The Annals of Statistics
1998, Vol. 26, No. 3, 850–878

ASYMPTOTIC EXPANSIONS OF THE *k* NEAREST NEIGHBOR RISK[1]

BY ROBERT R. SNAPP[2] AND SANTOSH S. VENKATESH[3]

University of Vermont and University of Pennsylvania

The finite-sample risk of the k nearest neighbor classifier that uses a weighted L^p-metric as a measure of class similarity is examined. For a family of classification problems with smooth distributions in $\mathbb{R}^n$, an asymptotic expansion for the risk is obtained in decreasing fractional powers of the reference sample size. An analysis of the leading expansion coefficients reveals that the optimal weighted L^p-metric, that is, the metric that minimizes the finite-sample risk, tends to a weighted Euclidean (i.e., L^2) metric as the sample size is increased. Numerical simulations corroborate this finding for a pattern recognition problem with normal class-conditional densities.

1. Introduction. In its original form [6], the k nearest neighbor classifier is perhaps the simplest pattern classification algorithm yet devised. Given a classification problem, in which a pattern, represented as an n-dimensional feature vector, is to be assigned to one of several pattern classes (e.g., states of nature), and any positive integer k, the k nearest neighbor algorithm requires two ingredients: (i) a finite reference sample of m (with $m \geq k$) feature vectors labeled (or classified) according to the pattern class of origin, and (ii) a metric, or pattern similarity function, in the space of the feature vectors. Given an input feature vector, the algorithm uses the given metric to first identify the k feature vectors from the reference sample that are closest to the input vector and then assigns the input feature vector to the pattern class that appears most frequently amongst the k nearest neighbors. (If no single class appears with greatest frequency, then an auxiliary procedure can be invoked to handle ties.) Despite its simplicity, this nonparametric algorithm is asymptotically consistent with a Bayes classifier [18], and is competitive with other popular classifiers in practical settings [1].

In the following, we consider the k nearest neighbor classifier in the context of weighted L^p-metrics of the form $d(\mathbf{x}, \mathbf{y}) = \|A(\mathbf{x} - \mathbf{y})\|_p$, where $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, A is a constant, nonsingular, linear transformation, and

Received June 1996; revised January 1998.

[1] Presented in part at the Twelfth International Conference on Pattern Recognition (1994) and at the IEEE International Symposium on Information Theory (1995).

[2] Supported in part by NSF Grant OSR-93-50540 and by Rome Laboratory, Air Force Material Command, USAF Grant F30602-94-1-0010.

[3] Supported in part by the Air Force Office of Scientific Research Grant F49620-93-1-0120.

AMS 1991 *subject classifications*. Primary 62G20, 62H30, 41A60.

Key words and phrases. k nearest neighbor classifier, finite-sample risk, asymptotic expansions, Laplace's method.

$\|\mathbf{x}\|_p = \sqrt[p]{|x_1|^p + \cdots + |x_n|^p}$ is the usual L^p-norm of $\mathbf{x}$. For a family of smooth classification problems, we seek a precise characterization of the finite-sample performance of the classifier utilizing such weighted L^p-metrics; in particular, we seek to explicitly detail the effects of the finite-sample size m, the dimensionality of the feature space n, and the choice of metric d on the performance of the classifier.

Our main results may be summarized as follows: for classification problems endowed with smooth distributions (the precise conditions are spelled out in Section 4), the finite-sample risk of the classifier has an asymptotic series expansion of the form

$$(1) \qquad R_m \sim R_\infty + \sum_{j=2}^{\infty} c_j m^{-j/n} \qquad (m \to \infty).$$

[The expansion holds in the sense of Poincaré: truncating the series after N terms leads to a residual error of order $\mathscr{O}(m^{-(N+1)/n})$ in the estimate for R_m.] The expansion coefficients c_j depend in general upon k, the metric parameters p and A and the chosen procedure for handling ties, in addition to the probability distributions that describe the classification problem under consideration, but are independent of the sample size m. The leading coefficient R_∞ is just the infinite-sample risk derived by Cover and Hart [3]; while R_∞ depends on k and the underlying distributions, it is independent of the choice of metric.

A further analysis of the leading coefficients in the asymptotic expansion for the finite-sample risk allows us to adduce the effects of the choice of metric on the finite-sample performance. Indeed, we show that for a large enough sample size m, the finite-sample risk R_m is minimized for a choice of weighted L^p-metric for specific values of the metric parameters $A = A_{\text{opt}}$ and $p = p_{\text{opt}}$. The optimal value p_{opt} depends on the sample size and lies in a neighborhood of 2; as m increases, the diameter of this neighborhood decreases, and $p_{\text{opt}} \to 2$ as $m \to \infty$. Numerical simulations confirm that the L^2-metric is superior to the L^1- and L^∞-metrics for normally distributed two-class problems in 12- and 25-dimensional feature spaces for practically attainable sample sizes. The asymptotically optimal linear transformation A_{opt} within the metric depends upon the underlying distributions that describe the given pattern recognition problem. Given the analytic forms of these distribution, this transformation can be obtained, in principle, as the solution of a constrained optimization problem using the Euler–Lagrange multiplier theorem.

To our knowledge, this is the first study to rigorously evaluate the finite-sample risk of the k nearest neighbor classifier under different weighted L^p-metrics. In previous work, Fukunaga and Flick [9] obtained a heuristic expression for the optimal quadratic weight matrix which differs somewhat, however, from our rigorous estimates. Earlier studies of the finite-sample risk of a nearest neighbor classifier include a rate of convergence result for a one-dimensional, two-class problem [2], a heuristic estimate of the bias of the classifier [8] and a previous study by the present authors of the risk of a nearest neighbor classifier ($k = 1$) under the usual Euclidean (L^2) metric for a two-class problem [13].

The k nearest neighbor algorithm is formally described in Section 2, along with different deterministic tie-breaking algorithms. Section 3 contains a derivation of an exact integral representation of the finite-sample risk for the k nearest neighbor classifier. In Section 4 we identify a family of smooth classification problems for which the integral representation for the finite-sample risk can be evaluated asymptotically using a generalization of Laplace's method and present the main results, including the existence of the asymptotic expansion (1) (and, more generally, truncated versions under weaker smoothness conditions), as well as the asymptotic optimality of the weighted L^2-metric. Section 5 describes the numerical simulations that validate the asymptotic optimality of the L^2-metric for a pattern recognition problem with two normally distributed class-conditional densities. Section 6 concludes the main body of the paper with some remarks on the significance of the Main Theorem. Technical lemmas are collected in Appendix A, and a constructive proof of the Main Theorem is sketched in Appendix B. A more detailed treatment appears in a technical report [16].

2. The *k* nearest neighbor classifier.

Let the elements of $\mathbb{L} = \{1, 2, \dots, C\}$ denote labels for C states of nature, or pattern classes. Patterns are represented by feature vectors $\mathbf{X} \in \mathbb{R}^n$. It is assumed that data is available in the form of a finite labeled reference sample $\mathscr{X}_m \triangleq \{(\mathbf{X}^1, L^1), \dots, (\mathbf{X}^m, L^m)\}$, where, for each i, L^i is a member of $\mathbb{L}$ and $\mathbf{X}^i$ is a feature vector representing class L^i. Given a metric $d(\mathbf{X}, \mathbf{X}')$ on $\mathbb{R}^n$ and a positive integer k, the k nearest neighbor classifier generates a map from $\mathbb{R}^n$ into $\mathbb{L}$ as a function of the reference sample $\mathscr{X}_m$ wherein each point $\mathbf{X} \in \mathbb{R}^n$ is mapped into one of the C classes according to the majority of the labels of its k nearest neighbors in the reference sample. More particularly, fix any $\mathbf{X} \in \mathbb{R}^n$. We may suppose, without loss of generality, that the indices of the labeled feature vectors in $\mathscr{X}_m$ are permuted to satisfy

$$\begin{aligned} &d(\mathbf{X}, \mathbf{X}^1) \le d(\mathbf{X}, \mathbf{X}^2) \le \cdots \le d(\mathbf{X}, \mathbf{X}^k), \\ &d(\mathbf{X}, \mathbf{X}^j) \ge d(\mathbf{X}, \mathbf{X}^k) \quad \text{for } j = k+1, \dots, m. \end{aligned} \tag{2}$$

The k nearest neighbors of $\mathbf{X}$ are identified as the labeled subset of feature vectors $\{(\mathbf{X}^1, L^1), \dots, (\mathbf{X}^k, L^k)\}$ of the reference sample. The k nearest neighbor classifier then assigns $\mathbf{X}$ to the class

$$L' = \operatorname{maj}(L^1, \dots, L^k), \tag{3}$$

namely, the most frequent class label exhibited by the k nearest neighbors of $\mathbf{X}$. Ties and equalities in (2) can be resolved by an arbitrary procedure.

We are interested in quantifying how the expected accuracy of the algorithm is influenced by the chosen metric. In particular, we consider the family of weighted L^p-metrics

$$d(\mathbf{x},\mathbf{x}') \triangleq \|A(\mathbf{x}-\mathbf{x}')\|_p, \qquad \mathbf{x},\mathbf{x}' \in \mathbb{R}^n, \tag{4}$$

where

$$\|\mathbf{x}\|_p \triangleq \begin{cases} \sqrt[p]{|x_1|^p + \cdots + |x_n|^p}, & \text{if } 1 \le p < \infty, \\ \max\limits_{1 \le i \le n} |x_i|, & \text{if } p = \infty, \end{cases}$$

is the usual L^p-norm, and A is a nonsingular, linear transformation. Thus, the weighted L^p-metric is determined by specifying a nonsingular n-by-n matrix, and a number $p \in [1,\infty]$. Without loss of generality, we assume that A has been normalized so that it is measure preserving, that is, $\det A = 1$. (Scaling the metric d by $|\det A|^{-1}$—and, in general, by any positive quantity —does not change the partition of $\mathbb{R}^n$ engendered by the k nearest neighbor classifier, hence does not affect the risk.)

2.1. *Deterministic tie-breaking algorithms.* If k is even, or if $C \ge 3$, it is necessary to define an auxiliary procedure to handle ties in (3). [Although the equalities in (2) can also be problematic, they occur with zero probability if the class-conditional distributions are absolutely continuous, as we assume subsequently.] A simple method is to break the tie according to a deterministic rule; for example, in the event that more than one class occurs with greatest frequency in the subset of k nearest neighbors, then the input pattern could be assigned to the most prolific class in the subset with the smallest class label. A convenient way to describe deterministic tie-breaking rules such as this is to construct an *assignment partition* $\mathscr{L}_1,\dots,\mathscr{L}_C$ of the space $\mathbb{L}^k$ that describes the action of the k nearest neighbor classifier for every possible ordered k-tuple of class labels. Here, $\mathscr{L}_i$ contains every ordered k-tuple, $\mathbf{l} = (l^1,\dots,l^k)$, representing the respective class labels of $\mathbf{X}^1,\dots,\mathbf{X}^k$, for which an assignment to class i occurs. For example, if $C = 2$ and $k = 2$, then (3) with the smallest-class-label tie-breaking algorithm induces the assignment partition, $\mathscr{L}_1 = \{(1,1),(1,2),(2,1)\}$ and $\mathscr{L}_2 = \{(2,2)\}$. By introducing an extra element $\mathscr{L}_0$ into the partition, one can represent a k nearest neighbor classifier that rejects certain input patterns [11]. If the action of a k nearest neighbor classifier can be described by an assignment partition, we shall call it *deterministic*.

Random tie-breaking algorithms can be represented by a stochastic process in which an assignment partition is selected from an ensemble of partitions, according to a discrete probability distribution, prior to each assignment.

3. The finite-sample risk. The classifier's expected accuracy can be quantified in terms of the classification risk, which is determined by the underlying statistical assumptions governing the data. We suppose that $\{P_1,\dots,P_C\}$ is a discrete probability distribution representing the a priori probabilities of the pattern classes $\mathbb{L} = \{1,\dots,C\}$ and that, for each $l \in \mathbb{L}$, feature vectors or patterns originating from class l are generated by the fixed conditional distribution F_l on $\mathbb{R}^n$.

