C}$ is not necessary for the bound part (Theorem 2).

REMARK 2.2. It follows from the proof of Theorem 2 that in the statement of the theorem, the supremum need not be taken over the entire class $\mathscr{F}$. As with bound results in density estimation, we need only pick one element $\mathbf{f} \in \mathscr{F}$, and then, for each $N \in \mathbb{Z}^+$, consider only a finite number of small perturbations. Here it turns out, only f_1 needs to be perturbed and the rest of $\mathbf{f} = (f_1, \cdots, f_K)$ may be left fixed.

REMARK 2.3. The error criterion can easily be changed by inserting a weight function into the integrand. A natural choice of weight function is the marginal density of $\mathbf{X}$. Since $\mathbf{f} \in \mathscr{F}$, this marginal density is bounded above and below on $\mathscr{L}$, so both rate of convergence results would remain the same. In the commonly considered case of f supported on $\mathscr{L}$, the integral with respect to $\mathbf{x}$ gives expected value.

REMARK 2.4. Of course the usual warnings regarding asymptotic results apply here. In particular, in the case of p and d even moderately large, the sample size N will have to be very large for the asymptotics to "take effect." However, the main point of the theorems of this paper is a strong indication the classification rules based on density estimators perform at least as well as any other possible procedure. This seems to provide a convincing answer to the question: "Why study density estimation?"

As remarked in Section 1, for Theorem 1, the achievability result, a density estimation result is required. To simplify the notation, given $f \in \mathscr{F}_1$, suppose that for each $N \in \mathbb{Z}^+$, there is a sample $X^1, \cdots, X^N$ from f. Then an estimate of $f(\mathbf{x})$ will be denoted by a measurable function $\hat{f}_N(\mathbf{x}, \mathbf{x}^1, \cdots \mathbf{x}^N)$. The result is

THEOREM 3. *There is a constant $c_3 > 0$ and a density estimator $\hat{f}_N(x, \mathbf{X}^1, \cdots, \mathbf{X}^N)$ so that, when $r = 2p/(2p + d)$,*

$$\lim_{N\to\infty} \sup_{f\in\mathscr{F}_1} P_f\left[\int_{\mathscr{L}} [\hat{f}_N(\mathbf{x}) - f(\mathbf{x})]^2 \, d\mathbf{x} > c_3 N^{-r}\right] = 0.$$

To save space, the proof of Theorem 3 is omitted here. It is essentially a generalization of a result of Epanechnikov (1969), using some techniques that can be found in Stone (1982). Details are in Section 8 of Marron (1982).

3. Motivation for averaging over π. In this section, to show the need for averaging $R_{\mathbf{f}}(\hat{\theta}_N, \mathbf{x}, \boldsymbol{\pi}) - R_{\mathbf{f}}(\hat{\theta}_B, \mathbf{x}, \boldsymbol{\pi})$ over $\boldsymbol{\pi} \in \mathscr{S}_K$, a simple example is heuristically considered. Specifically, assume $d = 1$, $K = 2$, $p = 2$, and L is 0-1 loss.

So now there are just two densities, f_1 and f_2, on the real line, which are smooth in the sense of (nearly) having bounded second derivatives. It follows from (1.3) and (1.4) that, for $x \in \mathbb{R}$, $\boldsymbol{\pi} \in \mathscr{S}_2$,

$$R_f(1, x, \pi) = P_f[\theta = 2 \mid X = x] = \frac{\pi_2 f_2(x)}{\pi_1 f_1(x) + \pi_2 f_2(x)}$$

and

$$R_f(2, x, \pi) = P_f[\theta = 1 \mid X = x] = \frac{\pi_1 f_1(x)}{\pi_1 f_1(x) + \pi_2 f_2(x)}.$$

By conditions i) and ii) in the definition of $\mathscr{F}_1$, the denominators of the above fractions are bounded above and below, so they will not affect convergence rates. Hence, for $\boldsymbol{\pi} \in \mathscr{S}_2$,

$$\int_{\mathscr{L}} [R_{\mathbf{f}}(\hat{\theta}_N, x, \pi) - R_{\mathbf{f}}(\hat{\theta}_B, x, \pi)] \, dx \sim \int_{\mathscr{L}} [\pi_{\hat{\theta}_B} f_{\hat{\theta}_B}(x) - \pi_{\hat{\theta}_N} f_{\hat{\theta}_N}(x)] \, dx. \tag{3.1}$$

Note that the integrand is 0 when $\hat{\theta}_N = \hat{\theta}_B$, and otherwise it is $| \pi_1 f_1(x) - \pi_2 f_2(x) |$, the "weighted difference" of the densities.

Now for $x \in \mathscr{L}$, $\boldsymbol{\pi} \in \mathscr{S}_2$, by the achievability results from density estimation, there are

estimates of $\pi_1 f_1(x)$ and $\pi_2 f_2(x)$ which have error of the order (as $N \to \infty$)$N^{-2/5}$. The bound results from density estimation imply that no estimator can do better. Hence, heuristically speaking, the "information" available about $\pi_1 f_1(x)$ and $\pi_2 f_2(x)$ is "accurate to the order $N^{-2/5}$."

