On the Finite Sample Performance of the Nearest Neighbor Classifier

Demetri Psaltis, Robert R. Snapp, and Santosh S. Venkatesh

***Abstract*—The finite sample performance of a nearest neighbor classifier is analyzed for a two-class pattern recognition problem. An exact integral expression is derived for the m-sample risk R_m given that a reference m-sample of labeled points is available to the classifier. The statistical setup assumes that the pattern classes arise in nature with fixed *a priori* probabilities and that points representing the classes are drawn from Euclidean n-space according to fixed class-conditional probability distributions. The sample is assumed to consist of m independently generated class-labeled points. For a family of smooth class-conditional distributions characterized by asymptotic expansions in general form, it is shown that the m-sample risk R_m has a complete asymptotic series expansion**

$$R_m \sim R_\infty + \sum_{k=2}^{\infty} c_k m^{-k/n} \qquad (m \to \infty),$$

where R_∞ denotes the nearest neighbor risk in the infinite-sample limit and the coefficients c_k are distribution-dependent constants independent of the sample size m. This analysis thus provides further analytic validation of Bellman's curse of dimensionality. Numerical simulations corroborating the formal results are included, and extensions of the theory discussed. The analysis also contains a novel application of Laplace's asymptotic method of integration to a multidimensional integral where the integrand attains its maximum on a continuum of points.

***Index Terms*—Nearest neighbor classifier, Pattern classification, Curse of dimensionality, Laplace's method of integration.**

I. Introduction

Because of its simplicity and nearly optimal performance in the large sample limit, the nearest neighbor classifier (see Duda and Hart [1], for example) endures as a fundamental algorithm for pattern recognition and signal identification. In its classical manifestation, pattern classes are assumed to generate random points, or feature vectors, in some n-dimensional metric space. First, a reference sample of m labeled feature vectors is constructed, each label indicating the pattern class from which the associated vector originated. The nearest neighbor classifier then assigns any input feature vector to the class indicated by the label of the nearest reference vector.

The simplicity of this nonparametric classifier belies its performance. When the reference sample is drawn independently according to a stationary underlying distribution, a classical result of Cover and Hart [2] asserts that in the infinite-sample limit ($m \to \infty$), the probability that an independently selected feature vector (drawn again from the same underlying distribution) is misclassified, is no more than twice the (optimal) Bayes error. Thus, if the Bayes error is small, the nearest neighbor classifier performs nearly optimally in the large sample limit. In practice, however, the sample must be finite. Furthermore, data storage and access costs favor small samples. Thus we are led to the following question of theoretical and practical import: *How rapidly does the statistical risk R_m of a nearest neighbor classifier approach its infinite-sample limit R_∞?*

For problems with two pattern classes and a one-dimensional feature space, Cover [3] has shown that the infinite-sample limit is approached as rapidly as $R_m = R_\infty + O(m^{-2})$ ($m \to \infty$) if the probability distributions that define the classification problem are sufficiently smooth.[1] More recently, Fukunaga and Hummels [7] have studied the rate of convergence of the statistical risk in an n-dimensional feature space. Using a series of nonrigorous approximations based on a second-order Taylor series expansion, they obtained the heuristic estimate

$$R_m \approx R_\infty + B\frac{\Gamma(m+1)}{\Gamma\left(m+1+\dfrac{2}{n}\right)}, \tag{1}$$

where Γ is the gamma function and B is a distribution-dependent constant. For large sample sizes m, the approximation gives $m^{-2/n}$ as the rate of convergence of R_m to

Manuscript received September 10, 1991; revised manuscript received May 3, 1993. This work was presented in part at the *IEEE International Symposium on Information Theory*, San Antonio, Texas, January 1993. The work reported here was supported in part by the Air Force Office of Scientific Research under grant AFOSR 89-0523 to Santosh S. Venkatesh.

D. Psaltis is with the Department of Electrical Engineering, California, Institute of Technology, Pasadena, CA 91125.

R. R. Snapp is with the Department of Computer Science and Electrical Engineering, University of Vermont, Burlington, VT 05405.

S. S. Ventatesh is with the Department of Electrical Engineering, University of Pennsylvania, Philadelphia PA 19104.

IEEE Log Number 9402031.

[1] On an alternative track, early studies have also estimated how the conditional finite risk, i.e., the probability of error of a nearest neighbor classifier with a given reference sample of m random patterns, converges in probability to R_∞. This has been investigated under a variety of conditions by Wagner [4], Fritz [5], and Györfi [6].

0018-9448/94$04.00

R_∞, which is in accord with Cover's result for one dimension ($n = 1$). While the approximation (1) certainly cannot hold without qualification—convergence can be arbitrarily slow, even in one dimension if smoothness conditions are not mandated on the distributions [3]—nonetheless, Cover's result for one dimension, together with simulation results for smooth distributions in higher dimensions suggest that the approximation (1) can be made rigorous, at least for sufficiently large sample sizes, for a suitably constrained family of smooth distributions.

In the sequel, we delineate a nontrivial family of "smooth" pattern recognition problems for which local approximation methods are valid, resulting in rigorous large-sample approximations of the form (1). More generally, we describe a formal analytic analytic technique, using higher-order series expansions, that yields more accurate approximations of the statistical risk, over a wide range of sample sizes, for the specified problem class.

Our main result is the demonstration that, if the class-conditional distributions are absolutely continuous with densities admitting of uniform asymptotic expansions, then the m-sample risk of the nearest neighbor classifier can be written as a complete asymptotic series expansion:

$$R_m \sim R_\infty + \sum_{k=2}^{\infty} c_k m^{-k/n} \qquad (m \to \infty). \tag{2}$$

(Anciliary conditions required in the demonstration of (2) are developed in the body of the paper.) This series converges in the sense of Poincaré: the approximation error of a truncated series is bounded by the magnitude of the first neglected term. Thus, for example, the third term in the above expansion can be used to estimate the approximation error of Fukunaga and Hummel's second-order truncation. The expansion coefficients c_k depend upon the underlying probability distributions, and are determined within the body of the analysis.

Within its realm of applicability, the formulation (2) constitutes a complete solution to the problem in that not only is the rate of approach, for large m, of the finite sample risk to the limiting behavior determined, but that the solution permits the investigation of classifier behavior for small sample sizes as well by inclusion of a suitable number of higher-order terms. Note that the lower-order terms of the expansion (2) indicate that an increase in the dimensionality of the feature space results in a dramatic decrease in the rate of convergence. Indeed, if at least one of the coefficients c_k is nonzero, then the sample size required to achieve a given level of performance grows exponentially with n. Thus the leading term in the expansion provides an analytic validation of Bellman's curse of dimensionality.

Formula (2) is derived by asymptotically integrating an exact expression for the m-sample risk, using a generalization of Laplace's method to multidimensional integrals. This method may have some intrinsic interest as it includes an asymptotic evaluation of a multidimensional Laplace-type integral of the form $\int g e^{-mh}$, where m is an integer parameter, g and h are functions of the variables of integration, and the function h has its minimum on a *continuum* of points in a linear subspace. (In a typical Laplace integral, h would have a minimum only at a discrete set of points.)

In Section II we review the nearest neighbor classifier in the context of a two class problem. The asymptotic results and the hypotheses under which they hold are formally stated in Section III. The proofs, which include explicit expressions for the leading coefficients in the asymptotic expansions, appear in Section IV. The utility of this asymptotic formula is confirmed by several numerical simulations described in Section V. Additional numerical experiments suggest that the results can be extended to a wider class of smooth problems than those captured by our hypotheses. Section VI contains a discussion of the hypotheses and outlines extensions.

On Notation: Logarithms are taken to the base e. We denote by Γ the gamma function $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt$. We use the usual notation $\nabla^2 \equiv \sum_{i=1}^n (\partial^2/\partial x_i^2)$ for the n-dimensional Laplacian, and

$$\nabla^{2k} \equiv (\nabla^2)^k = \sum_{i_1=1}^{n} \cdots \sum_{i_k=1}^{n} \frac{\partial^{2k}}{\partial x_{i_1}^2 \cdots \partial x_{i_k}^2},$$

for the kth power of the n-dimensional Laplacian differential operator.

Boldface letters such as $\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{y}, \cdots$ denote points in Euclidean n-space $\mathbb{R}^n$. Boldface capitals such as $\boldsymbol{X}, \boldsymbol{X}', \cdots$ denote random vectors drawn from $\mathbb{R}^n$. For $\boldsymbol{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n$, we use the usual norm $|\boldsymbol{x}| = (x_1^2 + \cdots + x_n^2)^{1/2}$; the induced metric $|\boldsymbol{x}' - \boldsymbol{x}|$ denotes the distance between any two points $\boldsymbol{x}'$ and $\boldsymbol{x}$. If $A \subset \mathbb{R}^n$ is any subset and $\boldsymbol{x} \in \mathbb{R}^n$ any point, we denote the distance between $\boldsymbol{x}$ and the closest point in A to $\boldsymbol{x}$ by $|\boldsymbol{x} - A| = \inf_{\boldsymbol{y} \in A} |\boldsymbol{x} - \boldsymbol{y}|$. We denote the smallest closed set that contains the subset $A \subset \mathbb{R}^n$ by $\mathrm{cl}(A)$. For any $\rho \geq 0$, we denote the (closed) ball of radius ρ at $\boldsymbol{x}$ by

$$B(\rho, \boldsymbol{x}) = \{\boldsymbol{y} \in \mathbb{R}^n : |\boldsymbol{y} - \boldsymbol{x}| \leq \rho\}.$$

With a slight abuse of notation, for any two points $\boldsymbol{x}$ and $\boldsymbol{x}'$ in $\mathbb{R}^n$ we hereafter denote their difference in spherical coordinates as $\boldsymbol{x}' - \boldsymbol{x} = (\rho, \Omega)$. Here $\rho = |\boldsymbol{x}' - \boldsymbol{x}|$ is the distance between $\boldsymbol{x}$ and $\boldsymbol{x}'$ and $\Omega = (\boldsymbol{x}' - \boldsymbol{x})/|\boldsymbol{x}' - \boldsymbol{x}|$ is a point on the surface S_{n-1} of the unit ball:

$$S_{n-1} = \{\Omega \in \mathbb{R}^n : |\Omega| = 1\}.$$

If $\mathscr{S}$ is any subset of points in $\mathbb{R}^n$ we denote $\boldsymbol{x}' \xrightarrow{\mathscr{S}} \boldsymbol{x}$ and $\rho \xrightarrow{\mathscr{S}} 0$, interchangeably, to mean that $\boldsymbol{x}'$ approaches $\boldsymbol{x}$ through points in $\mathscr{S}$. By $d\boldsymbol{x}, d\boldsymbol{x}', \cdots$, we mean the usual (Lebesgue) element of measure in $\mathbb{R}^n$. We also use $d\Omega$ to denote the element of measure in S_{n-1}. (Note that S_0 corresponds to the endpoints of the real interval $[-1, 1]$, and integrals over S_0 reduce to discrete sums.)

Finally, we summarize our asymptotic notation. Let x be a variable taking values in a subset X of a metric space, and let x_0 be a limit point of X (which may or may not be in X). Let f and g be functions defined on X. We denote: $f(x) = O(g(x))\,(x \to x_0)$ if there exists a constant K and a neighborhood U of x_0 such that $|f| \le K|g|$ for all $x \in U \cap X$; $f(x) = o(g(x))\,(x \to x_0)$ if for any $\epsilon > 0$ there exists a neighborhood U_ϵ of x_0 such that $|f| \le \epsilon|g|$ for all $x \in U_\epsilon \cap X$; $f(x) \sim g(x)\ (x \to x_0)$ if $f(x) - g(x) = o(g(x))\,(x \to x_0)$. If the functions f and g depend upon additional parameters, but the constants K and neighborhoods U and U_ϵ defined above may be chosen independent of the parameters, then we say that the order relations hold *uniformly* in the parameters.

We conclude by reviewing the definition of an asymptotic power series (cf. Erdélyi [8]). Let x denote a real parameter. We say that a formal power series $\sum_{i=0}^{\infty} a_i x^{-i}$ is *asymptotic to the function $f(x)$ as $x \to \infty$* and we write $f(x) \sim \sum_{i=0}^{\infty} a_i x^{-i}\ (x \to \infty)$ iff $f(x) - \sum_{i=0}^{N} a_i x^{-i} = o(x^{-N})\,(x \to \infty)$ for every N.

II. Nearest Neighbor Classifier

Let "1" and "2" denote two states of nature corresponding to two pattern classes, and let P_1 and P_2 denote their respective prior probability of occurrence. (In what follows, we assume that $0 < P_1,\ P_2 < 1$.) The patterns themselves are represented by *feature vectors* $\boldsymbol{X} \in \mathbb{R}^n$ which, conditioned on class $j \in \{1, 2\}$, are drawn according to the class-conditional probability distribution F_j. The mixture distribution

$$F = P_1 F_1 + P_2 F_2$$

is then the unconditional distribution for the feature vector $\boldsymbol{X}$.

Labeled feature vectors, $(\boldsymbol{X}, \theta)$, are generated from the mixture distribution by the following process: first, a pattern class $\theta \in \{1, 2\}$ is chosen at random in accordance with the prior probability for each class, $\boldsymbol{P}\{\theta = j\} = P_j$; then, a feature vector $\boldsymbol{X} \in \mathbb{R}^n$, conditioned upon θ, is drawn according to F_θ. After m independent repetitions of this process, a reference sample

$$\Sigma_m = \{(\boldsymbol{X}^{(1)}, \theta^{(1)}), \cdots, (\boldsymbol{X}^{(m)}, \theta^{(m)})\},$$

is constructed. Here, each $\boldsymbol{X}^{(i)} \in \mathbb{R}^n$ denotes a feature vector, with $\theta^{(i)} \in \{1, 2\}$ the corresponding class label.