In this statistical framework, we suppose that labeled feature vectors $(\mathbf{X}, L)$ are generated by a two-step process. First, a class $L \in \mathbb{L}$ is chosen at random according to the distribution $\{P_1, \dots, P_C\}$ on pattern classes; then a random feature vector $\mathbf{X}$ is drawn according to F_L from $\mathbb{R}^n$. We assume that the reference sample $\mathscr{X}_m = \{(\mathbf{X}^i, L^i), 1 \le i \le m\}$ is generated by m independent repetitions of this process.

To analyze the expected performance of the algorithm, let $(\mathbf{X}, L)$ represent an independent, labeled feature vector (a random "test point") that is obtained by the same random process used to construct the m reference vectors. The feature vector $\mathbf{X}$ is presented to a deterministic k nearest neighbor classifier, which in turn assigns $\mathbf{X}$ to class $L' = L'(\mathbf{X}, \mathscr{X}_m)$, an estimator of the class with maximum a posteriori probability. We define the expected m-sample risk of the classifier as

$$R_m \triangleq \sum_{i=1}^{C} \sum_{j=1}^{C} \Lambda_{i,j} \mathbb{P}\{L' = i, L = j\},$$

where $\Lambda_{i,j}$ is the cost incurred in assigning a feature vector that originates from class j to class i. Although, in practice, different errors may incur different costs, we shall henceforth assume the *zero–one* cost matrix $\Lambda_{i,j} = 1 - \delta_{i,j}$, where $\delta_{i,j}$ is the Kronecker delta function. (For other cost matrices, it is desirable to modify the original algorithm so that the resulting risk function is minimized, e.g., let $L' = \arg\min_{i \in \mathbb{L}} \sum_j \Lambda_{i,j} \nu_j$, where ν_j denotes the number of times class label j appears in the subsample of k nearest neighbors. For these cases, the ensuing analysis should apply with only minor modifications.) Using the zero–one cost matrix, the risk function reduces to $R_m = \mathbb{P}\{L' \ne L\}$, the probability that the k nearest neighbor algorithm assigns $\mathbf{X}$ to an incorrect class.

The finite-sample risk R_m can be written in integral form by taking the expectation of the probability of the event $L' \ne L$ conditioned on the reference sample and the test feature vector. More specifically, for each $l \in \mathbb{L}$, suppose the class-conditional distributions F_l are absolutely continuous with corresponding densities f_l. Let $f = \sum_{l=1}^{C} P_l f_l$ denote the mixture density, and let $S = \{\mathbf{x} \in \mathbb{R}^n\colon f(\mathbf{x}) > 0\}$ be its probability-one support in $\mathbb{R}^n$. Then

$$R_m = \mathbb{P}\{L' \ne L\} = \int_S \mathbb{P}\{L' \ne L \mid \mathbf{x}\} f(\mathbf{x})\, d\mathbf{x}, \tag{5}$$

where $\mathbb{P}\{L' \ne L \mid \mathbf{x}\}$ denotes the probability of error conditioned on the event $\{\mathbf{X} = \mathbf{x}\}$. With the reference sample permuted as necessary, let $\mathbf{X}^1, \dots, \mathbf{X}^k$ be the k nearest neighbors of $\mathbf{X} = \mathbf{x}$ as per the indexing convention set out in (2). Write $\mathbb{P}\{L' \ne L \mid \mathbf{x}, \mathbf{x}^1, \dots, \mathbf{x}^k\}$ for the probability of error conditioned jointly on the events $\{\mathbf{X} = \mathbf{x}, \mathbf{X}^1 = \mathbf{x}^1, \dots, \mathbf{X}^k = \mathbf{x}^k\}$ and let $f_m(\mathbf{x}^1, \dots, \mathbf{x}^k \mid \mathbf{x})$ denote the conditional probability density associated with the event that the k nearest neighbors of $\mathbf{x}$ in a random m-sample take values $\mathbf{X}^1 = \mathbf{x}^1, \dots,$ $\mathbf{X}^k = \mathbf{x}^k$. Introduce the notation $B(\rho, \mathbf{x}) \triangleq \{\mathbf{x}' \in \mathbb{R}^n\colon d(\mathbf{x}, \mathbf{x}') \le \rho\}$, where $d(\mathbf{x}, \mathbf{x}')$ is the weighted L^p-metric defined in (4), for the closed ball of radius ρ

at $\mathbf{x}$, and write $S^{(j)} \triangleq B(d(\mathbf{x},\mathbf{x}^j),\mathbf{x}) \cap S, (1 \le j \le k)$ for points in the probability-one support of f at distance no more than $d(\mathbf{x},\mathbf{x}^j)$ from $\mathbf{x}$. The inequalities in (2) imply that $\mathbf{x}^j \in S^{(j+1)}$ for $j = 1, 2, \dots, k-1$. See Figure 1. Taking expectation with respect to the values of the k nearest neighbors of $\mathbf{x}$, we hence obtain

$$\mathbb{P}\{L' \neq L \mid \mathbf{x}\} = \int_S \int_{S^{(k)}} \cdots \int_{S^{(2)}} \mathbb{P}\{L' \neq L \mid \mathbf{x}, \mathbf{x}^1, \dots, \mathbf{x}^k\} \\ \times f_m(\mathbf{x}^1, \dots, \mathbf{x}^k \mid \mathbf{x})\, d\mathbf{x}^1 \cdots d\mathbf{x}^{k-1}\, d\mathbf{x}^k.$$

After substituting the above into (5), and assuming statistical independence, the risk can be represented as

$$R_m = (m)_k \int_S \int_S \mathfrak{g}(\mathbf{x},\mathbf{x}^k) \exp\bigl(-m\mathfrak{h}\bigl(d(\mathbf{x},\mathbf{x}^k),\mathbf{x}\bigr)\bigr)\, d\mathbf{x}^k d\mathbf{x}, \tag{6}$$

where

$$(m)_k \triangleq m(m-1)\cdots(m-k+1),$$

$$\mathfrak{g}(\mathbf{x},\mathbf{x}^k) \triangleq \sum_{l=1}^{C} \sum_{\mathbf{l} \notin \mathscr{L}_l} P_l P_{l^1} \cdots P_{l^k} \frac{f_l(\mathbf{x}) f_{l^k}(\mathbf{x}^k)}{\bigl(1-\psi\bigl(d(\mathbf{x},\mathbf{x}^k),\mathbf{x}\bigr)\bigr)^k} \\ \times \int_{S^{(k)}} \cdots \int_{S^{(3)}} \int_{S^{(2)}} f_{l^1}(\mathbf{x}^1) f_{l^2}(\mathbf{x}^2) \\ \cdots f_{l^{k-1}}(\mathbf{x}^{k-1})\, d\mathbf{x}^1\, d\mathbf{x}^2 \cdots d\mathbf{x}^{k-1}, \tag{7}$$

$$\mathfrak{h}(\rho,\mathbf{x}) \triangleq -\ln(1-\psi(\rho,\mathbf{x})), \tag{8}$$

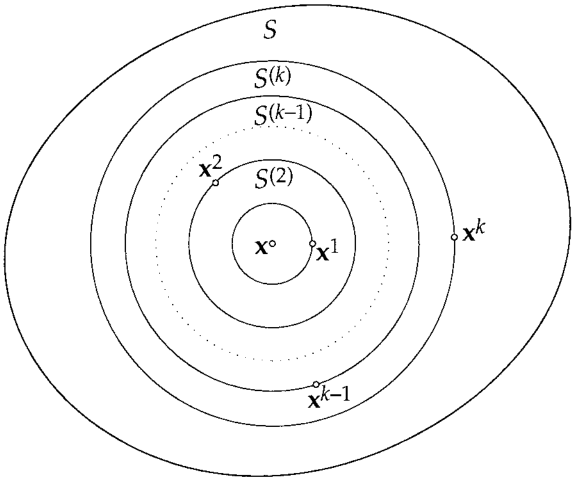

FIG. 1. *A hypothetical input feature vector* $\mathbf{x}$, *and reference feature vectors* $\mathbf{x}^1, \dots, \mathbf{x}^k$ *that represent the subset of k nearest neighbors. Note how these vectors induce a system of concentric balls,* $S^{(2)}, \dots, S^{(k)}$.

and

$$\psi(\rho, \mathbf{x}) \triangleq \int_{B(\rho, \mathbf{x})} f(\mathbf{x}')\, d\mathbf{x}'.$$

The integral in (6) is difficult to evaluate for finite m even if the analytic forms of the class-conditional densities are known. In the infinite-sample limit, $m \to \infty$, Cover and Hart [3] showed that under weak assumptions R_m tends to the well-defined limit

$$(9) \qquad R_\infty = R_\infty(k) = \sum_{l=1}^{C} \sum_{\mathbf{l} \notin \mathscr{L}_l} \int_S \hat{P}_l(\mathbf{x}) \hat{P}_{l^1}(\mathbf{x}) \cdots \hat{P}_{l^k}(\mathbf{x}) f(\mathbf{x})\, d\mathbf{x},$$

where

$$\hat{P}_l(\mathbf{x}) \triangleq \mathbb{P}\{L = l \mid \mathbf{x}\} = \frac{P_l f_l(\mathbf{x})}{f(\mathbf{x})}$$

denotes the posterior probability of class l, given feature vector $\mathbf{x}$. Moreover, they showed that R_∞ is related to the minimum achievable risk R_B of the Bayes classifier through the two-sided inequalities

$$R_B \le R_\infty(k) \le R_\infty(1) \le 2R_B\left(1 - \frac{C}{C-1} R_B\right).$$

When the Bayes risk is small, as is the case in several important pattern classification problems such as character recognition, the nearest neighbor classifier exhibits an asymptotically optimal character. The utility of this nonparametric approach as a *practical* classifier is, however, tempered by how rapidly the finite-sample risk R_m converges to R_∞. The next two examples demonstrate indeed that the rate of convergence of R_m to R_∞ depends critically on the underlying classification problem, and, moreover, can vary over a considerable range.

EXAMPLE 1 (Nonoverlapping distributions [3]). Consider a two-class problem in the n-dimensional feature space $\mathbb{R}^n$. For $l = 1, 2$, suppose f_l has probability-one support in a compact set A_l in $\mathbb{R}^n$. Suppose further that the distance between the sets A_1 and A_2 is larger than the diameter of either set, that is, $d(A_1, A_2) > \max\{\mathrm{diam}(A_1), \mathrm{diam}(A_2)\}$. As the two classes have nonoverlapping probability-one supports, it is clear that $R_\infty(k) = R_B = 0$ for every positive integer k. Moreover, the finite-sample risk R_m of the k nearest neighbor classifier approaches its infinite-sample limit R_∞ exponentially fast for fixed k as m increases. Indeed, for definiteness suppose that the two classes have equal a priori probabilities, $P_1 = P_2 = 1/2$, and that k is an odd integer. The classifier will make a mistake on a given class if, and only if, there are fewer than $k/2$ examples of that class in the reference sample, whence

$$R_m = 2^{-m} \sum_{i=0}^{(k-1)/2} \binom{m}{i} = \mathscr{O}(m^{(k-1)/2} 2^{-m}), \qquad m \to \infty.$$

Note that the exponentially fast rate of convergence of R_m to 0 is independent of the feature space dimension n.