So, for large N, and for those $x \in \mathscr{C}$, which satisfy

$$| \pi_1 f_1(x) - \pi_2 f_2(x) | > N^{-2/5}, \tag{3.2}$$

there is enough information available so that $\hat{\theta}_N(x, \boldsymbol{\pi})$ is (usually) the same as $\hat{\theta}_B(x, \boldsymbol{\pi})$, hence the contribution to the integrand in (3.1) is 0. For the rest of the $x \in \mathscr{C}$, it is expected that sometimes $\hat{\theta}_N = \hat{\theta}_B$ and sometimes $\hat{\theta}_N \neq \hat{\theta}_B$, however the probability of some contribution is bounded above 0. From (3.2), the $x \in \mathscr{C}$ which may contribute to the integrand of (3.1) are in neighborhoods of the zeroes of $\pi_1 f_1(x) - \pi_2 f_2(x)$.

Now suppose f_1, f_2, and $\boldsymbol{\pi}$ are such that $\pi_1 f_1(x) - \pi_2 f_2(x)$ has a zero of the first order (i.e.: nonzero first derivative) at $x = 0$. Then the x near 0, which may contribute to the integrand of (3.1), constitute a neighborhood whose diameter is of the order $n^{-2/5}$. Thus, since each x contributes with positive probability, the integral in (3.1) is expected to be of the order $N^{-4/5}$.

Unfortunately, the above analysis depends heavily on the fact that the zero of $\pi_1 f_1(x) - \pi_2 f_2(x)$ is of the first order. Suppose, instead, that for some $j \in \mathbb{Z}^+$, on some neighborhood of 0, $\pi_1 f_1(x) = \pi_2 f_2(x) = x^j$. Then an argument similar to the above shows that the integral of (3.1) is expected to be of the order $N^{-2/5-2/5j}$. Thus the rate of convergence depends not only on the smoothness and the dimension, but also on the order of the zeroes of $\pi_1 f_1(x) - \pi_2 f_2(x)$.

To verify these heuristics, one might be tempted to formulate a theorem that takes the order of the zeroes of $\pi_1 f_1(x) - \pi_2 f_2(x)$ into account. But note that even in the present simple case, the formulation is very awkward, and for $d > 1$, $K > 2$, the difficulties become prohibitive.

A way around these difficulties is to take the conservative approach of considering only the worst possible case (i.e.: j arbitrarily large). This viewpoint may be used to understand the rates obtained in the achievability results of van Ryzin (1966), Györfi (1978), Györfi (1981), Greblicki (1981) and Greblicki and Pawlak (1982). Some care is required in interpreting the rates obtained in the last two of these papers, because the authors combine dimensionality and smoothness in such a way that the results appear to be independent of the dimensionality. Although their smoothness conditions are not precisely comparable with those of this paper, the closest connection seems to be that the quantity p in the present paper corresponds to $m \cdot p$ in Greblicki (1981) and to $r \cdot d$ in Greblicki and Pawlak (1982).

To see why the conservative rates obtained by the above authors are not representative of the situation, return to the above example. Note that, for any j, if $\boldsymbol{\pi}$ is changed by a small amount, then the zero of $\pi_1 f_1(x) - \pi_2 f_2(x)$ is of the first order, and the rate again becomes $N^{-4/5}$. Thus it is apparent that the pathologies of higher order zeroes occur only on a set of $\boldsymbol{\pi}$ which have Lebesgue measure 0. Hence the rate $N^{-4/5}$ seems natural "almost everywhere with respect to $\boldsymbol{\pi}$."

With this in mind, a reasonable approach is that taken by Van Houwelingen (1980). In that paper it is assumed that the underlying densities $\mathbf{f}$ are known and that $\boldsymbol{\pi}$ is unknown (the reverse is assumed here). By the above considerations, and the fact that estimates of $\boldsymbol{\pi}$ are multinomial in character, it is not surprising that Van Houwelingen reports that an integral similar to (3.1) converges at the rate N^{-1} for almost all $\boldsymbol{\pi}$ and is bounded by $N^{-1/2}$ for all $\boldsymbol{\pi}$.

This approach is not taken in the present paper because, while achievability results (such as Theorem 1) provide a good indication of what is happening, bound type results (such as Theorem 2) are very difficult to formulate. Instead the pathological set of $\boldsymbol{\pi}$ is

nullified by averaging, or more precisely,

$$\int_{\mathscr{S}_K} \int_{\mathscr{C}} R_{\mathbf{f}}(\hat{\theta}_N, \mathbf{x}, \boldsymbol{\pi}) - R_{\mathbf{f}}(\hat{\theta}_B, \mathbf{x}, \boldsymbol{\pi})\, d\mathbf{x}\, d\boldsymbol{\pi}$$

is used as the error criterion. It should be noted that the choice of Lebesgue measure for averaging over $\mathscr{S}_K$ is not vital to the results of this paper. An inspection of the proofs shows that any measure which is mutually absolutely continuous with respect to Lebesgue measure and has bounded Radon-Nikodym derivative will suffice.

4. Proof of Theorem 1. In the course of this proof, it will be convenient to introduce a number of positive constants. These will be denoted by B_i, where $i \in \mathbb{Z}^+$. In each case, these are independent of $\mathbf{f}$, $\mathbf{x}$, $\boldsymbol{\pi}$, N, and Z^N, however they may depend on any or all of d, K, L, $\mathscr{C}$, m, β, and p.

It will also be conveneint to define,

$$\mathscr{E}_N = \int_{\mathscr{S}_K} \int_{\mathscr{C}} [R_{\mathbf{f}}(\hat{\theta}_N, \mathbf{x}, \boldsymbol{\pi}) - R_{\mathbf{f}}(\hat{\theta}_B, \mathbf{x}, \boldsymbol{\pi})]\, d\mathbf{x}\, d\boldsymbol{\pi}. \tag{4.1}$$

Now given $\mathbf{f} = (f_1, \cdots, f_K)$, recall, for $k = 1, \cdots, K$, that Z^N contains a sample $\mathbf{X}^{k1}, \cdots, \mathbf{X}^{kN}$ from f_k. Use it to construct a density estimate $\hat{f}_k(\mathbf{x}, \mathbf{X}^{k1}, \cdots, \mathbf{X}^{kN})$ with the same convergence property as that in Theorem 3. Then let $\hat{\mathbf{f}} = (\hat{f}_1, \cdots, \hat{f}_K)$.

