BELLMAN STRIKES AGAIN!
The Growth Rate of Sample Complexity with Dimension for the Nearest Neighbor Classifier

Santosh S. Venkatesh
Electrical Engineering Department
University of Pennsylvania
Philadelphia, PA 19104
venkatesh@ee.upenn.edu

Robert R. Snapp
Department of Computer Science
and Electrical Engineering
University of Vermont
Burlington, VT 05405
snapp@uvm.edu

Demetri Psaltis
Electrical Engineering Department
California Institute of Technology
Pasadena, CA 91125
psaltis@sunoptics.caltech.edu

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, drawn independently from Euclidean n-space according to a fixed probability distribution, is available to the classifier. For a family of smooth distributions, it is shown that the m-sample risk R_m has a complete asymptotic expansion $R_m \sim R_\infty + \sum_{k=1}^{\infty} c_{2k} m^{-2k/n}$, where R_∞ denotes the nearest neighbor risk in the infinite sample limit. Explicit definitions of the expansion coefficients are given in terms of the underlying distribution. As the convergence rate of $R_m \to R_\infty$ dramatically slows down as n increases, this analysis provides an analytic validation of Bellman's curse of dimensionality. Numerical simulations corroborating the formal results are included. The rates of convergence for less restrictive families of distributions are also discussed.

COLT'92-7/92/PA,USA
© 1992 ACM 0-89791-498-8/92/0007/0093...$1.50

1 INTRODUCTION

Because of its simplicity and nearly optimal performance in the large sample limit, the nearest neighbor classifier (see Duda and Hart [DH73], for example) endures as a fundamental algorithm for pattern recognition and machine learning. 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, then a classical result of Cover and Hart [CH67] 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. Practical considerations such as storage and access costs, however, favor small sample sizes. Thus, we are led to the following question of theoretical and practical import: *How rapidly does the risk R_m of a nearest neighbor classifier with a finite reference sam-*

ple of size m approach its infinite sample limit R_∞? For problems with two pattern classes and a one-dimensional feature space, Cover [Cov68] has shown that this limit is approached as rapidly as $R_m = R_\infty + O(m^{-2})$ $(m \to \infty)$ if the underlying distributions are sufficiently smooth. A generalization of this convergence rate to multidimensional feature spaces, however, has been hitherto lacking.

For a family of probability densities with uniformly bounded partial derivatives up to order $N+1$ over their probability-one support, we show that

$$R_m = R_\infty + \sum_{k=1}^{N'} c_{2k} m^{-2k/n} + o\left(m^{-2N'/n}\right) \quad (1)$$

for every $N' \leq N$. Moreover, if every partial derivative is uniformly bounded, then the m-sample risk has a complete asymptotic expansion

$$R_m \sim R_\infty + \sum_{k=1}^{\infty} c_{2k} m^{-2k/n}.$$

In these expressions, the coefficients $(c_2, c_4, \ldots)$ depend upon the underlying probability densities, but not upon m.

We emphasize that in this context the statistical risk is averaged over both the ensemble of input vectors *and* the ensemble of m-samples. This is in marked contrast to many practical applications where a single m-sample is selected, usually by a more deterministic process. Nevertheless, these results prove the existence of an m-sample for which the misclassification rate, when averaged over only the ensemble of input vectors, does not exceed R_m.

Once m is moderately large, the leading terms in the asymptotic summation dominate. Consequently, as m increases, R_m steadily tends to R_∞ via an inverse power law, e.g.

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

Thus for classification problems in low-dimensional spaces, reasonable performance can be attained with a finite sample. In particular, for $n = 1$, the above coincides with Cover's result. Difficulties arise, however, for problems in high-dimensional spaces as the magnitude of the exponent of m decreases with n. For example, if $c_2 \neq 0$ then the sample size m^* required to achieve a specified convergence criterion $|R_m - R_\infty| < \delta$, satisfies

$$m^* > \left(\frac{c_2}{\delta}\right)^{n/2}$$

for sufficiently small $\delta > 0$. If c_2 is bounded away from zero for all n, then m^* grows exponentially with n, validating Bellman's curse of dimensionality [DH73].

In Section 2 we review the nearest neighbor algorithm and its performance in the infinite sample limit. In Section 3 we describe the hypotheses under which our asymptotic results apply, and present a formal statement of our convergence theorem. The proof of the theorem is outlined in Section 4, and includes an explicit procedure for obtaining the expansion coefficients c_2, c_4, etc, in terms of the probability densities. In Section 5 we present numerical evidence that suggests that the convergence theorem may be extended to a broader class of problems than is captured by our hypotheses. Finally, in Section 6 we summarize our results.