Example 2 (Discrete distributions [2]). Let the feature space consist of the positive integers $1, 2, 3, \dots,$ and assume that the integer i is selected with probability $p_i = c/i^{1+\varepsilon}$ with $\varepsilon > 0$ and

$$c = c(\varepsilon) = \left(\sum_{i=1}^{\infty} \frac{1}{i^{1+\varepsilon}} \right)^{-1}.$$

We now suppose that each integer i is assigned to one of two classes according to the outcome of an independent toss of a fair coin. As each integer corresponds to a unique class, the Bayes risk is zero. Suppose now that a sample of m reference integers and a single test integer are drawn by independent sampling from the discrete distribution $\{p_i\}$ and the test integer is labeled according to the class of the integer in the reference sample closest to it (resolve ties in any fashion, say, pick the class of the smaller integer). A classification error can occur only if the test pattern does not appear in the m-sample. Consequently, the risk of the nearest neighbor classifier is

$$R_m = \tfrac{1}{2} \sum_{i=1}^{\infty} p_i (1 - p_i)^m.$$

Note that the function $x(1-x)^m$ is concave in the unit interval, and assumes a maximum value at $x = 1/(m+1)$. Let j be an integer for which $p_j < 1/(m+1)$. Specifically, this requires that $j > (c(m+1))^{1/(1+\varepsilon)}$. Then,

$$\begin{aligned} R_m &\geq \frac{1}{2} \sum_{i=j}^{\infty} p_i (1-p_i)^m \geq \frac{1}{2} \left(\frac{m}{m+1} \right)^m \sum_{i=j}^{\infty} p_i \\ &\geq \frac{1}{2(1+1/m)^m} \int_j^{\infty} \frac{c}{x^{1+\varepsilon}}\, dx \geq \frac{c}{2e\varepsilon} j^{-\varepsilon}. \end{aligned}$$

Thus, $R_m \geq K m^{-\varepsilon/(1+\varepsilon)}$ where $K = K(\varepsilon)$ is independent of m. The risk hence converges to 0 at an arbitrarily slow rate as $m \to \infty$ when $\varepsilon > 0$ is small enough.

The wide disparity in the convergence rates exhibited in these two examples suggests that it may be futile to seek a general evaluation of (6) without the imposition of suitable regularity conditions. As we endeavor to describe how the finite-sample risk R_m depends on the sample size m, it seems appropriate to examine this risk integral under smoothness conditions that may more closely correspond to practical applications of this algorithm (e.g.,

[14]) and in the sequel we shall assume that the class-conditional densities in (6) satisfy smoothness constraints, which preclude highly irregular situations such as those exhibited in the second example.

The manner in which the integral in (6) is expressed suggests a method of attack: if the class-conditional distributions, hence the functions $\mathfrak{g}$ and $\mathfrak{h}$, are smooth enough, then R_m may be representable as an asymptotic expansion in fractional powers of m by applying Laplace's method of integration (cf. [5]) in a multidimensional setting. We begin the next section with a list of sufficient conditions that facilitate the exploitation of the technique.

4. Main results. The integral representation (6) for the finite-sample risk is valid when the class-conditional distributions F_l $(1 \le l \le C)$ are absolutely continuous (with corresponding densities f_l). We assume hereafter that the following additional properties are in force.

CONDITIONS. Let $S \subset \mathbb{R}^n$ denote the probability-one support of the mixture density $f = \sum_{l=1}^{C} P_l f_l$:

C1. There exists a positive integer $N \ge 2$ such that, for each $l \in \mathbb{L}$, the class-conditional densities f_l possess uniformly bounded partial derivatives up to order $N+1$ on S.

C2. The mixture density f is bounded away from zero over S. (We can thus assume, without loss of generality, that S is compact.)

C3. All but one of the class-conditional densities vanishes close to the boundary of S. More precisely, let $\partial S = \mathrm{cl}(S) \cap \mathrm{cl}(\mathbb{R}^n \setminus S)$ denote the boundary of S, and for $t \ge 0$ let $\overline{S}_t \subset S$ denote the set of points in S of distance no more than t from the boundary: (For any set $A \subseteq \mathbb{R}^n$ and any point $\mathbf{x} \in \mathbb{R}^n$, we define the point-to-set distance by $d(\mathbf{x}, A) \triangleq \inf_{\mathbf{y} \in A} d(\mathbf{x}, \mathbf{y})$.)

$$\overline{S}_t = \{\mathbf{x} \in S \colon d(\mathbf{x}, \partial S) \le t\}.$$

(Note, e.g., that $\overline{S}_0 = \partial S$.) Then there exists $t_0 > 0$ such that exactly one element of the set $\{f_1(\mathbf{x}), \dots, f_C(\mathbf{x})\}$ is nonzero for all $\mathbf{x} \in \overline{S}_{t_0}$.

EXAMPLE 3 (Radial distributions). Construct $f_l(\mathbf{x}) = N_l \phi_l(\|\mathbf{x}\|_p)$, for $l = 1$, and 2, such that for $0 < r_1 < r_2$, $\phi_1 \in C^{N+1}[0, r_2]$ satisfies

$$\phi_1(\rho)\begin{cases} \ge 0, & \text{if } 0 \le \rho \le r_1, \\ = 0, & \text{if } r_1 < \rho \le r_2, \end{cases}$$

while, for a positive quantity b, $\phi_2 \in C^{N+1}[0, r_2]$ satisfies

$$\phi_2(\rho) \ge b \quad \text{if } 0 \le \rho \le r_2,$$

and N_1 and N_2 are appropriately chosen normalization constants.

We denote by $\mathscr{F}_N$ the family of C-class problems in $\mathbb{R}^n$ for which conditions C1–C3 hold. Our main result, which is proved using Laplace's method in the Appendices, follows.

MAIN THEOREM. *For every $N \geq 2$ and every C-class problem in $\mathscr{F}_N$, the finite-sample risk of any deterministic k nearest neighbor classifier, using a weighted L^p-metric* (4) *to measure class similarity, admits the asymptotic expansion*

$$R_m = R_\infty + \sum_{j=2}^{N} c_j m^{-j/n} + \mathscr{O}(m^{-(N+1)/n}), \qquad m \to \infty, \tag{10}$$

where the coefficients $R_\infty, c_2, \dots, c_N$ are determined by the choices of problem and classifier but are independent of m. Moreover, the expansion (10) *is uniform with respect to the metric parameter p when the choice of metric is varied with the nonsingular linear transformation A fixed and p varying in the range $1 \leq p \leq \infty$.*

REMARK 1. If $c_2 \neq 0$, the rate of convergence of R_m to R_∞ in (10) is $\Theta(m^{-2/n})$. (Following [12], an expression $g(m)$ is said to be $\Theta(f(m))$ if there exist positive constants κ, K, and an integer M, such that $0 < \kappa f(m) < |g(m)| < Kf(m)$, for all $m > M$.) Thus, the coefficient c_2 encapsulates the features of a smooth classification problem that are hardest to learn. For example, c_2 is large if the class-conditional densities are small and vary rapidly around the Bayes decision boundary. While this is the general case, for particular choices of distributions c_2 (and perhaps other coefficients as well) can be identically zero resulting in faster rates of convergence. An extreme instance is illustrated in Example 1 wherein *all* coefficients c_j are identically zero. In this case the theorem yields the estimate $R_m = \mathscr{O}(m^{-(N+1)/n})$ which is true, but trite—as shown in Example 1, the true rate of convergence is $\Theta(m^{-(k-1)/2}2^{-m})$ which decays faster than any polynomial in m^{-1}.

REMARK 2. Note that (10) provides an analytical validation of the curse of dimensionality. In order to maintain $|R_m - R_\infty| < \varepsilon$ for some $\varepsilon > 0$, the leading terms of (10) suggests that the sample size m should scale exponentially with the feature-space dimension n. Note that the curse of dimensionality persists in this algorithm for all fixed values of k. Conversely, if n is fixed, then the requisite sample size scales with ε according to $m \propto \varepsilon^{-n/2}$.

REMARK 3. It is also worth noting two direct consequences of the Main Theorem—one in the direction of increasing smoothness, the other in the direction of decreasing smoothness. Let $\mathscr{F}_\infty$ denote the family of C-class problems in $\mathbb{R}^n$ for which, for each $l \in \mathbb{L}$, the class-conditional densities f_l possess uniformly bounded partial derivatives of all orders, and, in addition, satisfy Conditions C2 and C3. Problems in this class have a complete asymptotic series expansion of the form (1) alluded to in the Introduction. Then, for every C-class problem in $\mathscr{F}_\infty$, the finite-sample risk of any deterministic k nearest neighbor classifier endowed with a weighted L^p-metric has a complete asymptotic series expansion

$$R_m \sim R_\infty + \sum_{j=2}^{\infty} c_j m^{-j/n}, \qquad m \to \infty,$$

where the coefficients $R_\infty, c_2, c_3, \dots,$ are determined by the choices of problem and classifier but are independent of m.

At the other extreme, observe that the sequence $\{\mathscr{F}_N\}$ describes an increasingly smooth family of classification and problems with $\mathscr{F}_N \supset \mathscr{F}_{N+1}$ for every $N \geq 1$. In particular, $\mathscr{F}_2 = \bigcup_{N=2}^{\infty} \mathscr{F}_N$ (while $\mathscr{F}_\infty = \lim_{N\to\infty} \mathscr{F}_N = \bigcap_{N=2}^{\infty} \mathscr{F}_N$). Keeping only the metric parameter p and the sample size m as variable parameters, the uniformity of the expansion (10) with respect to p hence directly yields

$$R_m = R_\infty + c_2 m^{-2/n} + \mathscr{O}(m^{-3/n}), \qquad m \to \infty.$$

Remark 4. We have not attempted to put down the weakest conditions under which an asymptotic expansion of the form (10) will hold for the finite-sample risk. Indeed, only Condition C1 (or something akin to it) may be absolutely necessary in the ensuing development. Example 2 shows, for instance, that without some degree of smoothness arbitrarily slow convergence rates are possible (also see [4]).

Condition C2 is applied at two points in the proof of the theorem in the guise of a uniformity argument. Neither the boundedness of the densities, nor indeed the compactness of their supports, may be strictly necessary, however, for an expansion of the form (10) to hold. [The numerical risk analyses using multivariate normal mixtures in Section 5 suggest that (10) may be valid in a broader context.]

Condition C3 allows us to finesse technical difficulties in the estimation of the risk integral close to the boundary of the support of the mixture density f and, in particular, yields relatively compact expressions for the various expansion coefficients c_j. The following example suggests that Condition C3 is not necessary for an expansion of the form (10) to hold.

Example 4 (Triangular distributions [3]). For a two-class problem, consider the one-dimensional triangular distributions

$$f_1(x) = 2(1-x) \quad \text{and} \quad f_2(x) = 2x$$

over the unit interval $0 \leq x \leq 1$. The finite-sample risk when the classes are equiprobable for $k = 1$ evaluates to

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

More generally, the estimate

$$R_m = \frac{2P_1P_2}{(P_2 - P_1)^2}\left(1 + \frac{2P_1P_2}{P_2 - P_1}\log\frac{P_1}{P_2}\right) + \frac{3(1 - 3P_1P_2)}{8P_1^2P_2^2}m^{-2} + \mathcal{O}(m^{-3}),$$
$$m \to \infty$$

holds provided $P_1 > 0$, $P_2 > 0$ and $P_1 \neq P_2$.

In general, contributions to the risk integral close to the boundary of the support of the mixture density f pose delicate problems in estimation. The form of the expansion (10) is still valid, however, for a rather large class of problems for which the boundary ∂S of the support of the mixture density is smooth: it suffices if, for instance, for each point $\mathbf{x} \in \partial S$, near $\mathbf{x}$ the boundary is the graph of a smooth function. Informally, the finite-sample risk admits an asymptotic expansion of the form (10) if the mixture density f is smooth and has a compact probability-one support S whose boundary is smooth. From the leading terms in (10) derived under Conditions C1–C3 we demonstrate below how the choice of metric effects the finite-sampling risk. However, it is not clear what role the choice of metric plays in the general case.

REMARK 5. The Main Theorem can be generalized to k nearest neighbor classifiers that use random tie-breaking algorithms. This is a direct consequence of the fact that the linear combination of two asymptotic expansions in common fractional powers can be represented as a single asymptotic expansion. In this case, the coefficients in the sum are commensurate linear combinations of the corresponding coefficients in the original expansions.