Next let $\hat{\theta}_N(\mathbf{x}, \boldsymbol{\pi}, Z^N)$ be the "estimated Bayes rule" given by: for $\mathbf{x} \in \mathbb{R}^d$ and $\boldsymbol{\pi} \in \mathscr{S}_K$, take $\hat{\theta}_N$ in $\{1, \cdots, K\}$ so that

$$R_{\hat{\mathbf{f}}}(\hat{\theta}_N, \mathbf{x}, \boldsymbol{\pi}) = \min_{k=1,\cdots,K} R_{\hat{f}}(k, \mathbf{x}, \boldsymbol{\pi}). \tag{4.2}$$

Now using (1.3) and (1.4), note that

$$R_{\mathbf{f}}(\hat{\theta}_N, \mathbf{x}, \boldsymbol{\pi}) - R_{\mathbf{f}}(\hat{\theta}_B, \mathbf{x}, \boldsymbol{\pi}) = \frac{\sum_k (L(\hat{\theta}_N, k) - L(\hat{\theta}_B, k))\pi_k f_k(\mathbf{x})}{\sum_l \pi_l f_l(\mathbf{x})}.$$

But from (ii) in the definition of the class $\mathscr{F}_1$, for $\mathbf{x} \in \mathscr{C}$

$$\sum_l \pi_l f_l(\mathbf{x}) \geq \sum_l \pi_l (1/M) = 1/M.$$

Thus, since the integrand is nonnegative, Fubini's theorem applied to (4.1) gives

$$\mathscr{E}_N \leq M \int_{\mathscr{C}} \int_{\mathscr{S}_K} \sum_k (L(\hat{\theta}_N, k) - L(\hat{\theta}_B, k))\pi_k f_k(\mathbf{x})\, d\boldsymbol{\pi}\, d\mathbf{x}. \tag{4.3}$$

Next, for each $\mathbf{x} \in \mathscr{C}$, a bound will be obtained for the inside integral. For $i, j = 1, \cdots, K$, define the set

$$U(i, j) = \{\boldsymbol{\pi} \in \mathscr{S}_K \mid \hat{\theta}_N = i, \hat{\theta}_B = j\}.$$

Then

$$\int_{\mathscr{S}_K} \sum_k (L(\hat{\theta}_N, k) - L(\hat{\theta}_B, k))\, \pi_k f_k\, d\boldsymbol{\pi} = \sum_{i \neq j} \int_{U(i,j)} \sum_k (L(i, k) = L(j, k))\pi_k f_k\, d\boldsymbol{\pi}. \tag{4.4}$$

At this point the proof involves a lot of technical details which are not particularly enlightening. Hence this part of the proof is summarized in the following lemma. The ideas behind the lemma are most easily seen in the case $K = 2$, with 0-1 loss. So, the proof will be given here only in that case. The complete proof may be found in Marron (1982) as the proof of Lemma 6.1.

LEMMA 4.1. *There is a constant B_1, so that for $\mathbf{x} \in \mathscr{C}$, and $i, j = 1, \cdots, K$,*

$$\int_{U(i,j)} \sum_k (L(i, k) - L(j, k))\pi_k f_k(\mathbf{x})\, d\pi \leq B_1 \sum_k (\hat{f}_k(\mathbf{x}) - f_k(\mathbf{x}))^2.$$

To prove Lemma 4.1, first suppose L is 0-1 loss and $K = 2$. Since $\mathbf{x}$ may be considered fixed here, $f_k(\mathbf{x})$ and $\hat{f}_k(\mathbf{x})$ will be abbreviated to f_k and $\hat{f}_k$. Without loss of generality, let $i = 1$ and $j = 2$.

Now, since $K = 2$, $\pi = (\pi_1, \pi_2)$ is determind by π_1, so $U(1, 2)$ may be considered to be a subset of the unit interval. Since L is 0-1 loss,

$$R_f(1, \mathbf{x}, \pi) = \frac{(1 - \pi_1)f_2}{\pi_1 f_1 + (1 - \pi_1)f_2}, \quad R_f(2, \mathbf{x}, \pi) = \frac{\pi_1 f_1}{\pi_1 f_1 + (1 - \pi_1)f_2}.$$

Hence, by (1.5), $\hat{\theta}_B(\mathbf{x}, \pi) = 2$ for $\pi_1 \in (0, f_2/(f_1 + f_2))$, and $\hat{\theta}_B(\mathbf{x}, \pi) = 1$ for $\pi_1 \in (f_2/(f_1 + f_2), 1)$. To simplify the notation, let $a = \hat{f}_2/(\hat{f}_1 + \hat{f}_2)$ and $b = f_2/(f_1 + f_2)$. By (4.2), note that $\hat{\theta}_N(x, \pi) = 2$ for $\pi_1 \in (0, a)$, and $\hat{\theta}_N(\mathbf{x}, \pi) = 1$ for $\pi_1 \in (a, 1)$. Thus, $U(1, 2)$ is contained in the (possibly empty) interval $[a, b]$. So, in the case $a > b$, $U(1, 2)$ is empty and the proof of Lemma 4.1 is complete.