Given the m-sample Σ_m, the nearest neighbor classifier partitions the feature space $\mathbb{R}^n$ as follows:

Algorithm: To every $\boldsymbol{x} \in \mathbb{R}^n$, assign the class label $C(\boldsymbol{x}|\Sigma_m) \in \{1, 2\}$ given by

$$C(\boldsymbol{x}|\Sigma_m) = \theta^{(i)} \quad \text{if } |\boldsymbol{x} - \boldsymbol{X}^{(i)}| \le |\boldsymbol{x} - \boldsymbol{X}^{(j)}| \text{ for all } j \ne i.^2$$

A. Finite Sample Risk

A labeled test vector $(\boldsymbol{X}, \theta)$ is now drawn by the same process, independent of the m-sample Σ_m. Let $(\boldsymbol{X}', \theta')$ denote the element of Σ_m chosen by the classifier to be the nearest neighbor of $(\boldsymbol{X}, \theta)$. The classifier's average performance can be quantified in terms of the statistical risk. Specifically, the m-sample statistical risk is defined as

$$R_m = L_{12}\boldsymbol{P}[\theta' = 1, \theta = 2] + L_{21}\boldsymbol{P}[\theta' = 2, \theta = 1],$$

where L_{12} and L_{21} are given loss coefficients. Here, L_{11} and L_{22} are assumed to be zero. If $L_{12} = L_{21} = 1$, then $R_m = \boldsymbol{P}[\theta' \ne \theta]$, the probability that the nearest neighbor algorithm assigns $\boldsymbol{X}$ to the incorrect class. For notational simplicity, we will assume this *zero-one* risk function. Note, however, that because the risk depends linearly upon the loss coefficients, all that follows can immediately be extended to more general risk functions.

Let $\hat{P}(j|\boldsymbol{X})$ denote the *a posteriori* probability of class j conditioned on $\boldsymbol{X}$. We then have

$$R_m = \boldsymbol{E}(\boldsymbol{P}(\theta \ne \theta'|\boldsymbol{X}, \boldsymbol{X}')) = \boldsymbol{E}(\hat{P}(1|\boldsymbol{X})\hat{P}(2|\boldsymbol{X}')) + \boldsymbol{E}(\hat{P}(1|\boldsymbol{X}')\hat{P}(2|\boldsymbol{X})),$$

where the first equality obtains by conditioning the event $[\theta' \ne \theta]$ first over the values assumed by the test vector $\boldsymbol{X}$, and then over the values assumed by the nearest reference vector $\boldsymbol{X}'$, while the second equality follows from the independent generation of the test and reference vectors. Denoting by $F(\boldsymbol{x}', \boldsymbol{x})$ the joint distribution function of the pair of random vectors $(\boldsymbol{X}', \boldsymbol{X})$, we then have

$$R_m = \int \hat{P}(1|\boldsymbol{x})\hat{P}(2|\boldsymbol{x}')\, dF(\boldsymbol{x}', \boldsymbol{x}) + \int \hat{P}(1|\boldsymbol{x}')\hat{P}(2|\boldsymbol{x})\, dF(\boldsymbol{x}', \boldsymbol{x}). \tag{3}$$

In general this integral is quite difficult to evaluate. In Section III and IV we provide conditions under which this integral can be approximated over a large range of sample sizes m.

B. Infinite Sample Limit

Under rather general conditions, Cover and Hart [2] evaluated (3) in the large-sample limit ($m \to \infty$). We briefly sketch the main result. Note first that the Glivenko–Cantelli theorem (cf. Billingsley [9], for instance) readily yields $\boldsymbol{X}' \to \boldsymbol{X}$ a.e. Suppose now that the class-conditional distributions F_1 and F_2 are absolutely continuous, and the corresponding probability density functions f_1 and f_2 are continuous almost everywhere (Lebesgue measure) on their probability-one supports. Invoking the dominated convergence theorem, we then have

$$R_\infty \equiv \lim_{m \to \infty} R_m = 2\boldsymbol{E}(\hat{P}(1|\boldsymbol{X})\hat{P}(2|\boldsymbol{X})). \tag{4}$$

This expression for R_∞ can be readily expressed in terms of the Bayes risk. Let

$$r_B(\boldsymbol{X}) = \min\{\hat{P}(1|\boldsymbol{X}), \hat{P}(2|\boldsymbol{X})\} = \min\{\hat{P}(1|\boldsymbol{X}), 1 - \hat{P}(1|\boldsymbol{X})\}$$

[2] Ties may be resolved by any procedure. Under our subsequent assumptions, ties occur with zero probability.

denote the conditional Bayes risk given X, and let

$$R_B = E(r_B(X))$$

denote the (unconditional) Bayes risk. Note that by symmetry,

$$\begin{aligned}\hat{P}(1|X)\hat{P}(2|X) &= \hat{P}(1|X)(1-\hat{P}(1|X))\\ &= r_B(X)(1-r_B(X)),\end{aligned}$$

hence from (4),

$$\begin{aligned}R_\infty &= 2\big(E(r_B(X)) - E\big(r_B(X)^2\big)\big)\\ &= 2(R_B - \operatorname{Var}(r_B(X)) - R_B^2) \le 2R_B(1-R_B).\end{aligned}$$

As the Bayes classifier is optimal, R_∞ must satisfy $R_B \le R_\infty \le 2R_B$. Thus if $R_B \ll 1$, then the nearest neighbor classifier is nearly optimal in the infinite sample limit. This optimistic assessment, however, is of practical benefit only if R_m converges to R_∞ at a reasonable rate. In the next few sections we examine some nontrivial conditions under which this may occur.

III. Convergence Rate

Given the near optimality of the infinite-sample nearest neighbor classifier, it is of interest to know how the nearest neighbor classifier performs with a finite sample. In particular, how rapidly does R_m approaches R_∞? For one-dimensional feature spaces, Cover [3] showed the following:

- If the class-conditional distribution functions F_j are absolutely continuous, the class-conditional densities f_j have uniformly bounded third derivatives, and the mixture density f is bounded away from zero on its probability-one support set, then $R_m = R_\infty + O(m^{-2})$ $(m \to \infty)$.
- If the class-conditional densities are not smooth, then the convergence of R_m to R_∞ can be arbitrarily slow.

Our focus here is on obtaining a full asymptotic series expansion of the m-sample risk for any given dimensionality n of the feature space. In particular, this would enable a study of the behaviour of R_m for relatively small values of m. In the following, we list a set of conditions on the class-conditional densities under which we prove our main theorem. These conditions are not the least restrictive of their kind for which the theorem will hold; we present them in this form, however, so that technical difficulties do not obscure the main thread of the ideas. Moreover, for this class of problems, we can obtain relatively simple expressions for the coefficients of the asymptotic expansion for R_m. In subsequent sections we provide numerical examples indicating a range of applicability outside that formally covered in the theorem, and provide discussions on the implications of the constraints listed below, and the effect of relaxing them.

Hypotheses:

H1. For $j \in \{1,2\}$, the class-conditional distributions F_j are absolutely continuous over $\mathbb{R}^n$ and have corresponding densities f_j.

H2. The mixture density $f = P_1 f_1 + P_2 f_2$ is bounded away from zero a.e.[3] over its probability-one support $\mathscr{S} \subset \mathbb{R}^n$. (We can thus assume, without loss of generality, that $\mathscr{S}$ is compact.)

H3. There exists a fixed integer $N \ge 1$ such that for each $j \in \{1,2\}$ there exist continuous functions $f_{j,k}(\Omega, x)$ defined a.e., on $S_{n-1} \times \mathscr{S}$, for which the following asymptotic expansion holds uniformly in x as $x' \overset{\mathscr{S}}{\to} x$ $[x' - x = (\rho, \Omega)]$:

$$f_j(x') = f_j(x) + \sum_{k=1}^{N} f_{j,k}(\Omega, x)\rho^k + o(\rho^N) \qquad (\rho \overset{\mathscr{S}}{\to} 0). \quad (5)$$

Equivalently, for every $\epsilon > 0$, positive integer $N' \le N$, and a.e. $x \in \mathscr{S}$, there is a $\rho_0(\epsilon)$ independent of x, such that

$$\left| f_j(x') - f_j(x) - \sum_{k=1}^{N'} f_{j,k}(\Omega, x)\rho^k \right| < \epsilon\rho^{N'}$$

whenever $\rho < \rho_0(\epsilon)$ with $x' \in \mathscr{S}$. (To avoid repetition, we will take it as understood that all asymptotic expansions involving a parameter $x \in \mathscr{S}$ hold for a.e. $x \in \mathscr{S}$, the expansions being uniform in x, holding as $x' \overset{\mathscr{S}}{\to} x$, i.e., as x is approached through points in $\mathscr{S}$, and with coefficients in the expansion being continuous a.e. on their domain.)

H4. One or the other of the class-conditional densities vanishes close to the boundary of $\mathscr{S}$. More precisely, let $\partial\mathscr{S} = \operatorname{cl}(\mathscr{S}) \cap \operatorname{cl}(\mathbb{R}^n \setminus \mathscr{S})$ denote the boundary of $\mathscr{S}$, and for $t \ge 0$, let $\bar{\mathscr{S}}_t \subset \mathscr{S}$ denote the set of points in $\mathscr{S}$ of distance no more than t from the boundary:

$$\bar{\mathscr{S}}_t = \{x \in \mathscr{S}: |x - \partial\mathscr{S}| \le t\}.$$

(Note, for example, $\bar{\mathscr{S}}_0 = \partial\mathscr{S}$.) Then there exists a $t_0 > 0$ such that for a.e. $x \in \bar{\mathscr{S}}_{t_0}$, either $f_1(x) = 0$ or $f_2(x) = 0$.

Example: (*Overlapping power-law densities*). Consider a one-dimensional feature space. Let $0 < x_0 < T$, and let N be a positive integer. Then

$$f_1(x) = \begin{cases} \dfrac{N+2}{T^{N+2}}(x - x_0)^{N+1}, & \text{if } x_0 \le x \le x_0 + T, \\ 0, & \text{otherwise,} \end{cases}$$

$$f_2(x) = \begin{cases} \dfrac{N+2}{T^{N+2}}(T - x)^{N+1}, & \text{if } 0 \le x \le T, \\ 0, & \text{otherwise,} \end{cases}$$

define class-conditional probability densities that satisfy Hypotheses H1–H4. ■

Remarks: Hypothesis H1 is relatively innocuous, but does preclude discrete distributions from this analysis. It also relegates "ties" to zero-probability events.

Hypothesis H2 arises out of a uniformity requirement in our proof. In particular, this excludes many standard distributions, such as mixtures of normal distributions, whose support is infinite. In practice however, these cases

[3] Here and elsewhere, with respect to Lebesgue measure.

can be well approximated by problems that satisfy Hypothesis H2 by truncating the tails of the original distribution, i.e., replacing the given distribution by one with compact support that agrees with the original distribution everywhere except the tails.

The requisite smoothness in the class-conditional densities mandated by Cover's results for the one-dimensional case is incorporated in Hypothesis H3. Additional smoothness constraints on the class-conditional densities f_j can lead to simple expressions for the functions $f_{j,k}(\Omega, \boldsymbol{x})$ as we see in the following.

Example: (Taylor series). We assert that if the functions f_j possess a convergent Taylor series everywhere then they will have asymptotic expansions of the form (5). First, we introduce some notation to simplify subsequent expressions. By a *multiset*, I, of indices from $\{1,\cdots,n\}$ we mean a collection of indices where order does not matter and indices can repeat. Let $\boldsymbol{\xi} = (\xi_1,\cdots,\xi_n)$ be any point in $\mathbb{R}^n$. For any multiset $I = (i_1,\cdots,i_k)$ of indices from $\{1,\cdots,n\}$ denote $\xi_I = \prod_{j=1}^k \xi_{i_j}$. For $k = 1,2,\cdots$, let $\mathscr{I}_k$ denote the family of all multisets of cardinality k from $\{1,\cdots,n\}$. If the functions f_j possess uniformly bounded partial derivatives of order $N + 1$ for all $\boldsymbol{x} \in \mathscr{S}$, then we can expand f_j locally in terms of a Taylor series with remainder. Denoting $\boldsymbol{\xi} = \boldsymbol{x}' - \boldsymbol{x}$, we have

$$f_j(\boldsymbol{x}') = f_j(\boldsymbol{x}) + \sum_{k=1}^{N} \sum_{I \in \mathscr{I}_k} \frac{f_{j,I}(\boldsymbol{x})}{k!} \xi_I + o\left(\max_i |\xi_i|^N\right) \qquad (\boldsymbol{x}' \overset{\mathscr{S}}{\to} \boldsymbol{x}) \quad (6)$$

where, for any multiset $I = (i_1,\cdots,i_k)$, we define

$$f_{j,I}(\boldsymbol{x}) = \frac{\partial^k}{\partial x_{i_1} \cdots \partial x_{i_k}} f_j(\boldsymbol{x}). \quad (7)$$

In spherical coordinates let us represent $\boldsymbol{\xi} = (\rho, \Omega) = (\rho, \phi_1,\cdots,\phi_{n-1})$ where the angles ϕ_i take values in the following ranges: $-\pi/2 \leq \phi_i \leq \pi/2$ for $i \in \{1,\cdots,n-2\}$, and $0 \leq \phi_{n-1} \leq 2\pi$. We then have

$$\xi_i = \rho \sin \phi_{n-i+1} \prod_{l=i}^{n-1} \cos \phi_{n-l}, \qquad i = 1,\cdots,n,$$

where, for $i = 1$, we define $\sin \phi_n \equiv 1$. With this transformation we can rewrite (6) in the form (5) where, as per our convention, $\boldsymbol{x}' - \boldsymbol{x} = (\rho, \Omega)$, and

$$f_{j,k}(\Omega, \boldsymbol{x}) = \frac{1}{k!} \sum_{I \in \mathscr{I}_k} f_{j,I}(\boldsymbol{x}) \prod_{i \in I} \sin \phi_{n-i+1} \prod_{l=i}^{n-1} \cos \phi_{n-l} \qquad k = 1,\cdots,N. \quad (8)$$

Note in particular that the functions $f_{j,k}(\Omega, \boldsymbol{x})$ are continuous on $S_{n-1} \times \mathscr{S}$. ■

The following assertion is hence proved.