We now examine how the leading terms in (10) depend upon the metric parameters p and A in (4). First, observe that the leading term, R_∞ defined in (9), does not depend upon the choice of metric. Thus the entire asymptotic influence of the metric parameters A and p is revealed in the coefficient c_2. To facilitate this analysis, let a_{st}^- denote the element in the sth row and tth column of the matrix A^{-1}, the inverse of the weight matrix A in (4), and introduce the "Laplacian-style" differential operator

$$\mathfrak{D}_2^2 \triangleq \sum_{r=1}^{n}\sum_{s=1}^{n}\sum_{t=1}^{n} a_{rt}^- a_{st}^- \frac{\partial^2}{\partial x_r \partial x_s}.$$

The operator $\mathfrak{D}_2^2$ is actually a special case of a more general differential operator, $\mathfrak{D}_p^{2r}$, to be defined in the sequel. If $A = I$, the identity matrix, then $\mathfrak{D}_2^2$ reduces to ∇^2, the Laplacian differential operator. In the proof of the Main Theorem, we explicitly evaluate c_2 and show it to be of the form

$$c_2 = D_n(p)\mathfrak{C}_n(k, A), \tag{11}$$

where, with the usual notation $\Gamma(x) \triangleq \int_0^\infty t^{x-1}e^{-t}\,dt$ for Euler's gamma function,

$$D_n(p) \triangleq \frac{\Gamma(3/p + 1)\Gamma(n/p + 1)^{1+2/n}}{\Gamma((n+2)/p + 1)\Gamma(1/p + 1)^3},$$

and, with the convention that $\hat{P}_l(\mathbf{x})/f_l(\mathbf{x})$ is to be identified as $P_l/f(\mathbf{x})$ for all $\mathbf{x} \in S$, (this is merely a notational convenience extending the Bayes identity $\hat{P}_l(\mathbf{x})/f_l(\mathbf{x}) = P_l/f(\mathbf{x})$ which is well defined for all points $\mathbf{x} \in S$ for which $f_l(\mathbf{x}) > 0$ to points $\mathbf{x} \in S$ for which $\hat{P}_l(\mathbf{x}) = f_l(\mathbf{x}) = 0$)

$$\mathfrak{C}_n(k, A) = \frac{\Gamma(k+1+2/n)}{24\Gamma(k)} \sum_{l=1}^{C} \sum_{\mathbf{l} \notin \mathscr{L}_l} \int_S \hat{P}_l(\mathbf{x}) \hat{P}_{l^1}(\mathbf{x}) \cdots \hat{P}_{l^k}(\mathbf{x}) f(\mathbf{x})^{1-2/n} \times \left(\frac{1}{f_{l^k}(\mathbf{x})} \mathfrak{D}_2^2 f_{l^k}(\mathbf{x}) - \frac{1}{f(\mathbf{x})} \mathfrak{D}_2^2 f(\mathbf{x}) \right) d\mathbf{x}.$$

Observe that $\mathfrak{C}_n(k, A)$ is independent of p and m and determined solely by k, the weight matrix A, the dimension n, the distributions engendered by the problem class and the assignment partition of the classifier.

Note that the entire dependence of c_2 on the metric parameter p is confined to the factor $D_n(p)$. Graphs of $D_n(p)$ for $n = 1, 2, 5, 12$ and 25, and

$$D_\infty(p) \triangleq \lim_{n \to \infty} D_n(p) = \frac{\Gamma(3/p + 1)}{\Gamma(1/p + 1)^3} e^{-2/p},$$

appear on a semilog scale in Figure 2. As all one-dimensional L^p-metrics are equivalent, we obtain $D_1(p) = 1$ for all values of p. It is not difficult to show that for all values of n,

$$D_n(2) = \frac{12\Gamma(n/2 + 1)^{2/n}}{\pi(n+2)} \le D_n(p) \le \lim_{p \to \infty} D_n(p) = 1.$$

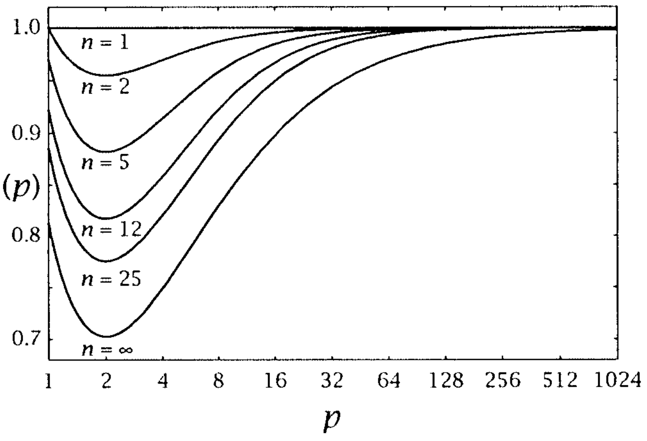

FIG. 2. *The variation of the metric factor $D_n(p)$ for different L^p-metrics. For each dimension $n > 1$, a global minimum occurs at $p = 2$, corresponding to a Euclidean metric.*

Thus, as Figure 2 indicates, $D_n(p)$ has a global minimum at $p = 2$ if $n > 1$. The minimum of $D_n(p)$ is most pronounced in high dimensions as $D_\infty(2) < D_n(2) \leq D_1(2) = 1$, where $D_\infty(2) = 6/(\pi e) \approx 0.702598$.

For any fixed weight matrix A and sufficiently large m, the term $c_2 m^{-2/n}$ will uniformly dominate the sum of the remaining terms in (10) for all admissible L^p-metrics ($1 \leq p \leq \infty$). In this case the finite-sample risk is practically minimized by selecting the Euclidean metric, since R_∞ is independent of the choice of metric. The benefits of optimal metric selection wane, however, if the reference sample is extremely large: although the optimal value p_m^* of p tends to 2 as $m \to \infty$, this limit is extremely degenerate as $\lim_{m\to\infty} R_m = R_\infty$ for all metrics; maximal benefits accrue for a finite intermediate range of m, the benefits being most pronounced in high dimensions. The graph of $D_n(p)$ also demonstrates that with A fixed, the weighted maximum-component (i.e., L^∞) metric is the least efficient weighted L^p-metric, especially in high-dimensional feature spaces. This result should complement other studies that demonstrate the computational efficiency of the L^∞-metric for nearest neighbor algorithms (cf. [7]).

Using the Euler–Lagrange multiplier theorem, it is also possible to determine the weight matrix A that minimizes the value of the coefficient c_2. For example, to find the optimal A for a given two-class problem, one can minimize

$$c_2 = c_2(A) = \sum_{r=1}^{n} \sum_{s=1}^{n} \sum_{t=1}^{n} a_{rs}^{-} a_{ts}^{-} H_{rt},$$

where

$$H_{rt} = D_n(p) \frac{\Gamma(k+1+2/n)}{24\Gamma((k+1)/2)^2} \int_S \left(\hat{P}_1 \hat{P}_2\right)^{(k+1)/2} \left(\hat{P}_2 - \hat{P}_1\right) \times \left(\frac{1}{f_1} \frac{\partial^2 f_1}{\partial x_r \partial x_t} - \frac{1}{f_2} \frac{\partial^2 f_2}{\partial x_r \partial x_t} \right) f^{1-2/n}\, d\mathbf{x},$$

subject to the constraint $\det A^{-1} = \det A = 1$. Write A_{rs}^{-1} for the $(n-1) \times (n-1)$ matrix obtained by removing the rth row and sth column from A^{-1}, and observe that we can write the constraint in the form

$$\det A^{-1} = \sum_{s=1}^{n} (-1)^{r+s} a_{rs}^{-1} \det A_{rs}^{-1} = 1,$$

for any $1 \leq r \leq n$. Introducing the Lagrange multiplier μ, we find that the optimal choice of weight matrix A satisfies

$$2 \sum_{t=1}^{n} H_{rt} a_{ts}^{-} = \mu(-1)^{r+s} \det A_{rs}^{-1} = \mu a_{sr},$$

for every $1 \leq r$, $s \leq n$, or equivalently, in matrix form, $2HA^{-1} = \mu A^T$. Since, $\det A = \det A^T = 1$, we find that $\mu = 2\sqrt[n]{\det H}$, whence,

$$A^T A = \frac{H}{\sqrt[n]{\det H}} \triangleq \hat{H}.$$

This equation is simple under the additional assumption that A is symmetric, whence A and $\hat{H}$ are diagonalized simultaneously. Thus, if $\hat{H} = R^T \Lambda R$, where R is an orthogonal matrix, and Λ is diagonal, then the optimal transformation is $A = R^T \Lambda^{1/2} R$.

5. Numerical simulations. As the results of the previous sections are asymptotic in nature, it is important to ascertain if they describe k nearest neighbor classifiers with practically realizable reference sample sizes. Thus, we shall empirically examine the finite-sample risk of the nearest neighbor classifier under the L^1, L^2 and L^∞ metrics. (Other experiments using only the L^2 metric have been described in [15] and [17].) In this investigation we consider the recognition problem, consisting of two equiprobable, normally distributed classes. Thus, we assume the class-conditional densities

$$f_j(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2} \left(\left(x_1 + (-1)^j \right)^2 + \sum_{i=2}^{n} x_i^2 \right) \right),$$

for $j = 1$ and 2, with $P_1 = P_2 = 1/2$. Observe this two-class problem violates Conditions C2 and C3, so the theorems in Section 4 may not apply. Nevertheless, the following empirical study suggests that the expected finite-sample risk converges to its infinite-sample limit in a fashion consistent with (10).

As Figure 2 suggests that the asymptotic optimality of the L^2-norm is most pronounced in high-dimensional spaces, we shall consider experiments with $n = 12$ and $n = 25$. In each experiment, values for R_m are estimated empirically for a nearest neighbor classifier using a large number of pseudorandom Bernoulli trials with values of m ranging from 20 through 10,000. In each trial a random test pattern is generated by a simulation of the process described in Section 2. Then, this pattern is classified with a similarly generated independent reference m-sample, using the metric $d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_p$. The empirical risk, $\hat{R}_m$, is then computed as the relative frequency of misclassified patterns. For each estimate, 95% confidence intervals are constructed. A numerical integration of the infinite-sample risk (9) with $C = 2$ and $k = 1$ yields $R_\infty = 0.224800$. Then after plotting $(\hat{R}_m - R_\infty)m^{2/n}$ against $\log_{10} m$, we obtain the results shown in Figure 3.

By scaling the ordinate with the factor $m^{2/n}$, the value of c_2 is revealed from the apparent horizontal asymptotes for each metric. (Note that these plots appear on a semilog scale.) Finally, we note that in both series of plots, the horizontal asymptotes corresponding to the L^∞, L^1 and L^2 metrics appear in decreasing order in agreement with Figure 2.

6. Conclusion. Although much is known about the classification risk of the k nearest neighbor classifier in the infinite-sample limit, $m \to \infty$ (the classic paper is [3]), analytical results concerning its finite-sample behavior are rare and, except for Cover's early paper [2] on the one-dimensional case, are recent (cf. [8] and [13]). In part, this is because of the wide range of

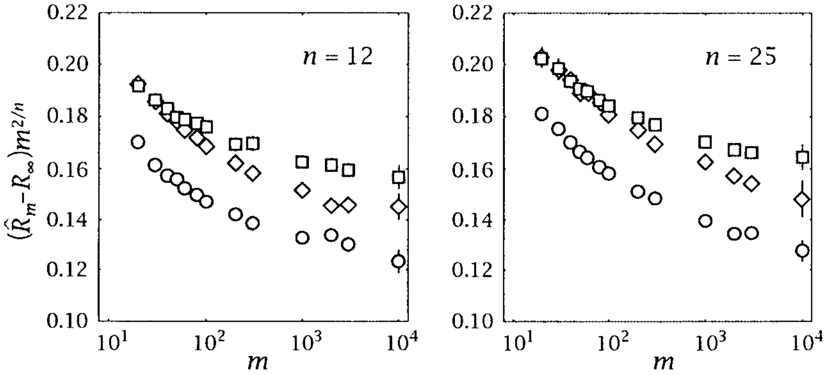

FIG. 3. *Empirical estimates of the function $(\hat{R}_m - R_\infty)m^{2/n}$, as a function of m, for a nearest neighbor classifier with two normally distributed classes in $\mathbb{R}^{12}$ (left) and $\mathbb{R}^{25}$ (right), using the L^1- "◇," L^2- "○," and L^∞- "□" metrics as measures of pattern similarity. Unless shown, the 95% confidence bars for each estimate are smaller than the size of each marker.*

possible dependencies of the misclassification rate on the reference sample size, as seen in Examples 1 and 2 in Section 3.