Suppose now that $a < b$. Note that

$$b - a = \frac{\hat{f}_1 f_2 - f_1 \hat{f}_2}{(f_1 + f_2)(\hat{f}_1 + \hat{f}_2)} = \frac{(\hat{f}_1 - f_1)f_2 + f_1(f_2 - \hat{f}_2)}{(f_1 + f_2)(\hat{f}_1 + \hat{f}_2)}.$$

By property ii) of the class $\mathscr{F}_1$, it may be assumed that $\hat{f}_1, \hat{f}_2 \geq 1/2M$. Hence by properties i) and ii) in the definition of $\mathscr{F}_1$, there is a constant, B_2, so that

$$b - a \leq B_2(|\hat{f}_1 - f_1| + |\hat{f}_2 - f_2|). \tag{4.5}$$

Next note that

$$\int_{U(1,2)} \sum_k (L(1, k) - L(2, k))\pi_k f_k\, d\pi = \int_a^b (1 - \pi_1)f_2 - \pi_1 f_1\, d\pi_1. \tag{4.6}$$

But $(1 - \pi_1)f_2 - \pi_1 f_1$ is a linear function which has bounded slope—by i) from the definition of $\mathscr{F}_1$—and has its 0 at b. Thus, there is a constant, B_3, so that

$$\int_a^b (1 - \pi_1)f_2 - \pi_1 f_1\, d\pi_1 \leq B_3(a - b)^2. \tag{4.7}$$

But now, by (4.5), (4.6), and (4.7), the proof of Lemma 4.1 is complete.

To finish the proof of Theorem 1, note that by Lemma 4.1, (4.3), and (4.4), there is a constant B_4, so that,

$$\mathscr{E}_N \leq B_4 \sum_k \int_{\mathscr{C}} (\hat{f}_k(\mathbf{x}) - f_k(\mathbf{x}))^2\, d\mathbf{x}. \tag{4.8}$$

Next, for $k = 1, \cdots, K$ and for $c \in \mathbb{R}$, let A_k denote the event that

$$\int_{\mathscr{C}} (\hat{f}_k(\mathbf{x}) - f_k(\mathbf{x}))^2\, d\mathbf{x} \leq \frac{c}{B_4} N^{-r}.$$

Recall from the structure of Z^N and from the definition of the $\hat{f}_k$, that the A_k are

independent. Thus, from (4.8)

$$P_{\mathbf{f}}[\mathscr{L}_N \leq cN^{-r}] \leq P_{\mathbf{f}}[\cap_k A_k] = \prod_k P_{\mathbf{f}}[A_k].$$

But now, from Theorem 3, for $c = c_3 B_4$, $k = 1, \cdots, K$,

$$\lim_{N\to\infty} \sup_{f_k \in \mathscr{F}_1} P_{f_k}[A_k] = 1.$$

From which it follows that,

$$\lim_{N\to\infty} \sup_{\mathbf{f}\in\mathscr{F}} P_f[\mathscr{L}_N > cN^{-r}] = 0.$$

This completes the proof of Theorem 1.

5. Proof of Theorem 2. Suppose a classification rule, $\hat{\theta}_N(\mathbf{x}, \pi, Z^N)$, is given. For each N, it is desired to show that $\hat{\theta}_N$ behaves poorly for some choice of the underlying densities.

To do this, let $\mathbf{f}$ be any fixed element of $\mathscr{F}$, which is in the interior of $\mathscr{F}$ in the sense that f_1 satisfies the bounds in the definition of $\mathscr{F}_1$ with M replaced by a constant $M' < M$. Then, for each N, a finite family of perturbations of f_1 will be constructed, and it will be shown $\hat{\theta}_N$ behaves poorly for at least one of these. The perturbations will be small in the sense that they will converge uniformly to f_1 (as $N \to \infty$).

Since the compact set $\mathscr{L}$ has nonempty interior, assume, without loss of generality, that $\mathscr{L}$ is the unit in $\mathbb{R}^d$.

Given $\alpha > 0$, which will be specified later, it will be convenient to define $\tilde{N} = [N^\alpha]^d$, where [ ] denotes greatest integer.

In the following, for each $N \in \mathbb{Z}^+$, a number of quantities will be defined which will also be indexed by $l = 1, \cdots, \tilde{N}$. For notational convenience, these quantities will be subscripted only by l, with the dependence on N understood.

Given $N \in \mathbb{Z}^+$, let $\mathscr{L}_1, \cdots, \mathscr{L}_{\tilde{N}}$ denote a partitioning of $\mathscr{L}$ into subcubes, each having sidelength $1/[N^\alpha]$. For $l = 1, \cdots, \tilde{N}$,

$$\text{vol}_d(\mathscr{L}_l) = 1/\tilde{N},$$

where vol_d denotes the usual d-dimensional Euclidean volume.

For $l = 1, \cdots, \tilde{N}$, let $\mathbf{x}^l$ be the centerpoint of the cube $\mathscr{L}_l$, and let $\mathbf{y}^l$ be the vertex closest to the origin in $\mathbb{R}^d$.

Let $\psi: \mathbb{R}^d \to [0, \infty)$ be a function with the following properties:

i) ψ is m times continuously differentiable.
ii) for $|\alpha| \leq m$, $|D_\alpha \psi(\mathbf{x})| \leq 1$ on $\mathscr{L}$.
iii) for $|\alpha| \leq m$, $D_\alpha \psi(\mathbf{x})$ is supported inside $\mathscr{L}$.
iv) there is a constant, $\varepsilon > 0$, and a set, $U \subset \mathbb{R}^d$, so that $\psi(\mathbf{x}) \geq \varepsilon$ on U and $\text{vol}_d(U) \geq (\frac{1}{2})^d$.