Assertion 1: The class-conditional densities f_j will satisfy uniform asymptotic expansions of the form (5) if they possess uniformly bounded partial derivatives up to order $N + 1$ on their probability one support.

To prove a converse to the assertion will call for the imposition of Tauberian-style conditions to allow for differentiation of the asymptotic series expansion. The uniformity of the expansion does help, however, at least in establishing the existence of the first partial derivatives of f_j.

Assertion 2: The coefficients $f_{j,1}$ in the expansion satisfy

$$f_{j,1}(\Omega, \boldsymbol{x}) = -f_{j,1}(-\Omega, \boldsymbol{x})$$

for a.e. $\boldsymbol{x}$. In particular, under Hypothesis H3, the class-conditional densities f_j have continuous first-order partial derivatives a.e. on their probability one support.

Proof: Let $\boldsymbol{x}$ and $\boldsymbol{x}'$ denote two distinct points with $\rho = |\boldsymbol{x} - \boldsymbol{x}'|$ and $\Omega = (\boldsymbol{x}' - \boldsymbol{x})/\rho$. Then by (5), as $\rho \to 0$,

$$f_j(\boldsymbol{x}') = f_j(\boldsymbol{x}) + f_{j,1}(\Omega, \boldsymbol{x})\rho + o(\rho),$$
$$f_j(\boldsymbol{x}) = f_j(\boldsymbol{x}') + f_{j,1}(-\Omega, \boldsymbol{x}')\rho + o(\rho).$$

Although the remainders of these two expressions, designated by "$o(\rho)$," need not be equal, the uniformity requirement in Hypothesis H3 implies that the sum of the above two expressions satisfies

$$0 = \left(f_{j,1}(\Omega, \boldsymbol{x}) + f_{j,1}(-\Omega, \boldsymbol{x}')\right)\rho + o(\rho).$$

After dividing through by ρ, and taking the limit $\rho \to 0$, we obtain

$$f_{j,1}(\Omega, \boldsymbol{x}) = -f_{j,1}(-\Omega, \boldsymbol{x}),$$

where we have used the continuity of $f_{j,1}(\Omega, \boldsymbol{x})$ with respect to $\boldsymbol{x}$. Now let $\boldsymbol{e}_k$ denote the kth basis vector in the current coordinate system. Then, by Hypothesis H3,

$$\begin{aligned} f_j(\boldsymbol{x} + \epsilon \boldsymbol{e}_k) &= f_j(\boldsymbol{x}) + f_{j,1}(\operatorname{sgn}(\epsilon)\boldsymbol{e}_k, \boldsymbol{x})|\epsilon| + o(\epsilon) \\ &= f_j(\boldsymbol{x}) + \operatorname{sgn}(\epsilon) f_{j,1}(\boldsymbol{e}_k, \boldsymbol{x})|\epsilon| + o(\epsilon) \\ &= f_j(\boldsymbol{x}) + f_{j,1}(\boldsymbol{e}_k, \boldsymbol{x})\epsilon + o(\epsilon). \end{aligned}$$

The partial derivative in the direction of $\boldsymbol{e}_k$ is found to be

$$\frac{\partial f_j}{\partial x_k}(x) = \lim_{\epsilon \to 0} \frac{f_j(\boldsymbol{x} + \epsilon \boldsymbol{e}_k) - f_j(\boldsymbol{x})}{\epsilon} = f_{j,1}(\boldsymbol{e}_k, \boldsymbol{x}).$$

Thus the n first-order partial derivatives exist and are continuous (by the assumed continuity of $f_{j,1}$) a.e. on the probability one support of f_j. ■

We will not pursue the issue of a Tauberian converse to Assertion 1 any further.

It is readily seen that Hypothesis H3 implies as asymptotic expansion for the mixture density f of the form

$$f(\boldsymbol{x}') = f(\boldsymbol{x}) + \sum_{k=1}^{N} f_k(\Omega, \boldsymbol{x})\rho^k + o(\rho^N) \qquad (\rho \overset{\mathscr{S}}{\to} 0) \quad (9)$$

where

$$f_k(\Omega, \boldsymbol{x}) = P_1 f_{1,k}(\Omega, \boldsymbol{x}) + P_2 f_{2,k}(\Omega, \boldsymbol{x}), \qquad k = 1,2,\cdots,N. \quad (10)$$

Finally, Hypothesis H4 is introduced to avoid technical complications arising from boundary effects. The point here is essentially this: if both $\boldsymbol{x}$ and its nearest neighbour $\boldsymbol{x}'$ happen to fall within $\bar{\mathscr{S}}_{t_0}$, then the classification will be correct (with probability one), and there is no contribution to R_m. Relaxing the requirement of Hypothesis H4 would necessitate placing smoothness constraints on the boundary $\mathscr{S}$, and would also result in more awkward expressions for the coefficients for R_m in an asymptotic expansion.

We are now ready to state the main result of this paper.

Theorem 1: Under Hypotheses H1–H4, there exists a unique set of constants $\{c_k\}$ such that

$$R_m = R_\infty + \sum_{k=2}^{N} c_k m^{-k/n} + o(m^{-N/n}) \qquad (m \to \infty).$$

Corollary 1: If, in Hypothesis H3, the asymptotic expansions (5) can be replaced by complete asymptotic series representations

$$f_j(\boldsymbol{x}') \sim f_j(\boldsymbol{x}) + \sum_{k=2}^{\infty} f_{j,k}(\Omega, \boldsymbol{x})\rho^k \qquad (\rho \overset{\mathscr{S}}{\to} 0),$$

then, with the remaining hypotheses intact, there exists a unique set of constants $\{c_k\}$ such that the following series is asymptotic to R_m:

$$R_m \sim R_\infty + \sum_{k=2}^{\infty} c_k m^{-k/n} \qquad (m \to \infty).$$

Remarks: The leading coefficient in the expansion is just the infinite-sample risk (see (4))

$$R_\infty = 2P_1P_2 \int_{\mathscr{S}} \frac{f_1(\boldsymbol{x})f_2(\boldsymbol{x})}{f(\boldsymbol{x})}\, d\boldsymbol{x}, \tag{11}$$

that was first derived by Cover and Hart [2]. We describe the evaluation of the succeeding coefficients c_k in the asymptotic expansion in the proof of the theorem. In particular, the coefficient c_2 evaluates to

$$c_2 = P_1P_2 \frac{\Gamma\left(1+\frac{2}{n}\right)\Gamma\left(1+\frac{n}{2}\right)^{1+2/n}}{n\pi^{1+n/2}} \int_{\mathscr{S}} \frac{f_1(\boldsymbol{x})f_2(\boldsymbol{x})}{[f(\boldsymbol{x})]^{1+2/n}} \cdot \Bigg[\Phi_{1,2}(\boldsymbol{x}) + \Phi_{2,2}(\boldsymbol{x}) - 2\Phi_2(\boldsymbol{x}) - \frac{n+2}{n(n+1)}\Phi_1(\boldsymbol{x})\{\Phi_{1,1}(\boldsymbol{x}) + \Phi_{2,1}(\boldsymbol{x}) - 2\Phi_1(\boldsymbol{x})\}\Bigg]\, d\boldsymbol{x}, \tag{12}$$

where, for $j \in \{1, 2\}$ and $k \geq 1$,

$$\Phi_k(\boldsymbol{x}) = \int_{S_{n-1}} \frac{f_k(\Omega, \boldsymbol{x})}{f(\boldsymbol{x})}\, d\Omega,$$

$$\Phi_{j,k}(\boldsymbol{x}) = \int_{S_{n-1}} \frac{f_{j,k}(\Omega, \boldsymbol{x})}{f_j(\boldsymbol{x})}\, d\Omega.$$

The above form of c_2 is somewhat more general than those presented in [7] and [10]. The form of c_2 is further simplified if it is assumed, in addition, that the class-conditional densities possess uniformly bounded partial derivatives of order $N + 1$; in this case the expression for c_2 can be cast in a form equivalent to those reported earlier. In general, under this smoothness assumption the forms of all the coefficients c_k are simplified and, in particular, explicit expressions can now be written for the c_k's involving only partial derivatives of the class-conditional densities. Expressions for the first seven coefficients are listed in the Appendix.

IV. Asymptotic Analysis

The proof of Theorem 1 proceeds in several stages. First, we derive an exact integral expression for R_m in the form

$$R_m = \int_{\mathscr{S}\times\mathscr{S}} ge^{-mh}, \tag{13}$$

where g and h are nonnegative functions in Euclidean $2n$-space. For large m, this integral appears to be in a form amenable to Laplace's asymptotic method. Recall that for one-dimensional integrals of the above form, Laplace's method (cf. Erdélyi [8, pp. 36–39], for instance) asserts that for large m the dominant contribution to the integral arises from a neighborhood of the point where h has a (discrete) minimum. (If h has more than one discrete minimum, then the domain of integration can be partitioned so that each subdomain contains only one minimum.) Furthermore, if g and h can be represented as asymptotic power series in a neighborhood of this minimum, then the integral itself may be represented as an asymptotic power series in reciprocal (noninteger, in general) powers of m, the coefficients of the asymptotic expansion being independent of the size of the neighborhood assumed around the minimum of h. The method can be extended to multidimensional integrals where, for instance, h has an interior minimum in the domain of integration (cf. Fulks and Sather [11]).

The evaluation of (13) by this asymptotic technique, however, is complicated by the fact that the minima of h (defined over $\mathbb{R}^{2n}$) occur on a continuum of points in a linear manifold consisting of the intersection of an n-dimensional linear subspace with the domain of integration $\mathscr{S}\times\mathscr{S}$. This has two consequences: (i) standard results on Laplace integrals when h has discrete minima cannot be carried over *in toto* to the case when h has a continuum of minima; (ii) contributions to the integral from a continuum of points where h its minima at and near the boundary of the domain of integration pose particular difficulties in evaluation, and depend, in general, on how smooth the boundary is.

The second difficulty is finessed by Hypothesis H4 which eliminates any contribution to the classifier's error rate in the boundary of the domain of integration. We reiterate that while we introduce Hypothesis H4 only to avoid certain ensuing analytical complications, the constraint has an artificial character. In Section V we present examples where Hypothesis H4 is flouted while asymptotic expansions for R_m of the form governed by the theorem continue to hold. This suggests that the hypothesis can be weakened or done away with altogether.

The principal tool in our resolution of the first difficulty is a technical result of Fulks and Sather which allows the replacement of the multiple integral (13) by a single Stieltjes integral. The key to the proof is the asymptotic representation of the functions g and h in a generalized cylindrical coordinate system. The asymptotic series expansion of the finite-sample risk R_m can now be obtained, forfending both difficulties, by identifying a dominant component to the integral (13) over a cylindrical domain, and evaluating the contributions from the interior and the boundary separately. The technique in evaluating each of these integral contributions follows the analysis of Fulks and Sather [11] for the discrete minimum case to effectively show that the Laplace method can be extended to this class of integrals. The coefficients of the asymptotic expansion for R_m will be explicitly evaluated in the course of the proof.

A. Technical Lemmas

We begin by presenting (without proof) two technical results in order to avoid breaking up the flow of the proof subsequently. The first result is due to Fulks and Sather [11].

Lemma 1: Let h be a measurable function on a set $\mathscr{R}$ in $\mathbb{R}^M$ taking values in the possibly infinite interval $\{a < s < b\}$. Let g be defined and integrable over $\mathscr{R}$, and define the function $G(z)$ by

$$G(z) = \int_{\{h \le z\}} g\,dx.$$

If $F(s)$ is a continuous function defined on $\{a < s < b\}$, and such that $F(h)g$ is integrable over $\mathscr{R}$, then

$$\int_{\mathscr{R}} F(h) g\,dx = \int_a^b F(s)\,dG(s).$$

The second result we will need is a classical result due to Watson [12], frequently referred to in the literature as *Watson's lemma.*

Lemma 2: Let $G(s)$ be a function of the positive real variable s, such that

$$G(s) \sim \sum_{k=0}^{\infty} \eta_k s^{(k+\tau-\mu)/\mu} \qquad (s \to 0),$$

where τ and μ are positive constants. Then

$$\int_0^\infty e^{-ms} G(s)\,ds \sim \sum_{k=0}^{\infty} \Gamma\left(\frac{k+\tau}{\mu}\right) \frac{\eta_k}{m^{(k+\tau)/\mu}} \qquad (m \to \infty)$$

provided that the integral converges throughout its range for all sufficiently large m.

B. Integral Representation of the Risk

Let $X' = \arg\min_{1 \le i \le m} |X - X^{(i)}|$ denote the feature vector in the reference sample Σ_m that is closest to the random test vector X, and let θ' be the class label associated with X'. Then

$$\begin{aligned} R_m &= \int_{\mathscr{S}} P[\theta \ne \theta' | X = x] f(x)\,dx \\ &= \int_{\mathscr{S}\times\mathscr{S}} P[\theta \ne \theta' | X' = x', X = x] \\ &\quad \cdot f_m(x'|x) f(x)\,dx'\,dx, \end{aligned}$$

where $f_m(x'|x)$ denotes the conditional density of X' given $X = x$.