In this paper, we have derived an asymptotic representation of the finite-sample risk of a k nearest neighbor classifier for a smooth family of classification problems. Simple analytical studies of, for example, the two-class problem with triangular distributions in [3], and multivariate normal mixtures, indicate that these smoothness conditions may be weakened further. In addition to clarifying the finite-sample risk, this asymptotic expansion can be put to practical use by serving as a parameterized model of the finite-sample risk. Thus, the coefficients in (10) can be estimated empirically by constructing ensembles of k nearest neighbor classifiers for a variety of sample sizes. The finite-sample risk of each classifier can then be estimated by resubstitution. Estimates for $R_\infty, c_2, \dots$ can then be obtained by a least-squares algorithm [17]. Finally, the preceding analysis clarifies how the choice of a weighted L^p-metric determines the finite-sample risk and suggests that the Euclidean metric may be the most accurate metric for smooth problems.

It is not at all clear that the most accurate weighted L_p metric is Euclidean for all sample sizes. For problems in $\mathscr{F}_N$ with $n > 1$, $c_3 = 0$ and c_4 appears to depend on p in a complicated manner [16]. Thus for small sample sizes, the optimality of $p = 2$ may depend on the given probability distributions. Analysis of the higher order coefficients in (10) may reveal the extent of these dependencies. Other interesting problems include generalizing the Main Theorem to pattern recognition problems that violate Conditions C2 and C3, and describing the optimal value of k in terms of the reference sample size and the feature space dimension. The latter problem is complicated by the fact that the truncation error of (10) may not be uniform in k.

APPENDIX A

Lemmas. We begin with a collection of technical results that will be needed subsequently in the proof of the Main Theorem. The first result is due to Fulks and Sather [10].

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*

$$G(z) = \int_{\{h \le z\}} g \, d\mathbf{x}.$$

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

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

The following classical result is frequently referred to in the literature as *Watson's lemma* [19]

LEMMA 2. *Let $G(s)$ be a function of the positive real variable s, such that*

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

where τ and μ are positive constants. Then

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

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

Proofs of Lemmas 1 and 2 may be found in the cited references.

For the next two lemmas, let $\mathbf{0}$ and $\mathbf{1}$ denote (column) vectors in $\mathbb{R}^n$ all of whose components are 0 and 1, respectively. Also, write $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_n)^T \in \mathbb{Z}^n$ for a generic lattice vector with integral coordinates. For vectors $\mathbf{u} = (u_1, \dots, u_n)^T$ and $\mathbf{v} = (v_1, \dots, v_n)^T$ in $\mathbb{R}^n$, interpret the vector inequality $\mathbf{u} \ge \mathbf{v}$ to mean that the component-wise inequalities $u_1 \ge v_1, \dots, u_n \ge v_n$ hold simultaneously; likewise, formally define the vector power $\mathbf{u}^{\mathbf{v}} \triangleq \prod_{i=1}^n u_i^{v_i}$. Finally, introduce the nonce notation $Q_p(R) \triangleq \{\mathbf{v}\colon \|\mathbf{v}\|_p \le R, \ \mathbf{v} \ge \mathbf{0}\}$ for the intersection of the L^p-ball of radius R at the origin with the first orthant.

LEMMA 3. *Let $F(p)$ denote any univariate function defined and integrable over the domain $0 \le \rho \le R$. Then for every choice of $p \ge 1$, and every lattice vector $\boldsymbol{\sigma} \ge \mathbf{1}$ with positive, integral coordinates,*

$$\int_{Q_p(R)} F(\|\mathbf{v}\|_p) \mathbf{v}^{\boldsymbol{\sigma}-\mathbf{1}} \, d\mathbf{v} = \frac{\prod_{j=1}^n \Gamma(\sigma_j/p)}{p^{n-1} \Gamma(\|\boldsymbol{\sigma}\|_1/p)} \int_0^R F(\rho) \rho^{\|\boldsymbol{\sigma}\|_1 - 1} \, d\rho.$$

PROOF. The integral is readily evaluated in spherical coordinates in L^p-norm. Set $\mathbf{v} = \rho\boldsymbol{\Omega}$ where $\rho \triangleq \|\mathbf{v}\|_p \geq 0$ is the L^p-norm of $\mathbf{v}$ and $\boldsymbol{\Omega} \triangleq \mathbf{v}/\|\mathbf{v}\|_p$ is a point on the surface of the unit L^p-ball $S_{n-1,p} \triangleq \{\boldsymbol{\Omega} \in \mathbb{R}^n: \|\boldsymbol{\Omega}\|_p = 1\}$. The orientation $\boldsymbol{\Omega} = (\Omega_1, \dots, \Omega_n)^T$ of $\mathbf{v}$ is readily represented by a transformation into spherical coordinates in L^p-norm,

$$\begin{aligned}
\Omega_1 &= \cos^{2/p}(\psi_1)\cos^{2/p}(\psi_2)\cdots\cos^{2/p}(\psi_{n-2})\cos^{2/p}(\psi_{n-1}),\\
\Omega_2 &= \cos^{2/p}(\psi_1)\cos^{2/p}(\psi_2)\cdots\cos^{2/p}(\psi_{n-2})\sin^{2/p}(\psi_{n-1}),\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
\Omega_{n-1} &= \cos^{2/p}(\psi_1)\sin^{2/p}(\psi_2),\\
\Omega_n &= \sin^{2/p}(\psi_1),
\end{aligned}$$

where the angles take values in the ranges $-\pi/2 \leq \psi_1, \dots, \psi_{n-2} \leq \pi/2$ and $-\pi \leq \psi_{n-1} \leq \pi$ as $\boldsymbol{\Omega}$ ranges over $S_{n-1,p}$. It is simple to verify that under this transformation, the Lebesgue element of measure in $\mathbb{R}^n$ can be written in the form $d\mathbf{v} = \rho^{n-1}\,d\rho\,d\boldsymbol{\Omega}$, where

$$d\boldsymbol{\Omega} = \prod_{j=1}^{n-1}\left(\frac{2}{p}\cos^{(2/p)(n-j)-1}(\psi_j)\sin^{2/p-1}(\psi_j)\,d\psi_j\right).$$

As $\|\boldsymbol{\sigma}\|_1 = \sum_{j=1}^n |\sigma_j| = \sum_{j=1}^n \sigma_j$ when $\boldsymbol{\sigma} \geq \mathbf{0}$, we obtain

$$\begin{aligned}
&\int_{Q_p(R)} F(\|\mathbf{v}\|_p)\mathbf{v}^{\boldsymbol{\sigma}-\mathbf{1}}\,d\mathbf{v}\\
&\quad= \frac{2^{n-1}}{p^{n-1}}\int_0^R F(\rho)\rho^{\|\boldsymbol{\sigma}\|_1 - 1}\,d\rho \prod_{j=1}^{n-1}\left(\int_0^{\pi/2}\cos^{(2/p)(\sigma_1+\cdots+\sigma_{n-j})-1}(\psi_j)\right.\\
&\qquad\qquad\qquad\qquad\qquad\qquad\left.\times\sin^{(2/p)\sigma_{n-j+1}-1}(\psi_j)\,d\psi_j\right)\\
&\quad= \frac{2^{n-1}}{p^{n-1}}\left(\int_0^R F(\rho)\rho^{\|\boldsymbol{\sigma}\|_1 - 1}\,d\rho\right)\\
&\qquad\times\prod_{j=1}^{n-1}\frac{\Gamma\big((\sigma_1+\cdots+\sigma_{n-j})/p\big)\Gamma(\sigma_{n-j+1}/p)}{2\Gamma\big((\sigma_1+\cdots+\sigma_{n-j+1})/p\big)}
\end{aligned}$$

Observe that the last product telescopes to complete the proof. □

Setting $F(\rho) = 1$, $R = 1$ and $\boldsymbol{\sigma} = \mathbf{1}$, the above integral yields the volume of the unit L^p-ball in one orthant. We hence have the following corollary.

COROLLARY 1. *The volume of the unit L^p-ball is given by*

$$V_{n,p} \triangleq \int_{\|\mathbf{v}\|_p \leq 1} d\mathbf{v} = 2^n\int_{Q_p(1)} d\mathbf{v} = \frac{2^n\Gamma(1/p)^n}{p^n\Gamma(n/p+1)}.$$

Now, let A be any measure preserving linear transformation on $\mathbb{R}^n$ and let $p \geq 1$ be any real scalar. Write $B_{A,p}(R, \mathbf{x}) \triangleq \{\mathbf{y} \in \mathbb{R}^n \colon \|A(\mathbf{y} - \mathbf{x})\|_p \leq R\}$ for the weighted L^p-ball of radius R at $\mathbf{x}$. (In earlier usages of this notation we had suppressed A and p for simplicity.) Let r be any nonnegative integer. For lattice vectors $\boldsymbol{\sigma} = (\sigma_1, \dots, \sigma_n)^T$ with integral coordinates, write

$$\binom{r}{\boldsymbol{\sigma}} \triangleq \begin{cases} \dfrac{r!}{\sigma_1! \cdots \sigma_n!}, & \text{if } \boldsymbol{\sigma} \geq \mathbf{0} \text{ and } \|\boldsymbol{\sigma}\|_1 = r, \\ 0, & \text{otherwise,} \end{cases}$$

for the multinomial coefficient, and define the differential operator

$$\mathfrak{D}_p^{2r} \triangleq \sum_{\boldsymbol{\sigma} \geq 0} \binom{r}{\boldsymbol{\sigma}} \prod_{j=1}^{n} \left[\frac{\Gamma(1/2)\Gamma\big((2\sigma_j + 1)/p\big)}{\Gamma(1/p)\Gamma\big((2\sigma_j + 1)/2\big)} \left(\sum_{i=1}^{n} a^{-}_{ij} \frac{\partial}{\partial x_i} \right)^{2\sigma_j} \right], \tag{12}$$

where the sum is over lattice vectors $\boldsymbol{\sigma}$ with integral coordinates. Note that the multinomial coefficients force the summands to be nonzero only over a finite range of lattice points. Finally, for nonnegative integers $1 \leq l \leq m$ and $0 \leq r \leq N/2$ and $\mathbf{x}$ ranging over the support S of the mixture density f, define the function

$$\Phi_{l,2r}(\mathbf{x}) \triangleq \frac{nV_{n,p}\Gamma(n/p)}{2^{2r} r!\, \Gamma\big((2r + n)/p\big)} \mathfrak{D}_p^{2r} f_l(\mathbf{x}), \tag{13}$$

where we suppose that the class-conditional densities f_l satisfy Condition C1, and $V_{n,p}$ denotes, as before, the volume of the unit L^p-ball. From the derivation sketched in the Appendix, we then obtain the following lemma.

Lemma 4. *Let $\mathbf{x}$ be any interior point of S and suppose that $R > 0$ is such that the weighted L^p-ball $B_{A,p}(R, \mathbf{x})$ of radius R at $\mathbf{x}$ is contained within S. Let $F(\rho)$ be any univariate function defined and integrable over the domain $\{0 \leq \rho \leq R\}$. Then*

$$\int_{B_{A,p}(R,\mathbf{x})} F(\|A(\mathbf{y} - \mathbf{x})\|_p) f_l(\mathbf{y})\, d\mathbf{y} = \sum_{r=0}^{[N/2]} \Phi_{l,2r}(\mathbf{x}) \int_0^R F(\rho) \rho^{2r+n-1}\, d\rho + \mathfrak{R}(\mathbf{x}),$$

where the remainder term $\mathfrak{R}(\mathbf{x})$ may be bounded uniformly with respect to p and $\mathbf{x}$ by

$$|\mathfrak{R}(\mathbf{x})| \leq K \int_0^R |F(\rho)| \rho^{N+n}\, d\rho$$

for a choice of constant K independent of p and $\mathbf{x}$.