Such a ψ may be constructed, for example, as a piecewise polynomial. Note that, by ii) and iii) with $\alpha = \mathbf{0}$, ψ is supported on $\mathscr{L}$ and $0 \leq \psi \leq 1$.

For $l = 1, \cdots, \tilde{N}$, given $a_l, a_l' > 0$, which will be specified later, define

$$\phi_l(\mathbf{x}) = a_l N^{-p\alpha} \psi(2[N^\alpha](\mathbf{x} - \mathbf{x}^l)) - a_l' N^{-p\alpha} \psi(2[N^\alpha](\mathbf{x} - \mathbf{y}^l)).$$

Note that ϕ_l vanishes everywhere, except on the part of $\mathscr{L}_l$ nearest the origin in $\mathbb{R}^d$, where it may be negative, and on the part of $\mathscr{L}_l$ farthest from the origin, where it may be positive.

To choose the a_l and a_l', first make them satisfy, for $l = 1, \cdots, \tilde{N}$.

$$\int_{\mathbb{R}^d} \phi_l(\mathbf{x}) f_1(\mathbf{x})\, d\mathbf{x} = 0. \tag{5.1}$$

This relationship is linear in a_l and a_l', so there is still "one degree of freedom" left. Let

$\partial > 0$ be a number which will be specified later, but which satisfies

$$0 < \partial < 1.$$

For $l = 1, \cdots, \tilde{N}$, choose a_l and a_l' so that

$$\partial = \max\{a_l, a_l'\}.$$

By (5.1) and the fact that $f_1 \in \mathscr{F}_1$, note that

$$\min\{a_l, a_l'\} \geq \partial/M^2.$$

By property iv) of the function ψ, for $l = 1, \cdots, \tilde{N}$, find a set $U_l \subset \mathscr{C}_l$, so that

$$\phi_l(x) \geq \partial \varepsilon N^{-p\alpha}/M^2 \quad \text{on} \quad U_l, \tag{5.2}$$

and

$$\text{vol}_d(U_l) \geq (\tfrac{1}{4}[N^\alpha])^d = \tfrac{1}{4}^d \tilde{N}. \tag{5.3}$$

Before defining the perturbations of f_1, a method of indexing is required. So, let

$$\mathscr{I}_N = \{0, 1\}^{\tilde{N}} = \{(b_1, \cdots, b_{\tilde{N}}) \text{: each } b_1 = 0 \text{ or } 1\},$$

Note that the cardinality of the set $\mathscr{I}_N$ is

$$\#(\mathscr{I}_N) = 2^{\tilde{N}}.$$

Now the family of perturbations will be defined. For $\mathbf{b} \in \mathscr{I}_N$, define the function

$$g_{\mathbf{b}}(x) = f_1(\mathbf{x}) + \textstyle\sum_{l=1}^{\tilde{N}} b_l \phi_l(\mathbf{x}) f_1(\mathbf{x}). \tag{5.4}$$

Also let $\mathbf{f}^{\mathbf{b}} = (g_{\mathbf{b}}, f_2, \cdots, f_k)$. In the following it will be convenient to let

$$R_{\mathbf{b}}, P_{\mathbf{b}}, E_{\mathbf{b}}, \quad \text{denote} \quad R_{\mathbf{f}}, P_{\mathbf{f}}, E_{\mathbf{f}} \quad \text{when} \quad \mathbf{f} = \mathbf{f}^{\mathbf{b}}. \tag{5.5}$$

As in Section 4, a number of constants, denoted by B_i, for $i \in \mathbb{Z}^+$, will be introduced. These will be independent of $\mathbf{f}$, $\mathbf{x}$, π, N, Z^N, $\hat{\theta}_N$, l, $\mathbf{b}$ and any quantities defined in terms of them. However, the B_i may depend on any, or all, of d, K, L, $\mathscr{C}$, m, β, p, M, M', α, ψ, ε, U, and ∂.

Next, for N sufficiently large, it is seen that $\mathbf{f}^{\mathbf{b}} \in \mathscr{F}$.

LEMMA 5.1. *There is a constant, B_5, so that, if $N > B_5$ and $\mathbf{b} \in \mathscr{I}_N$, then $g_{\mathbf{b}} \in \mathscr{F}_1$.*

The proof of this lemma is straightforward but tedious and hence is omitted. The details may be found in Section 7.1 of Marron (1982).

Now for any particular value of $l = 1, \cdots, \tilde{N}$ and any particular realization of Z^N, it will be useful to compare the function $R_{\mathbf{b}}(\hat{\theta}_N, \mathbf{x}, \pi) - R_{\mathbf{b}}(\hat{\theta}_B, \mathbf{x}, \pi)$ when $b_l = 0$ with the function when $b_l = 1$. Given $\mathbf{b} \in \mathscr{I}_N$ and $l = 1, \cdots, \tilde{N}$ it will be convenient (similar to (5.5)) to let

$$R_i \text{ and } g_i \text{ denote } R_{\mathbf{b}} \text{ and } g_{\mathbf{b}} \text{ when } b_l = i, \text{ for } i = 0, 1. \tag{5.6}$$

Recall the definition of U_l from (5.2).