We now obtain an explicit expression for $f_m(x'|x)$. The event $X' = x'$ occurs if one of the reference feature vectors $X^{(j)}$ assumes the value x' and every other feature vector $X^{(k)}$, $k \ne j$, assumes a value outside $B(\rho, x)$, the (closed) ball of radius $\rho = |x' - x|$ at x. (Ties occur with zero probability.) Because of the independent nature of the training set, the latter may occur with $j = 1, 2, \cdots, m$. We thus obtain,

$$f_m(x'|x) = \sum_{j=1}^{m} \left(\prod_{k \ne j} P[X^{(k)} \notin B(|x' - x|, x)] \right) f(x').$$

For $\rho \ge 0$ and $x \in \mathbb{R}^n$, let $\psi(\rho, x)$ denote the probability that a feature vector $Y \in \mathbb{R}^n$ drawn from the mixture distribution $F(y)$ lies in the ball of radius ρ at x:

$$\psi(\rho, x) = P[Y \in B(\rho, x)] = \int_{B(\rho, x)} f(y)\,dy. \tag{14}$$

It follows that

$$f_m(x'|x) = m(1 - \psi(|x' - x|, x))^{m-1} f(x').$$

We thus obtain the desired integral representation

$$R_m = m \int_{\mathscr{S}\times\mathscr{S}} g(x', x) e^{-m h(x', x)}\,dx'\,dx, \tag{15}$$

where

$$g(x', x) = \frac{P[\theta \ne \theta' | X' = x', X = x] f(x') f(x)}{1 - \psi(\rho, x)}, \tag{16}$$

and

$$h(x', x) = -\log(1 - \psi(\rho, x)). \tag{17}$$

Note the useful feature that h is independent of Ω. Moreover, the absolute continuity of the mixture distribution (Hypothesis H1) guarantees that h attains its minimum value of zero at $\rho = 0$, i.e., when $x' = x$. Furthermore, as f is bounded away from zero uniformly on $\mathscr{S}$ (Hypothesis H2), h strictly increases with $\rho(x', x)$, and, in particular, is bounded away from zero for $\rho > 0$. More formally, for each $x \in \mathscr{S}$ and $\rho_0 > 0$, there exists a constant $a > 0$, uniform with respect to x, such that $h(x', x) \ge a$ if $\rho \ge \rho_0$. Thus the minima of h consist of that portion of the n-dimensional linear subspace $x' - x = 0$ that lies within the $2n$-dimensional domain of integration, $(x', x) \in \mathscr{S} \times \mathscr{S}$.

C. Induced Asymptotic Expansions

We now compute asymptotic expansions for the functions g and h in (15). Letting $\boldsymbol{y} = \boldsymbol{x} + \boldsymbol{r}$, the integrand of (14) is first expanded, via the asymptotic expansions for the mixture density (cf. Hypothesis H3), about the point $\boldsymbol{x}$, as

$$f(\boldsymbol{y}) = f(\boldsymbol{x}) + \sum_{k=1}^{N} f_k(\Omega, \boldsymbol{x}) r^k + o(r^N) \qquad (r \overset{\mathscr{S}}{\to} 0),$$

where, $r = |\mathbf{r}|$, and $\Omega = \mathbf{r}/r \in S_{n-1}$. Writing $d\mathbf{r} = r^{n-1}\, dr d\Omega$, we can now express $\psi(\rho, \boldsymbol{x})$ as the following n-dimensional integral in spherical coordinates:

$$\begin{aligned}\psi(\rho, \boldsymbol{x}) &= \int_{B(\rho,0)} f(\boldsymbol{x} + \boldsymbol{r})\, d\boldsymbol{r} \\ &= \int_{S_{n-1}} \int_0^{\rho} f(\boldsymbol{x} + (r, \Omega)) r^{n-1}\, dr d\Omega \\ &= \rho^n \sum_{k=0}^{N} \psi_k(\boldsymbol{x}) \rho^k + o(\rho^{N+n}) \qquad (\rho \overset{\mathscr{S}}{\to} 0).\end{aligned} \tag{18}$$

Letting

$$V_n = \frac{\pi^{n/2}}{\Gamma(1 + n/2)}$$

denote the volume of the n-dimensional unit sphere, we find that

$$\psi_0(\boldsymbol{x}) = V_n f(\boldsymbol{x}), \tag{19}$$

and

$$\psi_k(\boldsymbol{x}) = \frac{1}{k+n} \int_{S_{n-1}} f_k(\Omega, \boldsymbol{x})\, d\Omega, \qquad k = 1, \cdots, N. \tag{20}$$

Note that, as a consequence of the skew symmetry of $f_1(\Omega, \boldsymbol{x})$ (Assertion 2), it follows that

$$\psi_1(\boldsymbol{x}) = 0. \tag{21}$$

A tabulation of the first few functions $\psi_k(\boldsymbol{x})$ can be found in Table I in the Appendix.

After representing the logarithm by its Taylor series in (17), we obtain the following asymptotic expansion for $h(\boldsymbol{x}', \boldsymbol{x})$,

$$\begin{aligned}h(\boldsymbol{x}', \boldsymbol{x}) &= \sum_{k=1}^{\infty} \frac{\psi(\rho, \boldsymbol{x})^k}{k} \\ &= \rho^n \sum_{k=0}^{N} h_k(\boldsymbol{x}) \rho^k + o(\rho^{N+n}), \qquad (\rho \overset{\mathscr{S}}{\to} 0),\end{aligned} \tag{22}$$

the functions $h_k(\boldsymbol{x})$ in the expansion being continuous and obtained by collecting like powers of ρ in the Taylor series expansion for the logarithm. In particular, Table II in the Appendix shows the form of these coefficients for $k \le 6$. Note that $h_k(\boldsymbol{x}) = \psi_k(\boldsymbol{x})$ if $k < n$ and that the coefficient $V_n f(\boldsymbol{x})$ of the leading term in (22) is strictly positive and bounded away from zero a.e. $\boldsymbol{x} \in \mathscr{S}$ as a consequence of Hypothesis H2. Note also that, as expected, the expansion (22) implies that the infimum of h over $\mathscr{S} \times \mathscr{S}$ is indeed zero.

We now develop a corresponding asymptotic expansion for the function g. It is easy to see by a simple independence argument that

$$\begin{aligned}\boldsymbol{P}[\theta \ne \theta' | \boldsymbol{X}' = \boldsymbol{x}', \boldsymbol{X} = \boldsymbol{x}] &= \hat{P}(1|\boldsymbol{x})\hat{P}(2|\boldsymbol{x}') + \hat{P}(1|\boldsymbol{x}')\hat{P}(2|\boldsymbol{x}) \\ &= \frac{P_1 P_2}{f(\boldsymbol{x}) f(\boldsymbol{x}')} (f_1(\boldsymbol{x}) f_2(\boldsymbol{x}') - f_1(\boldsymbol{x}') f_2(\boldsymbol{x})).\end{aligned} \tag{23}$$

Thus, from the asymptotic expansions in Hypothesis H3,

$$\begin{aligned}\boldsymbol{P}[\theta \ne \theta' | \boldsymbol{X}' = \boldsymbol{x}', \boldsymbol{X} = \boldsymbol{x}] f(\boldsymbol{x}') f(\boldsymbol{x}) &= \alpha_0(\boldsymbol{x}) + \sum_{k=1}^{N} \alpha_k(\Omega, \boldsymbol{x}) \rho^k + o(\rho^N),\end{aligned} \tag{24}$$

where

$$\begin{aligned}\alpha_0(\boldsymbol{x}) &= 2 P_1 P_2 f_1(\boldsymbol{x}) f_2(\boldsymbol{x}), \\ \alpha_k(\Omega, \boldsymbol{x}) &= P_1 P_2 (f_1(\boldsymbol{x}) f_{2,k}(\Omega, \boldsymbol{x}) + f_2(\boldsymbol{x}) f_{1,k}(\Omega, \boldsymbol{x})), \\ & \qquad k = 1, \cdots, N.\end{aligned} \tag{25}$$

For ρ sufficiently small, so that $\psi(\rho, \boldsymbol{x}) < 1$, the denominator of g may be expanded via

$$\frac{1}{1 - \psi(\rho, \boldsymbol{x})} = \sum_{k=0}^{\infty} \psi(\rho, \boldsymbol{x})^k.$$

Using the asymptotic expansion (18) in the geometric series above we get another asymptotic expansion in integer powers of ρ, and substituting this in conjunction with (24) in (16) results in the asymptotic expansion

$$g(\boldsymbol{x}', \boldsymbol{x}) = \sum_{k=0}^{N} g_k(\Omega, \boldsymbol{x}) \rho^k + o(\rho^N) \qquad (\rho \overset{\mathscr{S}}{\to} 0), \tag{26}$$

where, for instance, for $k < n$ the coefficient functions $g_k(\Omega, \boldsymbol{x})$ are independent of n; in particular, for all dimensionalities,

$$g_0(\boldsymbol{x}) = \alpha_0(\boldsymbol{x}),$$

and for $k < n$,

$$g_k(\Omega, \boldsymbol{x}) = \alpha_k(\Omega, \boldsymbol{x}).$$

The coefficients g_k, for $k \le 6$, are listed in Table III in the Appendix.

D. Laplace's Method

Now let us return to a consideration of (15). Recall that $h(\boldsymbol{x}', \boldsymbol{x})$ attains its minimum value of zero when $\boldsymbol{x}' = \boldsymbol{x}$. Consequently, the dominant contribution to the integral $\int_{\mathscr{S} \times \mathscr{S}} g e^{-mh}$ should occur in the immediate vicinity of the linear subspace $\boldsymbol{x}' = \boldsymbol{x}$. For $t > 0$, define the family of "cylinder" sets

$$C_t = \{(\boldsymbol{x}', \boldsymbol{x}) \in \mathbb{R}^{2n}: |\boldsymbol{x}' - \boldsymbol{x}| = \rho \le t\} \cap \mathscr{S}^2.$$

This is schematically the cylindrical area along the diagonal in Fig. 1. We can then partition the integral contribution to R_m into two parts:

$$R_m = m\int_{C_t} ge^{-mh} + m\int_{\mathscr{S}^2 \setminus C_t} ge^{-mh}.$$

We first show that for any fixed $t > 0$, the integral contribution from outside C_t is subdominant. Recall that the integral in (15) represents a probability and is hence convergent for every m. Following the argument at the conclusion of Section IV B, for any $t > 0$, there exists a fixed $a > 0$ such that $h(\boldsymbol{x}', \boldsymbol{x}) \ge a$ if $(\boldsymbol{x}', \boldsymbol{x}) \in \mathscr{S}^2 \setminus C_t$. Thus, $mh = (m-1)h + h \ge (m-1)a + h$, in $\mathscr{S}^2 \setminus C_t$. Since the functions g and h are non-negative everywhere, we have

$$\begin{aligned}&\left| m\int_{\mathscr{S}^2 \setminus C_t} ge^{-mh} \right| \\ &\quad \le me^{-(m-1)a} \int_{\mathscr{S}^2 \setminus C_t} g(\boldsymbol{x}', \boldsymbol{x}) e^{-h(\boldsymbol{x}', \boldsymbol{x})}\, d\boldsymbol{x}'\, d\boldsymbol{x} \\ &\quad = O(me^{-ma}) \qquad (m \to \infty)\end{aligned}$$

as the integral on the right is bounded above by $R_1 \le 1$.

Now recall that by Hypothesis H4, there exists a $t_0 > 0$ such that one or the other of the class-conditional densities f_j is identically zero at a.e. points in a set $\bar{\mathscr{S}}_{t_0}$ of points in $\mathscr{S}$ whose distance from the boundary $\partial\mathscr{S}$ of $\mathscr{S}$ is no more than t_0. Now choose $0 < t \le t_0/2$, and define the set

$$\mathscr{S}_t = \mathscr{S} \setminus \bar{\mathscr{S}}_t = \{\boldsymbol{x} \in \mathscr{S}: |\boldsymbol{x} - \partial\mathscr{S}| > t\}.$$

(See Fig. 2.) We now partition C_t into the sets

$$Q_t = C_t \cap (\mathscr{S} \times \mathscr{S}_t), \qquad \bar{Q}_t = C_t \cap \left(\mathscr{S} \times \bar{\mathscr{S}}_t\right).$$

Clearly, $Q_t \cap \bar{Q}_t = \varnothing$, and $Q_t \cup \bar{Q}_t = C_t$. We now further partition the dominant integral contribution to R_m according to whether $\boldsymbol{x}$ takes values in $\mathscr{S}_t$ or $\boldsymbol{x}$ takes values in $\bar{\mathscr{S}}_t$:

$$m\int_{C_t} ge^{-mh} = m\int_{Q_t} ge^{-mh} + m\int_{\bar{Q}_t} ge^{-mh} \equiv I_m + J_m,$$

where I_m and J_m denote the two integrals, respectively. Note that I_m is the part of the dominant contribution which arises from the interior points, while J_m is the part which arises from the boundary points. We evaluate these in turn.

Boundary Contribution: If $(\boldsymbol{x}', \boldsymbol{x}) \in \bar{Q}_{\underline{t}}$, then clearly $\boldsymbol{x} \in \bar{\mathscr{S}}_t$ by definition of $\bar{Q}_t = C_t \cap (\mathscr{S} \times \bar{\mathscr{S}}_t)$. Furthermore, in C_t we have $|\boldsymbol{x}' - \boldsymbol{x}| \le t$ so that by the triangle inequality we have $|\boldsymbol{x}' - \partial\mathscr{S}| \le |\boldsymbol{x}' - \boldsymbol{x}| + |\boldsymbol{x} - \partial\mathscr{S}| \le 2t \le t_0$ by choice of t. Consequently, both $\boldsymbol{x}$ and $\boldsymbol{x}'$ will lie in $\bar{S}_{t_0}$. It follows from Hypothesis H4 that for a.e. $(\boldsymbol{x}', \boldsymbol{x}) \in \bar{Q}_t$, $f_j(\boldsymbol{x}') = f_k(\boldsymbol{x}) = 0$ for some $j, k \in \{1, 2\}$. Now consider representation (16) for g. It is clear from (23) that

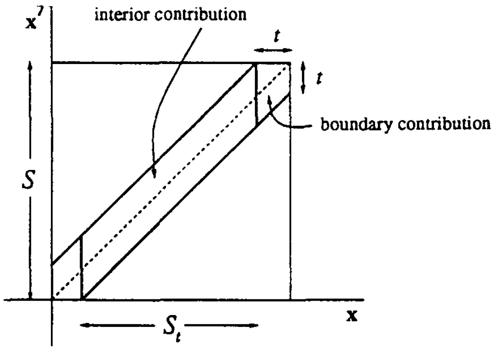

Fig. 1. A schematic of the cylinder set C_t, and the interior and boundary contributions to the dominant integral.