As a special case of the lemma, select $F(\rho) = \rho^\alpha$ to obtain the corollary.

COROLLARY 2. *Suppose $B_{A,p}(R,\mathbf{x}) \subseteq S$. Then, for every $\alpha > -n$,*

$$\int_{B_{A,p}(R,\mathbf{x})} \|A(\mathbf{y}-\mathbf{x})\|_p^\alpha f_l(\mathbf{y})\, d\mathbf{y} = \sum_{r=0}^{\lfloor N/2\rfloor} \frac{R^{\alpha+n+2r}}{\alpha+n+2r}\Phi_{l,2r}(\mathbf{x}) + \Re(\mathbf{x}),$$

where $|\Re(\mathbf{x})| \le K\, R^{\alpha+n+N+1}$ and K is a constant independent of p and $\mathbf{x}$.

PROOF OF LEMMA 4. The linear map A is measure preserving (whence $\det A = 1$) so that with the change of variables $\mathbf{v} = A(\mathbf{y}-\mathbf{x})$, we obtain

$$I \triangleq \int_{B_{A,p}(R,\mathbf{x})} F(\|A(\mathbf{y}-\mathbf{x})\|_p) f_l(\mathbf{y})\, d\mathbf{y} = \int_{\|\mathbf{v}\|_p \le R} F(\|\mathbf{v}\|_p) f_l(\mathbf{x}+A^{-1}\mathbf{v})\, d\mathbf{v}.$$

Since f_l has uniformly bounded derivatives up to order $N+1$, the integrand can be expanded to order N using Taylor's theorem, whence

$$I = \sum_{r=0}^{N} \sum_{\boldsymbol{\sigma}\ge \mathbf{0}} \frac{1}{r!}\binom{r}{\boldsymbol{\sigma}} \left\{ \prod_{j=1}^{n} \left(\sum_{i=1}^{n} a_{ij}^{-} \frac{\partial}{\partial x_i} \right)^{\sigma_j} \right\} f_l(\mathbf{x}) \int_{\|\mathbf{v}\|_p \le R} F(\|\mathbf{v}\|_p) \mathbf{v}^{\boldsymbol{\sigma}}\, d\mathbf{v} + \Re(\mathbf{x}).$$

Note that if $\boldsymbol{\sigma}$ contains an odd integral component, then by symmetry the corresponding integral above vanishes. Here $\Re(\mathbf{x})$ represents the remainder term. After invoking Lemma 3, we obtain

$$I = \sum_{r=0}^{\lfloor N/2\rfloor} \Phi_{l,2r}(\mathbf{x}) \int_0^R F(\rho)\rho^{2r+n-1}\, d\rho + \Re(\mathbf{x})$$

with $\Phi_{l,2r}(\mathbf{x})$ as defined in (13).

To complete the proof, we need to construct an upper bound for the absolute value of the remainder,

$$\Re(\mathbf{x}) = \int_{\|\mathbf{v}\|_p \le R} F(\|\mathbf{v}\|_p)\mathscr{R}(\mathbf{v},\mathbf{x})\, d\mathbf{v},$$

where

$$\mathscr{R}(\mathbf{v},\mathbf{x}) = \sum_{\boldsymbol{\sigma}\ge\mathbf{0}} \frac{1}{N!}\binom{N+1}{\boldsymbol{\sigma}} \mathbf{v}^{\sigma} \int_0^1 \left\{ \prod_{j=1}^{n} \left(\sum_{i=1}^{n} a_{ij}^{-} \frac{\partial}{\partial x_j} \right)^{\sigma_j} \right\} f_l(\mathbf{x}+\xi A^{-1}\mathbf{v})\xi^N\, d\xi,$$

that is uniform in p and $\mathbf{x}$. As f_l has uniformly bounded partial derivatives up through order $N+1$, there exists κ such that

$$\left| \frac{\partial^{N+1} f_l(\mathbf{x})}{\partial x_{j_1} \cdots \partial x_{j_{N+1}}} \right| \le \kappa$$

for all $\mathbf{x} \in S$ and all admissible values of $j_1, \dots, j_{N+1}$. Then

$$|\mathscr{R}(\mathbf{v},\mathbf{x})| \le \frac{\kappa}{N!}\|A^{-1}\mathbf{v}\|_1^{N+1} \int_0^1 \xi^N\, d\xi = \frac{\kappa}{(N+1)!}\|A^{-1}\mathbf{v}\|_1^{N+1}.$$

For $i = 1, \dots, n$, let $(\mathbf{a}_i^{-})^T = (a_{i1}^{-}, \dots, a_{in}^{-})$ denote the ith row of the matrix A^{-1} and let $\mathfrak{a} = \max_{i,j}|a_{ij}^{-}|$. Let p and q be conjugate exponents, $p^{-1} +$

$q^{-1} = 1$. Applying Hölder's inequality, we obtain

$$\|A^{-1}\mathbf{v}\|_1 = \sum_{i=1}^{n} |(\mathbf{a}_i^-, \mathbf{v})| \le \sum_{i=1}^{n} \|\mathbf{a}_i^-\|_q \|\mathbf{v}\|_p \le \sum_{i=1}^{n} \|\mathbf{a}_i^-\|_1 \|\mathbf{v}\|_p \le \alpha n^2 \|\mathbf{v}\|_p,$$

so that

$$|\mathscr{R}(\mathbf{v}, \mathbf{x})| \le \frac{\kappa \alpha^{N+1} n^{2N+2}}{(N+1)!} \|\mathbf{v}\|_p^{N+1}.$$

Consequently,

$$\begin{aligned} |\Re(\mathbf{x})| &= \left| \int_{\|\mathbf{v}\|_p \le R} F(\|\mathbf{v}\|_p) \mathscr{R}(\mathbf{v}, \mathbf{x})\, d\mathbf{v} \right| \\ &\le \int_{\|\mathbf{v}\|_p \le R} |F(\|\mathbf{v}\|_p)|\, |\mathscr{R}(\mathbf{v}, \mathbf{x})|\, d\mathbf{v} \\ &\le \frac{\kappa \alpha^{N+1} n^{2N+2}}{(N+1)!} \int_{\|\mathbf{v}\|_p \le R} |F(\|\mathbf{v}\|_p)|\, \|\mathbf{v}\|_p^{N+1}\, d\mathbf{v} \\ &\le K \int_0^R |F(\rho)|\, \rho^{N+n}\, d\rho, \end{aligned}$$

where $K = (\kappa \alpha^{N+1} n^{2N+3} 2^n)/(N+1)!$. The last step follow from Lemma 3 and the inequality $V_{n,p} \le V_{n,\infty} = 2^n$. Since K does not depend on p or $\mathbf{x}$, the bound is uniform. □

APPENDIX B

Proof of the Main Theorem. The representation (6) of the finite-sample risk in the form $R_m = \int \mathfrak{g} e^{-m\mathfrak{h}}$ suggests that, as in typical Laplace integrals, most of the contribution to the risk arises from a small neighborhood of the point(s) where $\mathfrak{h}$ attains its minimum value. Technical complications arise in the multidimensional setting considered here in part because the function $\mathfrak{h}(d(\mathbf{x}, \mathbf{x}^k), \mathbf{x})$ defined in (8) attains its minimum value of zero on a continuum of points in a linear manifold, and in part because of technical difficulties in evaluating contributions to the risk near the boundary of the support S of the mixture density f. The former difficulty is handled by adapting a technique of Fulks and Sather [10] to carry out an integration over a suitably chosen "skew cylindrical" neighborhood; the latter difficulty is obviated for the family considered here (see Remark 4 following the Main Theorem, however).

Introduce the notation $\rho_\nu \triangleq d(\mathbf{x}, \mathbf{x}^\nu) = \|A(\mathbf{x} - \mathbf{x}^\nu)\|_p$, for $\nu = 1, \dots, k$, and for $t > 0$ define the family of "cylinder" sets $C_t \triangleq \{(\mathbf{x}, \mathbf{x}^k) \in S \times S\colon \rho_k \le t\}$. Observe that $\mathfrak{h}(\rho_k, \mathbf{x})$ attains its minimum value of zero when $\mathbf{x}^k = \mathbf{x}$ so that one would expect the dominant contribution to the risk integral (6) to arise from the set C_t.

Partitioning the domain of integration of the risk integral we obtain

$$R_m = (m)_k \iint_{C_t} \mathfrak{g}(\mathbf{x}, \mathbf{x}^k) \exp(-m\mathfrak{h}(\rho_k, \mathbf{x}))\, d\mathbf{x}^k\, d\mathbf{x}$$
$$+ (m)_k \iint_{(S\times S)\setminus C_t} \mathfrak{g}(\mathbf{x}, \mathbf{x}^k) \exp(-m\mathfrak{h}(\rho_k, \mathbf{x}))\, d\mathbf{x}^k\, d\mathbf{x}.$$

Because of Condition C2, which asserts that the mixture density is uniformly bounded away from zero over its probability-one support S, the function $\mathfrak{h}(\rho_k, \mathbf{x})$ defined by (8) is an increasing function of ρ_k. In particular, for every $t > 0$, there exists an $a > 0$ such that $\mathfrak{h}(\rho_k, \mathbf{x}) > a$ whenever $\rho_k > t$. As $m\mathfrak{h} = (m-k)\mathfrak{h} + k\mathfrak{h} \geq (m-k)a + k\mathfrak{h}$, for $(\mathbf{x}, \mathbf{x}^k) \in (S \times S) \setminus C_t$, the contribution to the risk from the second integral satisfies

$$0 \leq (m)_k \iint_{(S\times S)\setminus C_t} \mathfrak{g}(\mathbf{x}, \mathbf{x}^k) \exp(-m\mathfrak{h}(\rho_k, \mathbf{x}))\, d\mathbf{x}^k\, d\mathbf{x}$$
$$\leq (m)_k \exp(-(m-k)a) \iint_{(S\times S)\setminus C_t} \mathfrak{g}(\mathbf{x}, \mathbf{x}^k) \exp(-k\mathfrak{h}(\rho_k, \mathbf{x}))\, d\mathbf{x}^k\, d\mathbf{x}$$
$$= \mathscr{O}(m^k \exp(-ma)),$$

which will turn out to be exponentially subdominant as $m \to \infty$.

We now distinguish between the boundary and interior contributions to the first integral. For $t > 0$, let $S_t \triangleq S \setminus \overline{S}_t = \{\mathbf{x} \in S\colon d(\mathbf{x}, \partial S) > t\}$, and partition C_t into the two sets $Q_t \triangleq C_t \cap (S_t \times S)$, and $\overline{Q}_t \triangleq C_t \cap (\overline{S}_t \times S)$. Define the interior integral

$$I_m \triangleq (m)_k \iint_{Q_t} \mathfrak{g}(\mathbf{x}, \mathbf{x}^k) \exp(-m\mathfrak{h}(\rho_k, \mathbf{x}))\, d\mathbf{x}^k\, d\mathbf{x}$$

and the boundary integral

$$J_m \triangleq (m)_k \iint_{\overline{Q}_t} \mathfrak{g}(\mathbf{x}, \mathbf{x}^k) \exp(-m\mathfrak{h}(\rho_k, \mathbf{x}))\, d\mathbf{x}^k\, d\mathbf{x}.$$

Separating the interior and boundary contributions to the risk integral we then obtain

$$R_m = I_m + J_m + \mathscr{O}(m^k \exp(-ma)), \qquad m \to \infty.$$

We dispose of the boundary contribution first. Condition C3 asserts that there exists a t_0 such that for all $\mathbf{x} \in \overline{S}_{t_0}$ exactly one class-conditional density is nonzero. Now select t so that $t < t_0/2$. If $(\mathbf{x}, \mathbf{x}^k) \in \overline{Q}_t$, then $\mathbf{x} \in \overline{S}_t \subset \overline{S}_{t_0}$. Moreover, $(\mathbf{x}, \mathbf{x}^k) \in C_t$ implies that $d(\mathbf{x}^k, \mathbf{x}) \leq t$. By the triangle inequality, $d(\mathbf{x}^k, \mathbf{y}) \leq d(\mathbf{x}^k, \mathbf{x}) + d(\mathbf{x}, \mathbf{y})$, for all $\mathbf{y} \in \mathbb{R}^n$. Since $\mathbf{x} \in \overline{S}_t$, we may choose $\mathbf{y} \in \partial S$ such that $d(\mathbf{x}, \mathbf{y}) < t$, whence $d(\mathbf{x}^k, \mathbf{y}) \leq t + t \leq t_0$. Thus, both $\mathbf{x}^k$ and $\mathbf{x}$ lie within $\overline{S}_{t_0}$. By the discussion pertaining to Figure 1, $\mathbf{x}, \mathbf{x}^1, \dots, \mathbf{x}^k \in \overline{S}_{t_0}$. It thus follows that $\mathbb{P}\{L' \neq L \mid (\mathbf{x}, \mathbf{x}^k) \in \overline{Q}_t\} = 0$. Thus, $\mathfrak{g} = 0$, and the boundary contribution, J_m, vanishes.