LEMMA 5.2. *There is a constant B_6 so that for $N > B_5$, for each realization of Z^N, for each $l = 1, \cdots, \tilde{N}$, and for each $\mathbf{b} \in \mathscr{I}_N$, there is a set $U_l' \subset U_l$ for which:*

a) $\text{vol}_d(U_l') \geq (\tfrac{1}{2})\text{vol}_d(U_l)$, *and*

b) *one of the following hold*:

i) $\text{vol}_{K-1}\{\pi \in \mathscr{S}_K : R_0(\hat{\theta}_N, \mathbf{x}, \pi) - R_0(\hat{\theta}_B, \mathbf{x}, \pi) > B_6 N^{-p\alpha}\} > B_6 N^{-p\alpha}$,

for all $\mathbf{x} \in U_l'$,

or

ii) $\text{vol}_{K-1}\{\pi \in \mathscr{S}_K : R_1(\hat{\theta}_N, \mathbf{x}, \pi) - R_1(\hat{\theta}_B, \mathbf{x}, \pi) > B_6 N^{-p\alpha}\} > B_6 N^{-p\alpha}$,

for all $\mathbf{x} \in U_l'$.

The proof of this lemma for general K may be found in Marron (1982), Section 7.5. It involves reducing the problem to a case that is only somewhat more complicated than the case $K = 2$. Hence, the lemma will be proven here only in the case $K = 2$.

To verify Lemma 5.2, note that it is enough to show that there is a constant B_7, so that for each $\mathbf{x} \in U_l$, either

$$\text{vol}_{K-1}\{\pi \in \mathscr{S}_K : R_0(\hat{\theta}_N, \mathbf{x}, \pi) - R_0(\hat{\theta}_B, \mathbf{x}, \pi) > B_7 N^{-p\alpha}\} > B_7 N^{-p\alpha}$$

or

$$\text{vol}_{K-1}\{\pi \in \mathscr{S}_K : R_1(\hat{\theta}_N, \mathbf{x}, \pi) - R_1(\hat{\theta}_B, \mathbf{x}, \pi) > B_7 N^{-p\alpha}\} > B_7 N^{-p\alpha}.$$

But since $K = 2$, $\mathscr{S}_K$ is just the line segment in $\mathbb{R}^2$ with endpoints (0, 1) and (1, 0). Also $\pi_2 = 1 - \pi_1$. Thus, it is enough to show that there is a constant B_8, so that for each $\mathbf{x} \in U_l$, either

$$\text{vol}_1\{\pi_1 \in (0, 1) : R_0(\hat{\theta}_N, \mathbf{x}, \pi) - R_0(\hat{\theta}_B, \mathbf{x}, \pi) > B_8 N^{-p\alpha}\} > B_8 N^{-p\alpha}, \tag{5.7}$$

or

$$\text{vol}_1\{\pi_1 \in (0, 1) : R_1(\hat{\theta}_N, \mathbf{x}, \pi) - R_1(\hat{\theta}_B, \mathbf{x}, \pi) > B_8 N^{-p\alpha}\} > B_8 N^{-p\alpha}.$$

Now $\mathbf{x}$ may be considered fixed, so dependence on it will be suppressed, hence $f_k(\mathbf{x})$ will be denoted f_k and so on. From properties i) and ii) in the definition of the class $\mathscr{F}_1$, for $k = 1, \cdots, K$, recall

$$1/M \leq f_k \leq M. \tag{5.8}$$

Also, since $\mathbf{x} \in U_l \subset \mathscr{C}_l$, from (5.4) and (5.6),

$$g_0 = f_1, \quad g_1 = f_1(1 + \phi_l),$$

and

$$1/M \leq f_1(1 + \phi_l) \leq M, \tag{5.9}$$

and by (5.2), there is a constant B_9, so that

$$f_1 \phi_l > B_9 N^{-p\alpha}. \tag{5.10}$$

From (1.3) and (1.4) note that for $k = 1, 2$,

$$R_0(k, \mathbf{x}, \pi) = \frac{\pi_1 L(k, 1) f_1 + (1 - \pi_1) L(k, 2) f_2}{\pi_1 f_1 + (1 - \pi_1) f_2},$$

$$R_1(k, \mathbf{x}, \pi) = \frac{\pi_1 L(k, 1) f_1(1 + \phi_l) + (1 - \pi_1) L(k, 2) f_2}{\pi_1 f_1(1 + \phi_l) + (1 - \pi_1) f_2}.$$

It will be convenient to define, for $k = 1, 2$,

$$\Delta_0(k, \pi_1) = (L(k, 1) f_1 - L(k, 2) f_2)\pi_1 + L(k, 2) f_2, \tag{5.11}$$

$$\Delta_1(k, \pi_1) = \Delta_0(k, \pi_1) + L(k, 1) f_1 \phi_l \pi_1. \tag{5.12}$$

Now from (5.7), by (5.8) and (5.9), the proof of Lemma 5.2 will be complete when it is shown that there is a constant, B_{10}, so that either:

$$\text{vol}_1\{\pi_1 \in (0, 1) : \Delta_0(\hat{\theta}_N, \pi_1) - \Delta_0(\hat{\theta}_B, \pi_1) > B_{10} N^{-p\alpha}\} > B_{10} N^{-p\alpha} \tag{5.13}$$

or,

$$\text{vol}_1\{\pi_1 \in (0, 1): \Delta_1(\hat{\theta}_N, \pi_1) - \Delta_1(\hat{\theta}_B, \pi_1) > B_{10}N^{-p\alpha}\} > B_{10}N^{-p\alpha}. \tag{5.14}$$

From (1.5), (5.11), and (5.12), for $i = 0, 1$, when $b_l = i$,

$$\Delta_i(1, \pi_1) \leq \Delta_i(2, \pi_1) \quad \text{implies} \quad \hat{\theta}_B = 1, \quad \text{and}$$

$$\Delta_i(1, \pi_1) > \Delta_i(2, \pi_1) \quad \text{implies} \quad \hat{\theta}_B = 2.$$