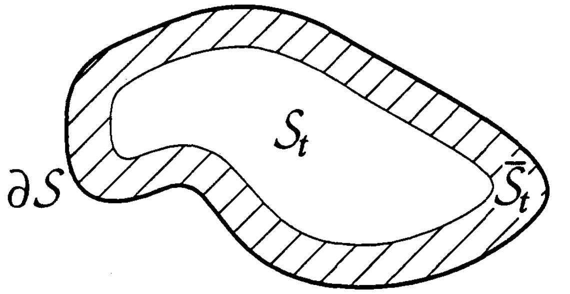

Fig. 2. A schematic of the support set $\mathscr{S}$, indicating its boundary $\partial\mathscr{S}$, the interior set $\mathscr{S}_t$, and the boundary set $\bar{\mathscr{S}}_t$ (hatched).

$$\boldsymbol{P}[\theta \ne \theta' | X' = \boldsymbol{x}', X = \boldsymbol{x}] = 0 \qquad \left(\boldsymbol{x}' \in \bar{\mathscr{S}}_{t_0}, \boldsymbol{x} \in \bar{\mathscr{S}}_{t_0}\right).$$

It follows that g is identically zero over $\bar{Q}_t$, and hence,

$$J_m = 0.$$

Interior Contribution: For every $\epsilon > 0$ define the functions h_+ and h_- on the domain $\mathbb{R}^{2n}$ by

$$h_\pm(\boldsymbol{x}', \boldsymbol{x}) \equiv \rho^n \sum_{k=0}^{N} h_k(\boldsymbol{x})\rho^k \pm \epsilon\rho^{N+n}, \qquad \boldsymbol{x}' - \boldsymbol{x} = (\rho, \Omega),$$

where the continuous functions $h_k(\boldsymbol{x})$ are as defined in (22). By uniformity of the asymptotic expansions (22) and (26), choose $0 < t \le t_0/2$ small enough so that:

(a) $|h(\boldsymbol{x}', \boldsymbol{x}) - \rho^n \Sigma_{k=0}^N h_k(\boldsymbol{x})\rho^k| < \epsilon\rho^{N+n}$ for a.e. $\boldsymbol{x} \in \mathscr{S}$ and $\rho < t$;

(b) $|g(\boldsymbol{x}', \boldsymbol{x}) - \Sigma_{k=0}^N g_k(\Omega, \boldsymbol{x})\rho^k| < \epsilon\rho^N$ for a.e. $\boldsymbol{x} \in \mathscr{S}$ and $\rho < t$;

(c) for a.e. $\boldsymbol{x} \in \mathscr{S}$, the functions $h_\pm(\boldsymbol{x}', \boldsymbol{x})$ are bounded away from zero and strictly increasing in ρ for $\{0 < \rho \le t\}$.

The last condition is made possible as $h_0(\boldsymbol{x}) = V_n f(\boldsymbol{x})$ is uniformly bounded away from zero on $\mathscr{S}$ by Hypothesis H2, and the other coefficients $h_k(\boldsymbol{x})$, $k \ge 1$, are bounded.

Now define I_m^+ and I_m^- by

$$I_m^{\pm} = m\int_{Q_t} g(x', x)e^{-mh_{\pm}(x',x)}\,dx'\,dx,$$

and note that

$$I_m^+ \leq I_m \leq I_m^-$$

as $g \geq 0$.

Let us first estimate I_m^+. For $\delta > 0$ chosen suitably small, define the subset of interior points $R_\delta \subset Q_t$ by

$$R_\delta = \{(x', x) \in \mathscr{S} \times \mathscr{S}_t \colon h_+(x', x) \leq \delta\}.$$

We then have

$$\begin{aligned} I_m^+ &= m\int_{R_\delta} ge^{-mh_+} + m\int_{Q_t \setminus R_\delta} ge^{-mh_+} \\ &= m\int_{R_\delta} ge^{-mh_+} + O(me^{-ma'}) \end{aligned}$$

for a choice of $a' > 0$; the second integration is subdominant by an argument similar to the one carried out earlier. Now define the function $G(s)$ by

$$G(s) = \int_{\{h_+ < s\} \cap R_\delta} g(x', x)\,dx'\,dx.$$

Now $0 < e^{-mh_+} < 1$ is bounded in R_δ as h_+ is bounded there, so that ge^{-mh_+} is integrable over R_δ. Invoking Lemma 1 and integrating by parts, we now have

$$\begin{aligned} I_m^+ &= m\int_0^\delta e^{-ms}\,dG(s) + O(me^{-ma'}) \\ &= m^2\int_0^\delta e^{-ms}G(s)\,ds + me^{-m\delta}G(\delta) + O(me^{-ma'}) \\ &= m^2\int_0^\delta e^{-ms}G(s)\,ds + O(me^{-ma''}) \qquad (m \to \infty), \end{aligned} \tag{27}$$

for a positive constant $a'' = \min\{\delta, a'\}$.

We now estimate $G(s)$. For $0 \leq s \leq \delta$, we first solve the equation $s = h_+(x'(\rho, \Omega, x), x)$ for $\rho = \rho(s, x)$. Note that a unique solution exists which is continuous in x and independent of Ω as h_+ is increasing in ρ and independent of Ω. We hence need to solve the equation

$$s^{1/n} = \rho\left(\sum_{k=0}^{N} h_k(x)\rho^k + \epsilon\rho^N\right)^{1/n}$$

for ρ. By condition (c) above, $s^{1/n}$ is an analytic function of ρ $(0 \leq \rho \leq t)$ for each x. Therefore, we may expand $\rho(s, x)$ as a Taylor series with remainder, obtaining

$$\begin{aligned} \rho(s, x) = &\sum_{k=0}^{N} \Upsilon_k(x)s^{(k+1)/n} + \epsilon\Upsilon_N'(x)s^{(N+1)/n} \\ &+ \Upsilon_{N+1}(x, \epsilon, s)s^{(N+2)/n} \end{aligned} \tag{28}$$

where: each Υ_k (and Υ_N') depends only on the h_j's for $j \leq k$; Υ_k (and Υ_N') is independent of ϵ for $k \leq N$; and Υ_{N+1} is uniformly bounded for $x \in \mathscr{S}_t$, $0 \leq \epsilon \leq 1$, and $0 \leq s \leq \delta$. The leading coefficients are tabulated in Table IV in the Appendix.

Now recall that by choice of $\delta > 0$ small enough, within R_δ we can use condition (b) to write

$$g(x', x) = \sum_{k=0}^{N} g_k(\Omega, x)\rho^k - \epsilon g_N'(\rho, \Omega, x)\rho^N,$$

where $|g_N'(\rho, \Omega, x)| < 1$. We can now estimate $G(s)$:

$$\begin{aligned} G(s) &= \int_{R_s} g\,dx'\,dx \\ &= \int_{S_t}\int_{S_{n-1}}\int_0^{\rho(s,x)} g(x'(\rho, \Omega, x), x)\rho^{n-1}\,d\rho\,d\Omega\,dx \\ &= \int_{S_t}\int_{S_{n-1}}\int_0^{\rho(s,x)} \sum_{k=0}^{N} g_k(\Omega, x)\rho^{n+k-1}\,d\rho\,d\Omega\,dx \\ &\quad - \epsilon\int_{S_t}\int_{S_{n-1}}\int_0^{\rho(s,x)} g_N'(\rho, \Omega, x)\rho^{n+N-1}\,d\rho d\Omega\,dx. \end{aligned}$$

Because $g_N'(\rho, \Omega, x)$ is bounded, the integral over ρ in the second term can be expressed as

$$\int_0^{\rho(s,x)} g_N'(\rho, \Omega, x)\rho^{N+n-1}\,d\rho = \frac{1}{N+n}g_N''(s, \Omega, x)\rho(s, x)^{N+n},$$

where $g_N''(s, \Omega, x)$ is an appropriately defined function that satisfies $|g_N''(s, \Omega, x)| < 1$. Consequently,

$$\begin{aligned} G(s) = \int_{\mathscr{S}_t}\int_{S_{n-1}} \rho(s, x)^n \Bigg(&\sum_{k=0}^{N} \frac{g_k(\Omega, x)}{k+n}\rho(s, x)^k \\ &- \frac{\epsilon}{N+n}g_N''(s, \Omega, x)\rho(s, x)^N \Bigg)\,d\Omega\,dx. \end{aligned}$$

If (28) is inserted into this expression, then, after collecting terms proportional to like powers of s, we obtain an expression of the form

$$\begin{aligned} G(s) = \int_{\mathscr{S}_t}\int_{S_{n-1}} \Bigg(&s\sum_{k=0}^{N} \lambda_k(\Omega, x)s^{k/n} \\ &- \epsilon\lambda_N''(s, \Omega, x)s^{(N+n)/n} + o(s^{(N+n)/n}) \Bigg)\,d\Omega\,dx. \end{aligned}$$

Here, the terms $\lambda_k(\Omega, x)$ contain the sum of the coefficients of all terms proportional to $s^{(1+k/n)}$, i.e.,

$$\lambda_k(\Omega, x) = \sum_{j=0}^{k} \frac{g_j(\Omega, x)}{j+n}B_{k-j}^{j+n}(x)$$

for $k \leq N$, and

$$\lambda_N'(s, \Omega, x) = \frac{g_N''(s, \Omega, x)}{N+n}B_0^{N+n}(x).$$

The coefficients B_j^k correspond to the coefficients of ξ^j

in the expansion of $(\Sigma_{l=0}^{N}\Upsilon_l(x)\xi^l)^k$, and are hence generated by

$$B_j^k(x) = \frac{1}{j!}\frac{\partial^j}{\partial\xi^j}\left(\sum_{l=0}^{N}\Upsilon_l(x)\xi^l\right)^k\Bigg|_{\xi=0}.$$

The coefficients λ_k can now be computed used the prior estimated coefficients Υ_l. The Appendix contains a tabulation of λ_k, for $k \le 6$, in Table V.

Now note that as a consequence of Hypothesis H4 and the choice $t \le t_0/2$, all coefficients $g_j(\Omega, x)$, $j \ge 0$ in (26) are identically zero for all x in $\bar{\mathcal{S}}_t$. As these coefficients are combined linearly to form the λ_k's, this in turn implies that each $\lambda_k(\Omega, x)$ is identically zero for all x in $\bar{\mathcal{S}}_t$. Hence, the domain of the x integral can be extended from $\mathcal{S}_t$ to $\mathcal{S} = \mathcal{S}_t \cup \bar{\mathcal{S}}_t$. Thus, we obtain,

$$\begin{aligned} G(s) &= \int_{\mathcal{S}}\int_{S_{n-1}}\left(\sum_{k=0}^{N}\lambda_k(\Omega, x)s^{1+k/n}\right.\\ &\qquad \left. - \epsilon\lambda_N''(s, \Omega, x)s^{1+N/n} + o(s^{1+N/n})\right) d\Omega\, dx\\ &= \sum_{k=0}^{N}\eta_k s^{1+k/n} - \epsilon\eta_N' s^{1+N/n} + o(s^{1+N/n}), \end{aligned}$$

where

$$\eta_k = \int_{\mathcal{S}}\int_{S_{n-1}}\lambda_k(\Omega, x)\, d\Omega\, dx, \qquad k = 0, \cdots, N,$$

$$\eta_N'' = \int_{\mathcal{S}}\int_{S_{n-1}}\lambda_N''(s, \Omega, x)\, d\Omega\, dx.$$

Note that

$$\eta_0 = 2\int_{\mathcal{S}}\frac{P_1P_2f_1(x)f_2(x)}{f(x)}\, dx, \tag{29}$$

$$\eta_1 = 0, \tag{30}$$

where we have invoked Assertion 2, (19), and (21).

We can now estimate I_m^+ by substituting the asymptotic expansion for $G(s)$ in (27). Noting that the integral $\int_\delta^\infty e^{-ms}s^L\, ds$ is subdominant for any finite L, we can expand the region of integration in (27) to $0 < s < \infty$ with an exponentially small correction factor. Applying Watson's lemma we finally obtain

$$I_m^+ = \sum_{k=0}^{N} c_k m^{-k/n} - \epsilon c_N' m^{-N/n} + o(m^{-N/n}),$$

where

$$c_k = \eta_k\Gamma(2 + k/n), \qquad k = 0, 1, \cdots, N.$$

Invoking (29) and (30) we obtain $c_0 = R_\infty$ [see (11)] and $c_1 = 0$. It is not difficult now to backtrack and write explicit expressions for the c_k's; the general form for c_2, for example is given in (12). The expressions for the c_k's simplify somewhat under a slightly stronger smoothness assumption than Hypothesis H3. This is described in the Appendix where tabulations of the first seven coefficients c_k are listed in Tables VI–XI under a slightly stronger smoothness assumption to keep the expressions from becoming too unwieldy.

An identical procedure yields

$$I_m^- = \sum_{k=0}^{N} c_k m^{-k/n} + \epsilon c_N'' m^{-N/n} + o(m^{-N/n}),$$

as h_- differs from h_+ only in the sign of ϵ. Recall that $I_m^+ \le I_m \le I_m^-$. Hence

$$\begin{aligned} I_m^+ - \sum_{k=0}^{N} c_k m^{-k/n} &\le I_m - \sum_{k=0}^{N} c_k m^{-k/n}\\ &\le I_m^- - \sum_{k=0}^{N} c_k m^{-k/n}. \end{aligned}$$

Also, collecting all the subdominant terms that we had dropped by the wayside gives

$$R_m = I_m + O(e^{-ma'''}) \qquad (m \to \infty)$$

for some fixed, positive a'''. Thus, by letting $m \to \infty$, we get

$$\begin{aligned} -\epsilon c_N' &\le \liminf\left\{\left(R_m - \sum_{k=0}^{N} c_k m^{-k/n}\right)m^{N/n}\right\}\\ &\le \limsup\left\{\left(R_m - \sum_{k=0}^{N} c_k m^{-k/n}\right)m^{N/n}\right\} \le \epsilon c_N'', \end{aligned}$$

the inequalities holding for every $\epsilon > 0$. Use $c_0 = R_\infty$ and $c_1 = 0$, and allow $\epsilon \to 0$ to complete the proof of Theorem 1.