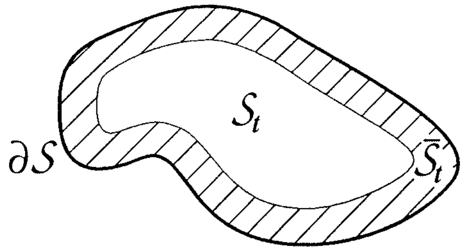

FIG. 4. *A schematic of the support set S, indicating the boundary ∂S, the interior set S_t, and the boundary set $\bar{S}_t$.*

The contribution I_m to the risk from the interior of the support set S can be evaluated asymptotically by adapting the method of [10]. Note that for any pair of interior points $(\mathbf{x},\mathbf{x}^k)\in Q_t$, for every $\nu=1,\dots,k$, the (weighted L^p-) balls

$$B_{A,p}(\rho_\nu,\mathbf{x})=\{\mathbf{y}\in\mathbb{R}^n\colon \|A(\mathbf{y}-\mathbf{x})\|_p\le\rho_\nu\}$$

are contained in S (as $0\le\rho_1\le\cdots\le\rho_k\le t$) so that

$$S^{(\nu)}=B_{A,p}(\rho_\nu,\mathbf{x})\cap S=B_{A,p}(\rho_\nu,\mathbf{x}),\qquad 1\le\nu\le k.$$

Applying Corollary 2, for $1\le\nu\le k-1$, we obtain

$$\int_{S^{(\nu+1)}}\rho_\nu^\alpha f_{l^\nu}(\mathbf{x}^\nu)\,d\mathbf{x}^\nu = \sum_{r=0}^{\lfloor N/2\rfloor}\frac{\rho_{\nu+1}^{\alpha+n+2r}}{\alpha+n+2r}\Phi_{l^\nu,2r}(\mathbf{x})+\mathfrak{o}\left(\rho_{\nu+1}^{\alpha+n+N}\right),\qquad \rho_k\to 0, \tag{14}$$

where the asymptotically small order term is uniform with respect to p.

It will be convenient to represent $\mathbf{x}^k$ in spherical coordinates centered at $\mathbf{x}$: accordingly, set $\mathbf{x}^k=\mathbf{x}+\rho_k\boldsymbol{\Omega}_k$, where, as before, $\rho_k=\|A(\mathbf{x}^k-\mathbf{x})\|_p$ is the weighted L^p-distance between $\mathbf{x}$ and $\mathbf{x}^k$, and $\boldsymbol{\Omega}_k\triangleq(\mathbf{x}^k-\mathbf{x})/\rho_k$ is a point on the surface of the unit weighted L^p-ball at $\mathbf{x}$ which describes the orientation of the point $\mathbf{x}^k$ vis à vis $\mathbf{x}$. Let $\boldsymbol{\sigma}=(\sigma_1,\dots,\sigma_{k-1})^T$ denote a generic $(k-1)$-dimensional lattice vector with integral coordinates. Recursively applying (14) to the nested integrals in (7) we then obtain

$$\mathfrak{g}(\mathbf{x},\mathbf{x}^k)=\sum_{l=1}^{C}\sum_{\mathbf{l}\notin\mathscr{L}_l}\sum_{\substack{\boldsymbol{\sigma}\ge\mathbf{0}\\ \|\boldsymbol{\sigma}\|_1\le N/2}}\frac{P_lP_{l^k}f_l(\mathbf{x})f_{l^k}(\mathbf{x}^k)}{(1-\psi(\rho_k,\mathbf{x}))^k}\times\left\{\prod_{j=1}^{k-1}\frac{P_{l^j}\Phi_{l^j,2\sigma_j}(\mathbf{x})}{jn+2(\sigma_1+\cdots+\sigma_j)}\right\}$$
$$\times\rho_k^{(k-1)n+2\|\boldsymbol{\sigma}\|_1}+\mathfrak{o}\left(\rho_k^{(k-1)n+N}\right)$$

which, on collection of common powers of ρ_k, can be put in the form

$$\triangleq\rho_k^{(k-1)n}\sum_{i=0}^{N}\Lambda_i(\boldsymbol{\Omega}_k,\mathbf{x})\rho_k^i+\mathfrak{o}\left(\rho_k^{(k-1)n+N}\right),\qquad\rho_k\to 0.$$

Observe that the asymptotically small order term is uniform in p.

A corresponding asymptotic expansion can be obtained for the function $\mathfrak{h}(\rho_k, \mathbf{x})$. Observe first that setting $\alpha = 0$ in (14), we obtain the asymptotic expansion

$$\psi(\rho_k, \mathbf{x}) = \int_{B_{A,p}(\rho_k, \mathbf{x})} f(\mathbf{x})\, d\mathbf{x} = \rho_k^n \sum_{r=0}^{\lfloor N/2 \rfloor} \psi_{2r}(\mathbf{x}) \rho_k^{2r} + \mathfrak{o}(\rho_k^{N+n}), \qquad \rho_k \to 0,$$

where, with $V_{n,p}$, the volume of the unit L^p-ball, given by the corollary to Lemma 3, and the differential operator $\mathfrak{D}_p^{2r}$ given by (12),

$$\psi_{2r}(\mathbf{x}) \triangleq \frac{V_{n,p}\Gamma(n/p+1)}{2^{2r} r! \Gamma((2r+n)/p+1)} \mathfrak{D}_p^{2r} f(\mathbf{x}).$$

Starting from (8), the Taylor expansion of the logarithm results in the asymptotic expansion

$$\mathfrak{h}(\rho_k, \mathbf{x}) = \sum_{i=1}^{\infty} \frac{1}{i} \psi(\rho_k, \mathbf{x})^i = \rho_k^n \sum_{i=0}^{N} h_i(\mathbf{x}) \rho_k^i + \mathfrak{o}(\rho_k^{N+n}), \qquad \rho_k \to 0,$$

where the functional forms of the coefficients $h_i(\mathbf{x})$ are determined by the functions $\psi_{2r}(\mathbf{x})$ and depend, in general, upon n, the dimension of the feature space. Again, note that the asymptotically small order term is uniform in p. Table 1 lists expressions for $h_i(\mathbf{x})$ up to third order.

Thus, for sufficiently small ρ_k, we can construct upper and lower bounds $\mathfrak{h}_{\pm}(\rho_k, \mathbf{x})$ for $\mathfrak{h}(\rho_k, \mathbf{x})$ in the form

$$\mathfrak{h}_{\pm}(\rho_k, \mathbf{x}) = \rho_k^n \sum_{i=0}^{N} h_i(\mathbf{x}) \rho_k^i \pm \varepsilon \rho_k^{N+n},$$

where ε is an arbitrary positive constant. We may now choose t from the interval $0 < t \le t_0/2$ sufficiently small, so that, for every $\mathbf{x} \in S$ and $0 < \rho_k \le t$, the following three conditions are simultaneously satisfied:

$$\left| \mathfrak{h}(\rho_k, \mathbf{x}) - \rho_k^n \sum_{i=0}^{N} h_i(\mathbf{x}) \rho_k^i \right| < \varepsilon \rho_k^{N+n};$$

(15)

$$\left| \mathfrak{g}(\mathbf{x}, \mathbf{x}^k) - \rho_k^{(k-1)n} \sum_{i=0}^{N} \Lambda_i(\boldsymbol{\Omega}_k, \mathbf{x}) \rho_k^i \right| < \varepsilon \rho_k^{(k-1)n+N};$$

(16) the functions $\mathfrak{h}_{\pm}(\rho_k, \mathbf{x})$ are bounded away from zero and strictly increasing in ρ_k.

TABLE 1

The low order coefficients $h_i(\mathbf{x})$ in the expansion of $\mathfrak{h}(\rho_k, \mathbf{x})$ expressed in terms of the functions $\psi_{2r}(\mathbf{x})$

Term	$\boldsymbol{n = 1}$	$\boldsymbol{n = 2}$	$\boldsymbol{n \ge 3}$
h_0	ψ_0	ψ_0	ψ_0
h_1	$\frac{1}{2}\psi_0^2$	0	0
h_2	$\psi_2 + \frac{1}{3}\psi_0^3$	$\psi_2 + \frac{1}{2}\psi_0^2$	ψ_2

The last condition is made possible as $h_0(\mathbf{x}) = V_{n,p} f(\mathbf{x})$ is uniformly bounded away from zero on S by Condition C2, and the other coefficients $h_i(\mathbf{x})$, $i \geq 1$, are bounded. Now define I_m^+ and I_m^- by

$$I_m^{\pm} = (m)_k \iint_{Q_t} \mathfrak{g}(\mathbf{x}, \mathbf{x}^k) \exp(-m \mathfrak{h}_{\pm}(\rho_k, \mathbf{x}))\, d\mathbf{x}^k\, d\mathbf{x}.$$

Note that $I_m^+ \leq I_m \leq I_m^-$ as $\mathfrak{g} \geq 0$.

Let us first estimate I_m^+. For $\delta > 0$ chosen suitably small, define the subset of interior points $R_\delta \subset Q_t$ by $R_\delta \triangleq \{(\mathbf{x}, \mathbf{x}^k) \in S_t \times S \colon \mathfrak{h}_+(\rho_k, \mathbf{x}) \leq \delta\}$. We then have

$$\begin{aligned} I_m^+ &= (m)_k \iint_{R_\delta} \mathfrak{g} \exp(-m\mathfrak{h}_+) + (m)_k \iint_{Q_t \setminus R_\delta} \mathfrak{g} \exp(-m\mathfrak{h}_+) \\ &= (m)_k \iint_{R_\delta} \mathfrak{g} \exp(-m\mathfrak{h}_+) + \mathscr{O}(m^k \exp(-ma')) \end{aligned}$$

for a choice of $a' > 0$; the second integral is again exponentially subdominant by an argument similar to that invoked earlier. Now define the function

$$G(s) \triangleq \iint_{\{\mathfrak{h}_+ \leq s\} \cap R_\delta} \mathfrak{g}(\mathbf{x}, \mathbf{x}^k)\, d\mathbf{x}^k\, d\mathbf{x}. \tag{17}$$

Observe that $0 < \exp(-m\mathfrak{h}_+) < 1$ is bounded in R_δ as $\mathfrak{h}_+$ is bounded there, so that $\mathfrak{g} \exp(-m\mathfrak{h}_+)$ is integrable over R_δ. Invoking Lemma 1 and integrating by parts, we hence have

$$\begin{aligned} I_m^+ &= (m)_k \int_0^\delta e^{-ms}\, dG(s) + \mathscr{O}(m^{k+1} e^{-ma'}) \\ &= m(m)_k \int_0^\delta e^{-ms} G(s)\, ds + \mathscr{O}(m^{k+1} e^{-ma''}), \qquad m \to \infty, \end{aligned} \tag{18}$$