Thus, by (1.1), (5.10), and (5.12), there is a constant B_{11}, so that an interval (a, b) can be found such that

$$b > a + B_{11}N^{-p\alpha},$$

and such that, for $\pi_1 \in (a, b)$,

$$b_l = 0 \quad \text{implies} \quad \hat{\theta}_B = 1, \quad \text{and} \quad b_l = 1 \quad \text{implies} \quad \hat{\theta}_B = 2. \tag{5.15}$$

But, by (1.1), (5.11), and (5.12), for $i = 0, 1$, $\Delta_0(1, \pi_1) - \Delta_0(2, \pi_1)$ is a linear function of π_1, whose slope is bounded away from 0. Hence, by (5.15), either (5.13) or (5.14) holds.

This completes the proof of Lemma 5.2 in the case $K = 2$.

Next, a method is needed to simultaneously take into account what is happening on $\mathscr{C}_1, \cdots, \mathscr{C}_{\tilde{N}}$. To do this, it is convenient to make a slight addition to the probability structure. Let β be an $\mathscr{I}_N$-valued random variable which takes on each of the $2^{\tilde{N}}$ values with equal probability. Then suppose the distribution of Z^N is determined by $\mathbf{f}^\beta$.

Next, consider the classification problem of guessing the value of β, using the observed value of Z^N. Let $\hat{\beta} = (\hat{\beta}_1, \cdots, \hat{\beta}_{\tilde{N}})$ denote a classification rule, or more precisely, a measurable function from $(\mathbb{R}^d)^{KN}$ to $\mathscr{I}_N$.

LEMMA 5.3. *For $\alpha \geq 1/(2p + d)$ and ∂ sufficiently small, there is a constant $B_{12} > 0$, so that*

$$\lim_{N\to\infty} \sup_{\mathbf{b}\in\mathscr{I}_N} P_{\mathbf{b}}[\#(l: \hat{\beta}_l \neq b_l) > B_{12}\tilde{N}] = 1.$$

A similar result has been proved in Stone (1982), hence the proof is omitted. The details may be found in Marron (1982).

Now, for $N > B_5$, define a classifier, $\hat{\beta}$, by, for $l = 1, \cdots \tilde{N}$, letting

$$\hat{\beta}_l = 1 \quad \text{when i) in Lemma 5.2 holds, and}$$

$$\hat{\beta}_l = 0 \quad \text{otherwise.}$$

Note that, for $N > B_5$, and for $l = 1, \cdots, \tilde{N}$, in the event $\{\beta_l = 0\}$, $\hat{\beta}_l \neq \beta_l$ implies i) holds. Similarly, by Lemma 5.2, in the event $\{\beta_l = 1\}$, $\hat{\beta}_l \neq \beta_l$ implies ii) holds. Hence, for each $\mathbf{b} \in \mathscr{I}_N$, $\hat{\beta}_l \neq b_l$ implies

$$\text{vol}_{K-1}\{\boldsymbol{\pi} \in \mathscr{S}_K: R_{\mathbf{b}}(\hat{\theta}_N, \mathbf{x}, \boldsymbol{\pi}) - R_{\mathbf{b}}(\hat{\theta}_B, \mathbf{x}, \boldsymbol{\pi}) > B_6N^{-p\alpha}\} > B_6N^{-p\alpha}, \quad \text{for every} \quad \mathbf{x} \in U_l',$$

which in turn implies

$$\int_{\mathscr{S}_K} R_{\mathbf{b}}(\hat{\theta}_N, \mathbf{x}, \boldsymbol{\pi}) - R_{\mathbf{b}}(\hat{\theta}_B, \mathbf{x}, \boldsymbol{\pi})\, d\boldsymbol{\pi} > B_6^2N^{-2p\alpha} \quad \text{for} \quad \mathbf{x} \in U_l'.$$

Therefore, by (5.3) and a) in Lemma 5.2, there is a constant B_{13}, so that $\hat{\beta}_l \neq b_l$ implies

$$\int_{U_l'} \int_{\mathscr{S}_K} R_{\mathbf{b}}(\hat{\theta}_N, \mathbf{x}, \boldsymbol{\pi}) - R_{\mathbf{b}}(\hat{\theta}_B, \mathbf{x}, \boldsymbol{\pi})\, d\boldsymbol{\pi}\, d\mathbf{x} > \frac{B_{13}N^{-2p\alpha}}{\tilde{N}}.$$

Next, for $N > B_5$, and for any $\mathbf{b} \in \mathscr{I}_N$, on the event $\{\#(l: \hat{\beta}_l \neq b_l) > B_{12}\tilde{N}\}$, (using

Fubini's theorem since the integrand is nonnegative), note that

$$\int_{\mathscr{S}_K}\int_{\mathscr{C}} R_{\mathbf{b}}(\hat{\theta}_N, \mathbf{x}, \pi) - R_{\mathbf{b}}(\hat{\theta}_B, \mathbf{x}, \pi)\, d\mathbf{x}\, d\pi$$

$$\geq \sum_{l:\hat{\beta}_l \neq \beta_l} \int_{U_l}\int_{\mathscr{S}_K} R_{\mathbf{b}}(\hat{\theta}_N, \mathbf{x}, \pi) - R_{\mathbf{b}}(\hat{\theta}_B, \mathbf{x}, \pi)\, d\pi\, d\mathbf{x} > B_{12}B_{13}N^{-2p\alpha}.$$

Thus, by Lemma 5.3, for $\alpha \geq 1/(2p + d)$,

$$\lim_{N\to\infty}\sup_{\mathbf{b}\in\mathscr{S}_N} P_{\mathbf{b}}\left[\int_{\mathscr{S}_K}\int_{\mathscr{C}} [R_{\mathbf{b}}(\hat{\theta}_N, \mathbf{x}, \pi) - R_{\mathbf{b}}(\hat{\theta}_B, \mathbf{x}, \pi)]\, d\mathbf{x}\, d\pi > B_{12}B_{13}N^{-2p\alpha}\right] = 1.$$

So, let $\alpha = 1/(2p + d)$ and the proof of the Theorem 2 is complete.