V. Discussion

In the preceding section, Theorem 1 was proved under a restrictive set of hypotheses so that the expansion coefficients c_2, c_3, etc., could be readily obtained. Unfortunately, in so doing, we have precluded many practical, well behaved, distributions; mixtures of normal distributions, for instance, violate Hypothesis H2. In this section we present evidence that suggests that the asymptotic convergence of the finite-sample risk, described in the statements of the theorems, applies to a broader set of classification problems. In generalizing the theorem, we emphasize that although the risk may be expanded as in (2), the expressions for the expansion coefficients (c_2, etc.) will be more complex. (However, for some problems, the expressions for c_k in the Appendix may provide useful approximations to the actual coefficients.)

Of the four restrictions assumed, Hypothesis H4 appears to be the most artificial as it was introduced solely to avoid analytical complications at the boundary of the domain of integration. In the proof, it was only used as a justification for neglecting the boundary contribution J_m to the finite-sample risk. Consequently, it could be replaced by the weaker requirement that one of the two class-conditional densities tends to zero sufficiently fast at every boundary point so that J_m is exponentially subdominant with respect to the interior contribution I_m. Even this weaker condition, however, may be unnecessary as the following example illustrates.

Example: (*Triangular distributions*). Consider the one-dimensional triangular distribution over the unit interval,

$$f_1(x) = 2(1 - x),$$
$$f_2(x) = 2x.$$

These densities clearly violate Hypothesis H4 as both are nonzero in every neighborhood of the boundary. If the classes occur with equal prior probability, the following exact expression can be obtained for the finite sample risk:[4]

$$R_m = \frac{1}{3} + \frac{3m + 5}{2(m + 1)(m + 2)(m + 3)}.$$

If $m \gg 1$, then the finite-sample risk for more general prior probabilities ($0 < P_1, P_2 < 1$, with $P_1 \neq P_2$), can be approximated as

$$R_m = \frac{2P_1P_2}{(P_2 - P_1)^2}\left(1 + \frac{2P_1P_2}{P_2 - P_1}\log\frac{P_1}{P_2}\right) + \frac{3(1 - 3P_1P_2)}{8(P_1P_2)^2}\frac{1}{m^2} + o\left(\frac{1}{m^2}\right).$$

These results agree with the m^{-2} convergence rate predicted by Theorem 1 with $n = 1$. Note, however, that as all second derivatives of f_1 and f_2 are identically zero, the interior contribution to the risk (cf. expression (12) for coefficient c_2) vanishes. Consequently, it is the boundary term alone that yields the order m^{-2} rate of convergence. ■

The next example provides another illustration of how the form of the asymptotic expansion predicted in Theorem 1 may provide an accurate approximation to the m-sample risk even though Hypothesis H4 is not strictly satisfied.

Example: (*Trigonometric distributions*). Consider a multidimensional two-class problem where the class-conditional densities are given by

$$f_1(x) = \frac{1}{2^{n-1}\pi^n}\sin^2 x_1,$$
$$f_2(x) = \frac{1}{2^{n-1}\pi^n}\cos^2 x_1,$$

over the feature space $[-\pi, \pi]^n \subset \mathbb{R}^n$. If $P_1 = P_2 = 1/2$, then $R_\infty = 1/4$, and $R_B = (\pi - 2)/2\pi \approx 0.1817$. Clearly, this problem satisfies Hypotheses H1 through H3 while violating Hypothesis H4.

In Fig. 3 we present numerical estimates of R_m as a function of m and n for $n = 1$ (circular markers), through $n = 5$ (diamond markers). Each marker represents the fraction of "failures" of a large number (10^5–10^8) of Bernoulli trials. In each trial a pseudo-random reference sample of m labeled patterns is constructed in accord with the above probability densities. Then a single input vector is generated by the same process and is classified according to the reference sample. (In practice, these two steps are best carried out in reverse order so that only one reference pattern need be stored at a time.) A trial is regarded as a failure if the input is assigned to the wrong class by the reference sample. For each marker, an error bar, representing 95% certainty, is estimated using the De Moivre–Laplace limit theorem.

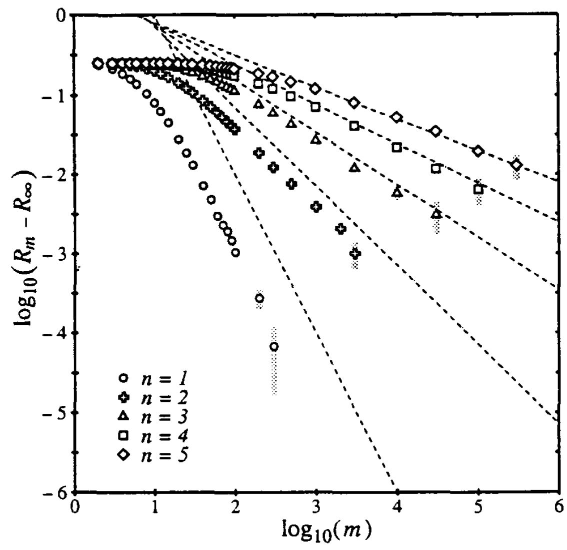

Fig. 3. Numerical evidence supporting the nearest neighbor scaling hypothesis for two trigonometric distributions. Here the dashed lines describe the leading asymptotic behavior predicted by the asymptotic expansion in Theorem 1. The convergence thus occurs at rates of order $m^{-2/n}$.

This data is compared to the asymptotic expansion derived in Theorem 1 which we truncate to second order,

$$R_m \approx R_\infty + c_2 m^{-2/n}. \tag{31}$$

Using the explicit expressions for the coefficients obtained under Hypotheses H1 through H4, a broken curve is plotted for each dimensionality. The close agreement suggests that in this case the boundary contribution is very small, if not exponentially subdominant. ■

Hypotheses H1–H3 would appear to incorporate needed uniformity and smoothness constraints. The following examples illustrate, however, that it may be possible to weaken the constraints of these hypotheses, or alternatively, trade one for the other. The next example demonstrates that Hypothesis H2 may not be necessary to obtain the asymptotic scaling law of Theorem 1 (though with somewhat different expressions for the coefficients of the expansion).

Example: (*Normal distributions*). Consider the classification problem described in $\mathbb{R}^n$ by the two normal class-

[4] The expression given here corrects minor errors in Cover and Hart's early treatment of the problem [2].

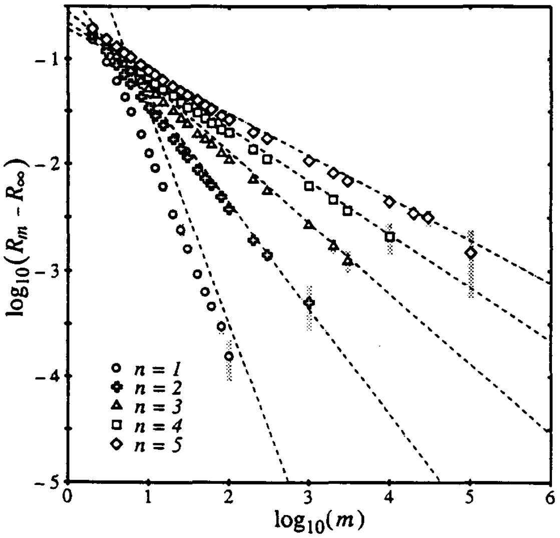

Fig. 4. Numerical evidence supporting the nearest neighbor scaling hypothesis for two normally distributed classes in $\mathbb{R}^n$. The data suggests that the risk converges at rates of order $m^{-2/n}$.

conditional densities,

$$f_1(x) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2}\left((x_1 - \mu)^2 + \sum_{j=2}^{n} x_j^2 \right) \right\},$$

$$f_2(x) = \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2}\left((x_1 + \mu)^2 + \sum_{j=2}^{n} x_j^2 \right) \right\},$$

and prior probabilities, $P_1 = P_2 = 1/2$. Using (11), the risk of the nearest neighbor classifier tends to

$$R_\infty = \frac{1}{\sigma\sqrt{2\pi}} e^{-\mu^2/2\sigma^2} \int_0^\infty e^{-x^2/2\sigma^2} \operatorname{sech}\left(\frac{\mu x}{\sigma^2}\right) dx.$$

For $\mu = \sigma = 1$, a numerical integration yields $R_\infty \approx 0.224\,80$, which is consistent with the Bayes risk, $R_B = (1/2)\operatorname{erfc}(1/\sqrt{2}) \approx 0.158\,65$.

Fig. 4 summarizes the outcomes of numerous simulations of nearest neighbor classifiers operating on data from this family of normal distributions with $n = 1$ through $n = 5$. The broken lines indicate the power law (31), where the coefficient c_2 was chosen to obtain a convincing fit. (For this example, the expression for c_2 in the remarks following Theorem 1 is inapplicable as the hypotheses fail.) Even so, the close agreement again suggests that an asymptotic expression of the form (12) for R_m exists. ■

The last example considers the effect of violating Hypothesis H3.

Example: (*Nonoverlapping uniform distributions*). This case can be illustrated by two nonoverlapping, uniform distributions over the n-dimensional unit cube $[0, 1)^n$. Explicitly, we assume that the *a priori* probabilities of the two classes are equal, $P_1 = P_2 = 1/2$, and that

TABLE I
A TABLE OF THE FIRST FEW FUNCTIONS $\psi_k(x)$ IN THE ASYMPTOTIC EXPANSION (18) FOR $\psi(\rho, x)$ WHEN THE CLASS-CONDITIONAL DENSITIES ARE ASSUMED TO HAVE UNIFORMLY BOUNDED PARTIAL DERIVATIVES OF ORDER $N + 1$

ψ_0	$V_n f$
ψ_1	0
ψ_2	$\frac{V_n}{2(n+2)} \nabla^2 f$
ψ_3	0
ψ_4	$\frac{V_n}{8(n+4)(n+2)} \nabla^4 f$
ψ_5	0
ψ_6	$\frac{V_n}{48(n+6)(n+4)(n+2)} \nabla^6 f$

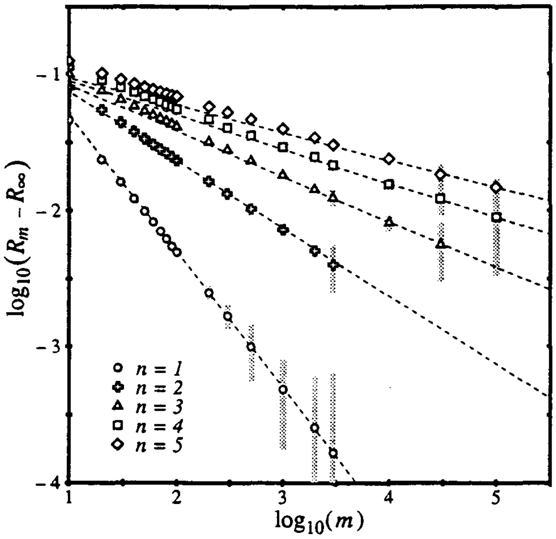

Fig. 5. Numerical evidence supporting the nearest neighbor scaling hypothesis for two, nonoverlapping, uniformly distributed classes in $\mathbb{R}^n$. The data suggests that the risk converges at rates of order $m^{-1/n}$.

$$f_1(x) = \begin{cases} 2, & \text{if } 0 \le x_1 < \frac{1}{2}, \\ 0, & \text{if } \frac{1}{2} \le x_1 < 1, \end{cases}$$

$$f_2(x) = \begin{cases} 0, & \text{if } 0 \le x_1 < \frac{1}{2}, \\ 2, & \text{if } \frac{1}{2} \le x_1 < 1. \end{cases}$$

For $n = 1$, a direct calculation yields

$$R_m = \frac{1}{2(m+1)} + \frac{1}{2^{m+1}},$$

which corresponds to the case $c_1 \neq 0$. Because of analytical complications at the boundary, it is much more difficult to obtain expressions for R_m in higher dimensions.