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

We now estimate $G(s)$. For $0 \leq s \leq \delta$, first solve the equation $s = h_+(\rho_k, \mathbf{x})$ for $\rho_k = \rho_k(s, \mathbf{x})$. Note that a unique solution exists which is continuous in $\mathbf{x}$ as h_+ is increasing in ρ. We hence need to solve the equation

$$s^{1/n} = \rho_k \left(\sum_{i=0}^{N} h_k(\mathbf{x}) \rho_k^i + \varepsilon \rho_k^N \right)^{1/n}$$

for ρ_k. Deploying Condition (16) above, observe that $s^{1/n}$ is a real analytic function of ρ_k $(0 \leq \rho_k \leq t)$ for each $\mathbf{x}$. We may hence expand $\rho_k(s, \mathbf{x})$ in a Taylor series with remainder, thus obtaining

$$\rho_k(s, \mathbf{x}) = \sum_{k=0}^{N} Y_k(\mathbf{x}) s^{(k+1)/n} + \varepsilon Y_N'(\mathbf{x}) s^{(N+1)/n} + Y_{N+1}(\mathbf{x}, \varepsilon, s) s^{(N+2)/n}, \tag{19}$$

where each Y_k (and Y_N') depends only on the h_j's for $j \leq k$, Y_k (and Y_N') is independent of ε for $k \leq N$ and Y_{N+1} is uniformly bounded for $\mathbf{x} \in S_t$, $0 \leq \varepsilon \leq 1$ and $0 \leq s \leq \delta$. In particular, $Y_0 = h_0^{-1/n}$, $Y_1 = -h_1/(nh_0^{1+2/n})$, and $Y_2 = ((n+3)h_1^2 - 2nh_0h_2)/(2n^2h_0^{2+2/n})$. Now recall that by choice of $\delta > 0$ small enough, within R_δ we can use Condition (15) to write

$$\mathfrak{g}(\mathbf{x}, \mathbf{x}^k) = \rho_k^{(k-1)n} \sum_{i=0}^{N} \Lambda_i(\mathbf{\Omega}_k, \mathbf{x}) \rho_k^i - \varepsilon \mathfrak{g}'_N(\mathbf{x}, \mathbf{x}^k) \rho_k^{(k-1)n+N},$$

where $|\mathfrak{g}'_N(\mathbf{x}, \mathbf{x}^k)| < 1$. We now substitute (19) into the above, and then the resultant into (17). Using Lemma 4 and Corollary 2, the inner integral of $G(s)$ is evaluated. Whence,

$$\begin{aligned} G(s) &= \int_{S_t} \int_{B_{A,p}(\rho_k(s,\mathbf{x}),\mathbf{x})} \mathfrak{g}(\mathbf{x}, \mathbf{x}^k)\, d\mathbf{x}^k\, d\mathbf{x}, \\ &= \int_{S_t} \left(\sum_{j=0}^{N} \lambda_j(\mathbf{x}) s^{k+(j/n)} - \varepsilon \lambda'_N(s, \mathbf{x}) s^{k+(N/n)} \right) d\mathbf{x} + \mathfrak{o}(s^{k+(N/n)}), \end{aligned}$$

where, for instance,

$$\begin{aligned} \lambda_0 &= g_0/\big(knh_0^k\big), \\ \lambda_1 &= \big(nh_0g_1 - (kn+1)g_0h_1\big)/\big(n(kn+1)h_0^{k+1+(1/n)}\big), \\ \lambda_2 &= \big\{2\Gamma(k+(2/n))h_0^2g_2 - 2\Gamma(k+1+(2/n))h_0(h_1g_1 + h_2g_0) \\ &\qquad + \Gamma(k+2+(2/n))h_1^2g_0\big\}/\big\{2n\Gamma(k+1+(2/n))h_0^{k+2+(2/n)}\big\}, \end{aligned}$$

and where $\lambda'_N(s, \mathbf{x})$ is uniformly bounded. (See [16] for details.) The coefficients $g_j(\mathbf{x})$ depend upon the dimensionality n, as shown in Table 2, and in turn upon

$$\begin{aligned} \alpha_0 &= \frac{n}{(k-1)!} V_{n,p}^k f^{k+1} \sum_{l=1}^{C} \sum_{\mathbf{l} \notin \mathscr{L}_l} \hat{P}_l \hat{P}_{l^1} \cdots \hat{P}_{l^k}, \\ \alpha_2 &= \frac{(kn+2)}{4(k-1)!} V_{n,p}^k f^{k+1} \\ &\quad \times \sum_{l=1}^{C} \sum_{\mathbf{l} \notin \mathscr{L}_l} \hat{P}_l \hat{P}_{l^1} \cdots \hat{P}_{l^k} \left[\frac{\Gamma(n/p+1)}{\Gamma((n+2)/p+1)} \left(\frac{\mathfrak{D}_p^2 f_{l^k}}{f_{l^k}} \right) \right]. \end{aligned}$$

TABLE 2
A tabulation of the functions $g_j(\mathbf{x})$ for $j = 0, 1,$ and 2

Term	$n = 1$	$n = 2$	$n \geq 3$
g_0	α_0	α_0	α_0
g_1	$k\alpha_0\psi_0$	0	0
g_2	$\alpha_2 + \binom{k+1}{2}\alpha_0\psi_0^2$	$\alpha_2 + k\alpha_0\psi_0$	α_2

As a consequence of Condition C3 and the choice $t \le t_0/2$, each $g_j(\mathbf{x})$, and hence each $\lambda_j(\mathbf{x})$, is identically zero for all $\mathbf{x}$ in $\overline{S}_t$. Consequently, the domain of the $\mathbf{x}$ integral in the above can be extended from S_t to $S = S_t \cup \overline{S}_t$. Letting,

$$\eta_j = \int_S \lambda_j(\mathbf{x})\, d\mathbf{x},\ 0 \le j \le N, \qquad \eta'_N(s) = \int_S \lambda'_N(s, \mathbf{x})\, d\mathbf{x},$$

we insert the asymptotic expansion for $G(s)$ into (18) to obtain

$$\begin{aligned} I_m^+ &= m(m)_k \int_0^\infty e^{-ms} \left(\sum_{j=0}^{N} \eta_j s^{k+(j/n)} - \varepsilon \eta'_N(s) s^{k+(N/n)} + \mathfrak{o}\left(s^{k+(N/n)}\right) \right) ds \\ &\quad - m(m)_k \int_\delta^\infty e^{-ms} \left(\sum_{j=0}^{N} \eta_j s^{k+(j/n)} - \varepsilon \eta'_N(s) s^{k+(N/n)} + \mathfrak{o}\left(s^{k+(N/n)}\right) \right) ds \\ &\quad + \mathscr{O}\left(m^{k+1} e^{-ma''}\right). \end{aligned}$$

As the integral $\int_\delta^\infty e^{-ms} s^\alpha\, ds$ is exponentially subdominant for $\delta > 0$ and finite α we may neglect the integrals confined to this domain. From Lemma 2 and the identity

$$(m)_k = m(m-1) \cdots (m-k+1) = \sum_{i=0}^{k} \begin{bmatrix} k \\ i \end{bmatrix} (-1)^{k-i} m^i,$$

where $[{k \atop i}]$ denotes a Stirling number of the first kind (i.e., the number of cyclic arrangements of k objects into i cycles), we obtain

$$I_m^+ = \sum_{j=0}^{N} c_j m^{-j/n} - \varepsilon c'_N m^{-N/n} + \mathfrak{o}\left(m^{-N/n}\right), \qquad m \to \infty,$$

where the asymptotically small order term is uniform in p and where $c_0 = R_\infty$ as defined by (9), $c_1 = 0$, and c_2 is defined by (11). (Higher order coefficients appear in [16].) An identical procedure yields

$$I_m^- = \sum_{j=0}^{N} c_j m^{-j/n} + \varepsilon c''_N m^{-N/n} + \mathfrak{o}\left(m^{-N/n}\right), \qquad m \to \infty,$$

as h_- differs from h_+ only in the sign of ε. Thus,

$$I_m^+ - \sum_{k=0}^{N} c_k m^{-k/n} \le I_m - \sum_{k=0}^{N} c_k m^{-k/n} \le I_m^- - \sum_{k=0}^{N} c_k m^{-k/n},$$

while, collecting all the subdominant terms that we had dropped by the wayside, we have

$$R_m = I_m + \mathscr{O}(m^{k+1}e^{-ma'''}), \qquad m \to \infty$$

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

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

the inequalities holding for every $\varepsilon > 0$. Letting $\varepsilon \to 0$ we obtain (10), the expansion uniform in p. □

REFERENCES

[1] Breiman, L., Friedman, J. H., Olshen, R. A. and Stone, C. J. (1984). *Classification and Regression Trees*. Wadsworth & Brooks/Cole, Pacific Grove, CA.

[2] Cover, T. M. (1968). Rates of convergence of nearest neighbor decision procedures. In *Proceedings First Annual Hawaii Conference on Systems Theory* 413–415.

[3] Cover, T. M. and Hart, P. E. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* **13** 21–27.

[4] Devroye, L. (1982). Any discrimination rule can have an arbitrarily bad probability of error for finite sample size. *IEEE Trans. Pattern Anal. Machine Intelligence* **4** 154–157.

[5] Erdélyi, A. (1956). *Asymptotic Expansions*. Dover, New York.

[6] Fix, E. and Hodges, J. L., Jr. (1951). Discriminatory analysis—nonparametric discrimination: consistency properties. Project 21-49-004, Report No. 4. 261–279. USAF School of Aviation Medicine, Randolf Field, TX.

[7] Friedman, J. H., Bentley, J. L. and Finkel, R. A. (1977). An algorithm for finding best matches in logarithmic expected time. *ACM Trans. Math. Software* **3** 209–226.

[8] Fukunaga, K. and Hummels, D. M. (1987). Bias of nearest neighbor estimates. *IEEE Trans. Pattern Anal. Machine Intelligence* **9** 103–112.

[9] Fukunaga, K. and Flick, T. E. (1984). An optimal global nearest neighbor metric. *IEEE Trans. Pattern Anal. Machine Intelligence* **6** 314–318.

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

[11] Hellman, M. E. (1970). The nearest-neighbor classification rule with a reject option. *IEEE Trans. Systems Man Cybernet.* **6** 179–185.

[12] Knuth, D. E. (1976). Big omicron and big omega and big theta. *ACM SIGACT News* **8** 18–23.

[13] Psaltis, D., Snapp, R. R. and Venkatesh, S. S. (1994). On the finite sample performance of the nearest neighbor classifier. *IEEE Trans. Inform. Theory* **40** 820–837.

[14] Smith, S. J., Bourgoin, M. O., Sims, K. and Voorhees, H. L. (1994). Handwritten character classification using nearest neighbor in large databases. *IEEE Trans. Pattern Anal. Machine Intelligence* **16** 915–919, 1994.

[15] Snapp, R. R. and Venkatesh, S. S. (1994). Asymptotic predictions of the finite-sample risk of the k-nearest-neighbor classifier. In *Proceedings of the 12th International Conference on Pattern Recognition* **2** 1–7. IEEE Computer Society Press, Los Alamitos, CA.

[16] SNAPP, R. R. and VENKATESH, S. S. (1998). Asymptotic derivation of the finite-sample risk of the k nearest neighbor classifier. Technical Report UVM-CS-1998-0101, Dept. Computer Science, Univ. Vermont.

[17] SNAPP, R. R. and XU, T. (1996). Estimating the Bayes risk from sample data. In *Advances in Neural Information Processing Systems* **8** (D. S. Touretzky, M. C. Moser, and M. E. Hasselmo, eds.) MIT Press.

[18] STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–645.

[19] WATSON, G. N. (1918). The harmonic functions associated with the parabolic cylinder. *Proc. London Math. Soc.* **17** 116–148.

DEPARTMENT OF COMPUTER SCIENCE
UNIVERSITY OF VERMONT
BURLINGTON, VERMONT 05405
E-MAIL: snapp@cs.uvm.edu

DEPARTMENT OF ELECTRICAL ENGINEERING
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
E-MAIL: venkatesh@ee.upenn.edu