Acknowledgment. The author is grateful to Charles J. Stone for posing the problem treated in this paper and for much help, criticism and encouragement during all phases of this research. Thanks are also due the anonymous referee for pointing the way to several recent references.

REFERENCES

BRETAGNOLLE, J. and HUBER, C. (1979). Estimation des densities: risque minimax. *Z. Wahrsch. verw. Gebiete* **47** 119–137.

DAS GUPTA, S. (1964). Nonparametric classification rules. *Sankyā, Ser. A* **26** 25–30.

DEVROYE, L. P. and WAGNER, T. J. (1977). Nonparametric discrimination and density estimation. Tech. Report 183, Electronics Research Center, the Univ. of Texas at Austin.

DEVROYE, L. P. and WAGNER, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Ann. Statist.* **8** 231–239.

EPANECHNIKOV, V. (1969). Nonparametric estimation of a multivariate probability density. *Theor. Probab. Appl.* **14** 153–158.

FARRELL, R. H. (1972). On the best obtainable asymptotic rates of convergence in estimation of a density function at a point. *Ann. Math. Statist.* **43** 170–180.

FIX, E. and HODGES, J. L., JR. (1951). Discriminatory analysis, nonparametric discrimination, consistency properties. Randolph Field, Texas, Project 21-49-004, Report No. 4.

GLICK, N. (1972). Sample-based classification procedures derived from density estimators. *J. Amer. Statist. Assoc.* **67** 116–122.

GORDON, L. and OLSHEN, R. A. (1978). Asymptotically efficient solutions to the classification problem. *Ann. Statist.* **6** 515–533.

GREBLICKI, W. (1978). Asymptotically optimal pattern recognition procedures with density estimates. *IEEE Trans. Information Theory.* **IT-24** 250–251.

GREBLICKI, W. (1981). Asymptotic efficiency of classifying procedures using the Hermite series estimate of multivariate probability densities. *IEEE Trans. Information Theory* **IT-27** 364–366.

GREBLICKI, W. and PAWLAK, M. (1982). A classification procedure using the multiple Fourier Series. *Information Sciences* **26** 115–126.

GYÖRFI, L. (1978). On the rate of convergence of nearest neighbor rules. *IEEE Trans. Information Theory.* **IT-24** 509–512.

GYÖRFI, L. (1981). The rate of convergence of k_n-*NN* regression estimates and classification rules. *IEEE Trans. Information Theory.* **IT-27** 362–364.

KHASMINSKII, R. Z. (1978). A lower bound on the risks of nonparametric estimates of densities in the uniform metric. *Theor. Probability Appl.* **23** 794–798.

MARRON, J. S. (1982). Optimal rates of convergence in nonparametric discrimination. (Ph.D. Dissertation, UCLA).

MÜLLER, H. and GASSER, T. (1979). Optimal convergence properties of kernel estimates of derivatives of a density function. *Smoothing Techniques for Curve Estimation. Lecture Notes in Mathematics* **757** 143–144. Springer, Berlin.

QUESENBERRY, C. P. and LOFTSGAARDEN, D. O. (1965). Nonparametric density estimation and classification. Tech. Note D-2699, NASA.

ROSENBLATT, M. (1956). Remarks on some nonparametric estimates of a density function. *Ann. Math. Statist.* **27** 832–837.

STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645.
STONE, C. J. (1980). Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8** 1348–1360.
STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.
TARTER, M. E. and KRONMAL, R. A. (1976). An introduction to the implementation and theory of nonparametric density estimation. *Amer. Statist.* **30** 105–112.
VAN HOUWELINGEN, H. (1980). Risk convergence rates in empirical Bayes classification. *Scand. J. Statist.* **7** 99–101.
VAN RYZIN, J. (1966). Bayes risk consistency of classification procedures using density estimation. *Sankyā, Ser. A.* **28** 261–270.
WAHBA, G. (1975). Optimal convergence properties of variable knot, kernel, and orthogonal series methods for density estimation. *Ann. Statist.* **3** 15–29.
WALTER, G. and BLUM, J. (1979). Probability density estimation using delta sequences. *Ann. Statist.* **7** 328–340.
WEGMAN, E. J. (1972a). Nonparametric probability density estimation: I. a summary of available methods. *Technometrics* **14** 533–546.
WEGMAN, E. J. (1972b). Nonparametric probability density estimation: II. a comparison of density estimation methods. *J. Statist. Comp. and Simulation* **1** 225–245.
WERTZ, W. (1978). Statistical density estimation: a survey. *Angewandte Statistique und Okonometrie* **13**. Vandenhoeck and Ruprecht, Göttinger.

THE UNIVERSITY OF NORTH CAROLINA
DEPARTMENT OF STATISTICS
321 PHILLIPS HALL 039A
CHAPEL HILL, NORTH CAROLINA 27514.