Fig. 5 indicates the asymptotic trends evidenced by a similar set of numerical experiments for the nearest neighbor classifier with this distribution. For each dimensionality, the discontinuity at $x_1 = 1/2$, rules out the

TABLE II
COEFFICIENTS h_k OF EXPANSION (22) FOR VARIOUS DIMENSIONALITIES. THE FUNCTION ψ_k ARE DEFINED IN (19) AND (20), AND ARE LISTED IN TABLE I UNDER THE STRONGER SMOOTHNESS ASSUMPTION H3′

Term	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n \geq 7$
h_0	ψ_0	ψ_0	ψ_0	ψ_0	ψ_0	ψ_0	ψ_0
h_1	$\frac{1}{2}\psi_0^2$	0	0	0	0	0	0
h_2	$\psi_2 + \frac{1}{3}\psi_0^3$	$\psi_2 + \frac{1}{2}\psi_0^2$	ψ_2	ψ_2	ψ_2	ψ_2	ψ_2
h_3	$\psi_0\psi_2 + \frac{1}{4}\psi_0^4$	0	$\frac{1}{2}\psi_0^2$	0	0	0	0
h_4	$\psi_4 + \psi_0^2\psi_2 +$ $\frac{1}{5}\psi_0^5$	$\psi_4 + \psi_0\psi_2 +$ $\frac{1}{3}\psi_0^3$	ψ_4	$\psi_4 + \frac{1}{2}\psi_0^2$	ψ_4	ψ_4	ψ_4
h_5	$\psi_0\psi_4 + \frac{1}{2}\psi_2^2 +$ $\psi_0^3\psi_2 + \frac{1}{6}\psi_0^6$	0	$\psi_0\psi_2$	0	$\frac{1}{2}\psi_0^2$	0	0
h_6	$\psi_6 + \psi_0^2\psi_4 +$ $\psi_0\psi_2^2 +$ $\psi_0^4\psi_2 + \frac{1}{7}\psi_0^7$	$\psi_6 + \psi_0\psi_4 +$ $\frac{1}{2}\psi_2^2 +$ $\psi_0^2\psi_2 + \frac{1}{4}\psi_0^4$	$\psi_6 + \frac{1}{3}\psi_0^3$	$\psi_6 + \psi_0\psi_2$	ψ_6	$\psi_6 + \frac{1}{2}\psi_0^2$	ψ_6

TABLE III
COEFFICIENTS g_k OF EXPANSION (26) FOR VARIOUS DIMENSIONALITIES.
THE FUNCTIONS α_k ARE DEFINED IN (25)

Term	$n = 1$	$n = 2$	$n = 3$	$n = 4$	$n = 5$	$n = 6$	$n \geq 7$
g_0	α_0	α_0	α_0	α_0	α_0	α_0	α_0
g_1	$\alpha_1 + \alpha_0\psi_0$	α_1	α_1	α_1	α_1	α_1	α_1
g_2	$\alpha_2 + \alpha_1\psi_0 + \alpha_0\psi_0^2$	$\alpha_2 + \alpha_0\psi_0$	α_2	α_2	α_2	α_2	α_2
g_3	$\alpha_3 + \alpha_2\psi_0 + \alpha_1\psi_0^2 +$ $\alpha_0(\psi_0^3 + \psi_2)$	$\alpha_3 + \alpha_1\psi_0$	$\alpha_3 +$ $\alpha_0\psi_0$	α_3	α_3	α_3	α_3
g_4	$\alpha_4 + \alpha_3\psi_0 + \alpha_2\psi_0^2 +$ $\alpha_1(\psi_0^3 + \psi_2) +$ $\alpha_0(\psi_0^4 + 2\psi_0\psi_2)$	$\alpha_4 + \alpha_2\psi_0 +$ $\alpha_0(\psi_0^2 + \psi_2)$	$\alpha_4 +$ $\alpha_1\psi_0$	$\alpha_4 +$ $\alpha_0\psi_0$	α_4	α_4	α_4
g_5	$\alpha_5 + \alpha_4\psi_0 + \alpha_3\psi_0^2 +$ $\alpha_2(\psi_0^3 + \psi_2) +$ $\alpha_1(\psi_0^4 + 2\psi_0\psi_2) +$ $\alpha_0(\psi_0^5 + 3\psi_0^3\psi_2 + \psi_4)$	$\alpha_5 + \alpha_3\psi_0 +$ $\alpha_1(\psi_0^2 + \psi_2)$	$\alpha_5 +$ $\alpha_2\psi_0 +$ $\alpha_0\psi_2$	$\alpha_5 +$ $\alpha_1\psi_0$	$\alpha_5 +$ $\alpha_0\psi_0$	α_5	α_5
g_6	$\alpha_6 + \alpha_5\psi_0 + \alpha_4\psi_0^2 +$ $\alpha_3(\psi_0^3 + \psi_2) +$ $\alpha_2(\psi_0^4 + 2\psi_0\psi_2) +$ $\alpha_1(\psi_0^5 + 3\psi_0^2\psi_2 + \psi_4) +$ $\alpha_0(\psi_0^6 + 4\psi_0^3\psi_2 + \psi_2^2 + 2\psi_0\psi_4)$	$\alpha_6 + \alpha_4\psi_0 +$ $\alpha_2(\psi_0^2 + \psi_2) +$ $\alpha_0(\psi_0^3 +$ $2\psi_0\psi_2 + \psi_4)$	$\alpha_6 +$ $\alpha_3\psi_0 +$ $\alpha_0\psi_0^2 +$ $\alpha_1\psi_2$	$\alpha_6 +$ $\alpha_2\psi_0 +$ $\alpha_0\psi_2$	$\alpha_6 +$ $\alpha_1\psi_0$	$\alpha_6 +$ $\alpha_0\psi_0$	α_6

TABLE IV
A TABLE OF THE FUNCTIONS $\Upsilon_k(x)$ FOR $k \leq 6$

Υ_0	$1/h_0^{\frac{1}{n}}$
Υ_1	$-h_1/nh_0^{1+\frac{2}{n}}$
Υ_2	$\left((n+3){h_1}^2 - 2nh_0h_2\right)/\,2n^2h_0^{2+\frac{3}{n}}$
Υ_3	$-\ \left((2n+4)(n+4){h_1}^3 - 6n(n+4)h_0h_1h_2 + 6n^2{h_0}^2h_3\right)/\,6n^3h_0^{3+\frac{4}{n}}$
Υ_4	$\left((3n+5)(2n+5)(n+5){h_1}^4 - 12n(2n+5)(n+5)h_0{h_1}^2h_2\right.$ $\left.+\ 12n^2(n+5){h_0}^2(2h_1h_3+{h_2}^2) - 24n^3{h_0}^3h_4\right)/\,24n^4h_0^{4+\frac{5}{n}}$
Υ_5	$-\left((4n+6)(3n+6)(2n+6)(n+6){h_1}^5\right.$ $-\ 20n(3n+6)(2n+6)(n+6)h_0{h_1}^3h_2 + 60n^2(2n+6)(n+6)h_0^2h_1(h_1h_3+h_2^2)$ $\left.-\ 120n^3(n+6)h_0^3(h_2h_3+h_1h_4) + 120n^4{h_0}^4h_5\right)/\,120n^5h_0^{5+\frac{6}{n}}$
Υ_6	$\left(-720n^5{h_0}^5h_6 + 360n^4(n+7)h_0^4(h_3^2+2h_1h_5+2h_2h_4)\right.$ $-\ 120n^3(2n+7)(n+7)h_0^3(h_2^3+3h_1^2h_4+6h_1h_2h_3)$ $+\ 60n^2(3n+7)(2n+7)(n+7)h_0^2h_1^2(2h_1h_3+3h_2^2)$ $-\ 30n(4n+7)(3n+7)(2n+7)(n+7)h_0{h_1}^4h_2$ $\left.+\ (5n+7)(4n+7)(3n+7)(2n+7)(n+7){h_1}^6\right)/\,720n^6h_0^{6+\frac{7}{n}}$

existence of the uniform asymptotic expansions assumed by Hypothesis H3 for each class-conditional density. In this case, the rates of convergence tend to be governed by a leading term proportional to $m^{-1/n}$ unlike the preceding examples where the convergence was as rapid as $O(m^{-2/n})$. This example, along with Cover's analysis [3] for the one-dimensional case, suggests that some uniform smoothness hypothesis is necessary to achieve a rapid rate of convergence. ■

VI. CONCLUSION

The examples developed in the previous section (see also the simulations in Fukunaga [13]) indicate that the conditions under which Theorem 1 was proved are not strictly essential. While some form of smoothness and uniformity requirements (such as Hypotheses H1–H3) would appear to be mandated if reasonable performance is to be attained, it may be possible to significantly weaken the constraints imposed by Hypothesis H4. The difficulties here appear to be purely technical involving boundary effects; in particular, even for smooth boundaries, it is difficult to determine even roughly the contribution to the rate of convergence from integral contributions at the boundary where the domain of integration consists of the intersection of a ball with the boundary. It is possible, however, that this is just a constraint imposed by the technique used in the paper.

In summary, the complete asymptotic series expansions for the finite-sample risk developed in the theorem and the corollaries allow us to not just investigate the large sample behavior of the classifier, but the small sample behavior as well. The main results proved in this paper also vividly illustrate Bellman's curse of dimensionality: the finite-sample nearest neighbor risk R_m approaches its infinite-sample limit R_∞ only as slowly as the order of $m^{-2/n}$. Conversely, this indicates that the sample complexity demanded by the nearest neighbor algorithm to achieve acceptable levels of performance grows exponentially with the dimension n for a typical classification problem. In particular, the sample complexity m needed to achieve a finite sample risk which is ϵ close to the infinite sample risk is asymptotically $m \sim (c_2/\epsilon)^{n/2}$.

ACKNOWLEDGMENTS

We would like to thank Leonid Gurvits for suggesting Assertion 2, and an anonymous referee for providing the reference to the work of Fukunaga and Hummels [7].

APPENDIX:

TABULATION OF THE COEFFICIENTS

General expressions for all the coefficients can be obtained as indicated in the proof of the theorem. The expressions for the coefficients can be substantially simplified, however, when Hypothesis H3 is replaced by the following, slightly stronger, smoothness assumption.

Hypothesis H3′: The class-conditional densities f_j possess uniformly bounded partial derivatives up to order $N + 1$ on their probability one support.

Assertion 1 shows that the above smoothness condition implies that Hypothesis H3 holds as well.

We will evaluate relevant coefficients in the proof of the theorem under the stronger Hypothesis H3′ to keep the expressions from growing too unwieldy. As a practical matter, the stronger assumption may be of most use as well.

Coefficients ψ_k: The functions ψ_k satisfy

$$\psi_0(x) = V_n f(x),$$

TABLE V
A TABLE OF THE FUNCTIONS $\lambda_k(\Omega, x)$ FOR $k \leq 6$

λ_0	g_0/nh_0
λ_1	$\left(nh_0g_1 - (n+1)g_0h_1\right) / n\,(n+1)h_0{}^{2+\frac{1}{n}}$
λ_2	$\left(n^2h_0^2g_2 - n(n+2)h_0(g_1h_1 + g_0h_2) + (n+2)(n+1)g_0h_1^2\right) / n^2\,(n+2)h_0{}^{3+\frac{2}{n}}$
λ_3	$\left(2n^3h_0^3g_3 - 2n^2(n+3)h_0^2(g_2h_1 + g_1h_2 + g_0h_3)\right.$ $\left. + n(2n+3)(n+3)h_0h_1(g_1h_1 + 2g_0h_2) - (2n+3)(n+3)(n+1)g_0h_1^3\right) / 2n^3\,(n+3)h_0{}^{4+\frac{3}{n}}$
λ_4	$\left(3n^4h_0^4g_4 - 3n^3(n+4)h_0^3(g_3h_1 + g_2h_2 + g_1h_3 + g_0h_4)\right.$ $+ 3n^2(n+4)(n+2)h_0^2(g_2h_1^2 + 2g_1h_1h_2 + g_0h_2^2 + 2g_0h_1h_3)$ $- n(3n+4)(n+4)(n+2)h_0(g_1h_1^3 + 3g_0h_1^2h_2)$ $\left. + (3n+4)(n+4)(n+2)(n+1)g_0h_1{}^4\right) / 3n^4\,(n+4)h_0{}^{5+\frac{4}{n}}$
λ_5	$\left(24n^5h_0^5g_5 - 24n^4(n+5)h_0^4(g_4h_1 + g_3h_2 + g_2h_3 + g_1h_4 + g_0h_5)\right.$ $+ 12n^3(2n+5)(n+5)h_0^3(g_3h_1^2 + 2g_2h_1h_2 + g_1h_2^2 + 2g_1h_1h_3 + 2g_0h_2h_3 + 2g_0h_1h_4)$ $- 4n^2(3n+5)(2n+5)(n+5)h_0^2(g_2h_1^3 + 3g_1h_1^2h_2 + 3g_0h_1h_2^2 + 3g_0h_1^2h_3)$ $+ n(4n+5)(3n+5)(2n+5)(n+5)h_0(g_1h_1^4 + 4g_0h_1^3h_2)$ $\left. - (4n+5)(3n+5)(2n+5)(n+5)(n+1)g_0h_1{}^5\right) / 24n^5\,(n+5)h_0{}^{6+\frac{5}{n}}$
λ_6	$\left(10n^6h_0^6g_6 - 10n^5(n+6)h_0^5(g_5h_1 + g_4h_2 + g_3h_3 + g_2h_4 + g_1h_5 + g_0h_6)\right.$ $+ 10n^4(n+6)(n+3)h_0^4(g_4h_1^2 + 2g_3h_1h_2 + g_2h_2^2$ $\quad + 2g_2h_1h_3 + 2g_1h_2h_3 + 2g_1h_1h_4 + g_0h_3^2 + 2g_0h_2h_4 + 2g_0h_1h_5)$ $- 10n^3(n+6)(n+3)(n+2)h_0^3(g_3h_1^3 + 3g_2h_1^2h_2$ $\quad + 3g_1h_1h_2^2 + 3g_1h_1^2h_3 + g_0h_2^3 + 6g_0h_1h_2h_3 + 3g_0h_1^2h_4)$ $+ 5n^2(2n+3)(n+6)(n+3)(n+2)h_0^2(g_2h_1^4 + 4g_1h_1^3h_2 + 6g_0h_1^2h_2^2 + 4g_0h_1^3h_3)$ $- n(5n+6)(2n+3)(n+6)(n+3)(n+2)h_0(g_1h_1^5 + 5g_0h_1^4h_2)$ $\left. + (5n+6)(2n+3)(n+6)(n+3)(n+2)(n+1)g_0h_1{}^6\right) / 10n^6\,(n+6)h_0{}^{7+\frac{6}{n}}$

and

$$\psi_k(x) = \frac{1}{k+n}\int_{S_{n-1}} f_k(\Omega, x)\, d\Omega, \qquad k = 1,\cdots, N,$$

where V_n is the volume of the unit ball in n-dimensions. For $k = 1,\cdots, N-1$, each $f_k(\Omega, x)$, can be written as a sum of partial derivatives of the mixture density as in (7) and (8). The element of measure of S_{n-1} is represented by

$$d\Omega = \cos^{n-2}\phi_1 \cos^{n-3}\phi_2 \cdots$$
$$\cos\phi_{n-2}\, d\phi_1\, d\phi_2 \cdots d\phi_{n-1},$$

where, $-\pi/2 \leq \phi_i \leq \pi/2$ $(i = 1,\cdots, n-2)$, and $0 \leq \phi_{n-1} \leq 2\pi$. It is readily verified that every integrand with odd parity, i.e., those having an odd power of $\sin\phi_j$ for at least one $j \in \{1,\cdots, n-1\}$, evaluates to zero. Consequently, $\psi_k(x) = 0$ for odd k, while for even k the values of ψ_k can be determined through direct integration. The results are tabulated in Table I.

Coefficients h_k: The coefficients h_k in (22) are obtained by substituting from (18) and collecting like powers of ρ. Table II lists coefficients h_k for $0 \leq k \leq 6$.

Coefficients g_k: The coefficients g_k in (26) are obtained by substituting from (18) and collecting like powers of ρ. Table III lists coefficients g_k for $0 \leq k \leq 6$.

Coefficients Υ_k: The coefficients $\Upsilon_k(x)$, $0 \leq k \leq 6$, are listed in Table IV. The higher-ordered coefficients, as well as the more arduous expressions that follow, were derived with the aid of *Mathematica* [14].

Coefficients λ_k: The coefficients $\lambda_k(\Omega, x)$, $0 \leq k \leq 6$, are listed in Table V. Again, we have resorted to *Mathematica* to obtain the higher-order terms.

TABLE VI
THE INFINITE-SAMPLE RISK OF THE NEAREST NEIGHBOR CLASSIFIER

Range	R_∞
$n \geq 1$	$2P_1P_2\int_S \frac{f_1f_2}{f}$

Coefficients c_k: Recall that under Hypotheses H1–H4, Theorem 1 gives the asymptotic expansion

$$R_m = R_\infty + \sum_{k=2}^{N} c_k m^{-k/n} + o(m^{-N/n}) \qquad (m \to \infty).$$

The general form of the coefficients c_k involves the asymptotic expansion coefficients $f_{j,k}(\Omega, x)$. The complete representation for c_2, for instance, is given in (12). Substantial simplifications in the expressions result when Hypothesis H3 is replaced by the stronger Hypothesis H3′, and Tables VI–XI list expressions for the coefficients c_k, $0 \leq k \leq 6$, under the stronger hypothesis. In particular, note that explicit expressions can now be written for the c_k's involving only partial derivatives of the class-conditional densities. The general forms for these coefficients can be readily derived in a similar recursive fashion (albeit with somewhat more algebraic detail) as indicated in the proof of the theorem.

The coefficient $c_0 \equiv R_\infty$ is given in Table VI. This general form is valid under Hypothesis H3 as well, and indeed, under more general conditions than those imposed in this paper (see Cover and Hart [2]).

The simplified expression for c_2 under Hypothesis H3′ is shown in Table VII and is equivalent to the second-order

TABLE VII
THE COEFFICIENT c_2 UNDER HYPOTHESIS H3′. THE GENERAL EXPRESSION FOR c_2 UNDER HYPOTHESIS H3 IS GIVEN IN (12)

Range	c_2
$n \geq 1$	$\frac{\Gamma(1+\frac{2}{n})\Gamma(1+\frac{n}{2})^{2/n}}{2n\pi} \int_S \frac{P_1P_2f_1f_2}{f^{1+2/n}} \left(\frac{1}{f_1}\nabla^2 f_1 + \frac{1}{f_2}\nabla^2 f_2 - \frac{2}{f}\nabla^2 f\right)$

TABLE VIII
THE COEFFICIENT c_3 UNDER HYPOTHESIS H3′

Range	c_3
$n = 1$	$-3c_2$
$n \geq 2$	0

TABLE IX
THE COEFFICIENT c_4 UNDER HYPOTHESIS H3′

Range	c_4
$n = 1$	$\frac{P_1P_2}{16} \int_S \frac{f_1f_2}{f^5} \left(\frac{1}{f_1}\frac{d^4f_1}{dx^4} + \frac{1}{f_2}\frac{d^4f_2}{dx^4} - \frac{2}{f}\frac{d^4f}{dx^4} - \frac{2}{f}\frac{d^4f}{dx^4}\right.$ $\left. + \left(28f^2 - \frac{10}{f}\frac{d^2f}{dx^2}\right)\left(\frac{1}{f_1}\frac{d^2f_1}{dx^2} + \frac{1}{f_2}\frac{d^2f_2}{dx^2} - \frac{2}{f}\frac{d^2f}{dx^2}\right)\right)$
$n = 2$	$\frac{P_1P_2}{32\pi^2} \int_S \frac{f_1f_2}{f^3} \left(\frac{1}{f_1}\nabla^4 f_1 + \frac{1}{f_2}\nabla^4 f_2 - \frac{2}{f}\nabla^4 f\right.$ $\left. - \left(8\pi f + \frac{6}{f}\nabla^2 f\right)\left(\frac{1}{f_1}\nabla^2 f_1 + \frac{1}{f_2}\nabla^2 f_2 - \frac{2}{f}\nabla^2 f\right)\right)$
$n \geq 3$	$\frac{P_1P_2\,\Gamma(1+\frac{4}{n})\,\Gamma(1+\frac{n}{2})^{4/n}}{8\pi^2 n(n+2)} \int_S \frac{f_1f_2}{f^{1+4/n}} \left(\frac{1}{f_1}\nabla^4 f_1 + \frac{1}{f_2}\nabla^4 f_2 - \frac{2}{f}\nabla^4 f\right.$ $\left. - \left(1+\frac{4}{n}\right)\frac{2}{f}\nabla^2 f\left(\frac{1}{f_1}\nabla^2 f_1 + \frac{1}{f_2}\nabla^2 f_2 - \frac{2}{f}\nabla^2 f\right)\right)$

TABLE X
THE COEFFICIENT c_5 UNDER HYPOTHESIS H3′

Range	c_5
$n = 1$	$-\frac{5P_1P_2}{8} \int_S dx \frac{f_1f_2}{f^5} \left(\frac{1}{f_1}\frac{d^4f_1}{dx^4} + \frac{1}{f_2}\frac{d^4f_2}{dx^4} - \frac{2}{f}\frac{d^4f}{dx^4}\right.$ $\left. + \left(6\,f^2 - \frac{10}{f}\frac{d^2f}{dx^2}\right)\left(\frac{1}{f_1}\frac{d^2f_1}{dx^2} + \frac{1}{f_2}\frac{d^2f_2}{dx^2} - \frac{2}{f}\frac{d^2f}{dx^2}\right)\right)$
$n = 2$	0
$n = 3$	$-\frac{5\,\Gamma(2/3)P_1P_2}{54\,(6\pi^2)^{1/3}} \int_S \frac{f_1f_2}{f^{8/3}} \left(\frac{1}{f_1}\nabla^2 f_1 + \frac{1}{f_2}\nabla^2 f_2 - \frac{2}{f}\nabla^2 f\right)$
$n \geq 4$	0

TABLE XI
THE COEFFICIENT c_6 UNDER HYPOTHESIS H3′

Range	c_6
$n = 1$	$\frac{P_1P_2}{64}\int_S \frac{f_1f_2}{f^7}\Big(\frac{1}{f_1}\frac{d^6f_1}{dx^6}+\frac{1}{f_2}\frac{d^6f_2}{dx^6}-\frac{2}{f}\frac{d^6f}{dx^6}$ $+\left(260\,f^2-35\,\frac{1}{f}\,\frac{d^2f}{dx^2}\right)\left(\frac{1}{f_1}\,\frac{d^4f_1}{dx^4}+\frac{1}{f_2}\,\frac{d^4f_2}{dx^4}-\frac{2}{f}\,\frac{d^4f}{dx^4}\right)$ $+\left(\frac{280}{f^2}\left(\frac{d^2f}{dx^2}\right)^2-\frac{21}{f}\,\frac{d^4f}{dx^4}-2600\,f\,\frac{d^2f}{dx^2}+496\,f^4\right)\left(\frac{1}{f_1}\,\frac{d^2f_1}{dx^2}+\frac{1}{f_2}\,\frac{d^2f_2}{dx^2}-\frac{2}{f}\,\frac{d^2f}{dx^2}\right)\Big)$
$n = 2$	$\frac{P_1P_2}{384\pi^3}\int_S \frac{f_1f_2}{f^4}\Big(\frac{1}{f_1}\nabla^6 f_1+\frac{1}{f_2}\nabla^6 f_2-\frac{2}{f}\nabla^6 f$ $-\left(36\pi f+\frac{18}{f}\nabla^2 f\right)\left(\frac{1}{f_1}\nabla^4 f_1+\frac{1}{f_2}\nabla^4 f_2-\frac{2}{f}\nabla^4 f\right)$ $+\left(90\left(\frac{1}{f}\nabla^2 f\right)^2-\frac{12}{f}\nabla^4 f+216\,\pi\,\nabla^2 f+96\,\pi^2\,f^2\right)\left(\frac{1}{f_1}\nabla^2 f_1+\frac{1}{f_2}\nabla^2 f_2-\frac{2}{f}\nabla^2 f\right)\Big)$
$n = 3$	$\frac{P_1P_2}{4480\pi^2}\int_S \frac{f_1f_2}{f^3}\Big(\frac{1}{f_1}\nabla^6 f_1+\frac{1}{f_2}\nabla^6 f_2-\frac{2}{f}\nabla^6 f-\frac{63}{5}\,\frac{1}{f}\nabla^2 f\left(\frac{1}{f_1}\nabla^4 f_1+\frac{1}{f_2}\nabla^4 f_2-\frac{2}{f}\nabla^4 f\right)$ $+\left(\frac{63}{5}\left(\frac{2}{f}\nabla^2 f\right)^2-\frac{9}{f}\nabla^4 f\right)\left(\frac{1}{f_1}\nabla^2 f_1+\frac{1}{f_2}\nabla^2 f_2-\frac{2}{f}\nabla^2 f\right)\Big)$
$n = 4$	$\frac{P_1P_2}{768\,(2\pi)^{5/2}}\int_S \frac{f_1f_2}{f^{5/2}}\Big(\frac{1}{f_1}\nabla^6 f_1+\frac{1}{f_2}\nabla^6 f_2-\frac{2}{f}\nabla^6 f-\frac{10}{f}\nabla^2 f\left(\frac{1}{f_1}\nabla^4 f_1+\frac{1}{f_2}\nabla^4 f_2-\frac{2}{f}\nabla^4 f\right)$ $+\left(35\left(\frac{1}{f}\nabla^2 f\right)^2-\frac{15}{2}\,\frac{1}{f}\nabla^4 f-144\,\pi^2\,f\right)\left(\frac{1}{f_1}\nabla^2 f_1+\frac{1}{f_2}\nabla^2 f_2-\frac{2}{f}\nabla^2 f\right)\Big)$
$n \geq 5$	$\frac{P_1P_2\,\Gamma\left(1+\frac{6}{n}\right)\Gamma\left(1+\frac{n}{2}\right)^{6/n}}{48\pi^3\,n(n+2)(n+4)}\int_S \frac{f_1f_2}{f^{1+6/n}}\Big(\frac{1}{f_1}\nabla^6 f_1+\frac{1}{f_2}\nabla^6 f_2-\frac{2}{f}\nabla^6 f$ $-3\left(1+\frac{6}{n}\right)\Big(\frac{n+4}{n+2}\,\frac{1}{f}\nabla^2 f\left(\frac{1}{f_1}\nabla^4 f_1+\frac{1}{f_2}\nabla^4 f_2-\frac{2}{f}\nabla^4 f\right)$ $+\left(\frac{1}{f}\nabla^4 f-2\,\frac{(n+4)(n+3)}{n+2}\left(\frac{1}{f}\nabla^2 f\right)^2\right)\left(\frac{1}{f_1}\nabla^2 f_1+\frac{1}{f_2}\nabla^2 f_2-\frac{2}{f}\nabla^2 f\right)\Big)\Big)$

expansion coefficient derived by Fukanaga and Hummels [7] [B in Formula (1)]. For the general form of the coefficient see (12).

Assertion 2 ensures that c_3 vanishes for all but one-dimensional feature spaces, in which case $c_3 = -3c_2$, as in Table VIII. In general, the odd numbered coefficients vanish when n becomes suitably large. Expressions for c_4, c_5, and c_6 are presented in Tables IX–XI. All the expressions are derived with Hypothesis H3′ replacing Hypothesis H3.

REFERENCES

[1] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*. New York: Wiley & Sons, 1973.

[2] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 21–27, 1967.

[3] T. M. Cover, "Rates of convergence of nearest neighbor decision procedures," *Proc. First Annual Hawaii Conf. on Systems Theory*, pp. 413–415, 1968.

[4] T. J. Wagner, "Convergence of the nearest neighbor rule," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 566–571, 1971.

[5] J. Fritz, "Distribution-free exponential error bound for nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 552–557, 1975.

[6] L. Györfi, "On the rate of convergence of nearest neighbor rules," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 509–512, 1978.

[7] K. Fukunaga and D. M. Hummels, "Bias of nearest neighbor estimates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. PAMI-9, pp. 103–112, 1987.

[8] A. Erdélyi, *Asymptotic Expansions*. New York: Dover, 1956.

[9] P. Billingsley, *Probability of Measure*. New York: Wiley & Sons, 1986.

[10] R. R. Snapp, D. Psaltis, and S. S. Venkatesh, "Asymptotic slowing down of the nearest-neighbor classifier," in *Advances in Neural Information Processing Systems, 3* (ed. R. Lippmann, J. Moody, and D. Touretzky). San Mateo, California: Morgan Kaufmann, 1991.

[11] W. Fulks and J. O. Sather, "Asymptotics II: Laplace's method for multiple integrals," *Pacific J. Math.*, vol. 11, pp. 185–192, 1961.

[12] G. N. Watson, "The harmonic functions associated with the parabolic cylinder," *Proc. London Math. Soc.*, vol. 17, pp. 116–148, 1918.

[13] K. Fukunaga, Introduction to Statistical Pattern Recognition, Second Edition, New York: Academic Press, 1990.

[14] S. Wolfram, *Mathematica: a system for doing mathematics by computer*, 2nd ed. Redwood City, CA: Addison-Wesley, 1991.