2 THE 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* $\mathbf{X} \in \mathbb{R}^n$ which are drawn according to the class-conditional probability distributions F_j, for class $j \in \{1, 2\}$. The mixture distribution

$$F = P_1 F_1 + P_2 F_2$$

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

Labeled feature vectors, $(\mathbf{X}, \theta)$, 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, $\mathbf{P}[\theta = j] = P_j$; then, a feature vector $\mathbf{X} \in \mathbb{R}^n$, conditioned upon θ, is drawn according to F_θ. After m independent repetitions of this process, a reference sample

$$(X_m, \Theta_m) = \{(\mathbf{X}^{(1)}, \theta^{(1)}), \ldots, (\mathbf{X}^{(m)}, \theta^{(m)})\},$$

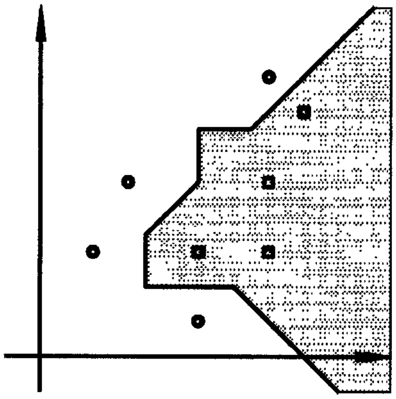

Figure 1: The decision regions for a two-class nearest neighbor classifier for a two-dimensional feature space. Circular and square markers denote the locations of the reference vectors belonging to the two pattern classes. The shaded area describes the region of feature space that the nearest neighbor algorithm assigns to the class of the square markers.

is constructed. Here, $X_m = \{\mathbf{X}^{(1)}, \ldots, \mathbf{X}^{(m)}\}$ ($\mathbf{X}^{(i)} \in \mathbb{R}^n$) denotes the set of m feature vectors, and $\Theta_m = \{\theta^{(1)}, \ldots, \theta^{(m)}\}$ ($\theta^{(i)} \in \{1, 2\}$), the set of corresponding class labels.

Given (X_m, Θ_m), with $m \geq 1$, the nearest neighbor classifier partitions the feature space $\mathbb{R}^n$ as follows:

> ALGORITHM: To every $\mathbf{x} \in \mathbb{R}^n$, assign the class label $C(\mathbf{x}; (X_m, \Theta_m)) \in \{1, 2\}$ given by
>
> $$C(\mathbf{x}; (X_m, \Theta_m)) = \theta^{(i)}$$
>
> if $|\mathbf{x} - \mathbf{X}^{(i)}| \leq |\mathbf{x} - \mathbf{X}^{(j)}|$ for all $j \neq i$.[1]

As the example of Fig. 1 illustrates, the nearest neighbor algorithm induces piecewise linear decision boundaries.

2.1 PERFORMANCE ESTIMATES

A labeled test vector $(\mathbf{X}, \theta)$ is now drawn by the same process, independent of (X_m, Θ_m). Let $(\mathbf{X}', \theta')$ denote the element of (X_m, Θ_m), chosen by the classifier to be the nearest neighbor of $(\mathbf{X}, \theta)$. The classifier's average performance can be quantified in terms of the statistical risk. For simplicity, we take $R_m = \mathbf{P}[\theta' \neq \theta]$, the probability that the nearest neighbor algorithm assigns $\mathbf{X}$ to the incorrect class. Conditioning the event $[\theta' \neq \theta]$ first over the values assumed by the test vector $\mathbf{X}$, and then over the values assumed by the nearest reference vector $\mathbf{X}'$, we obtain

$$\begin{aligned} R_m &= \int_{\mathcal{S}} \mathbf{P}[\theta \neq \theta' \,|\, \mathbf{X} = \mathbf{x}] f(\mathbf{x})\, d\mathbf{x}, \\ &= \int_{\mathcal{S}\times\mathcal{S}} \mathbf{P}[\theta \neq \theta' \,|\, \mathbf{X}' = \mathbf{x}', \mathbf{X} = \mathbf{x}] \\ &\qquad \times f_m(\mathbf{x}' \,|\, \mathbf{x}) f(\mathbf{x})\, d\mathbf{x}'\, d\mathbf{x}, \end{aligned}$$

where $f = P_1 f_1 + P_2 f_2$ denotes the mixture density, $\mathcal{S}$ denotes its probability one support, and $f_m(\mathbf{x}' \,|\, \mathbf{x})$ denotes the conditional density of $\mathbf{X}'$ in a random m-sample given $\mathbf{X} = \mathbf{x}$. Note that the event $[\theta' \neq \theta]$ occurs if and only if one of the disjoint events $[\theta' = 1, \theta = 2]$ or $[\theta' = 2, \theta = 1]$ occurs. Moreover, as the test and reference vectors are chosen independently,

$$\begin{aligned} R_m &= \int_{\mathcal{S}\times\mathcal{S}} \left(\widehat{P_1}(\mathbf{x})\widehat{P_2}(\mathbf{x}') + \widehat{P_1}(\mathbf{x}')\widehat{P_2}(\mathbf{x})\right) \\ &\qquad \times f_m(\mathbf{x}' \,|\, \mathbf{x}) f(\mathbf{x})\, d\mathbf{x}'\, d\mathbf{x}, \qquad (2) \end{aligned}$$

where $\widehat{P_j}(\mathbf{x}) = \mathbf{P}[\theta = j | \mathbf{X} = \mathbf{x}] = P_j f_j(\mathbf{x}) / f(\mathbf{x})$ denotes the posterior probability that the pattern $\mathbf{x}$ belongs to class j. In general, this integral is quite difficult to evaluate. In Sections 3 and 4 we provide conditions under which this integral can be evaluated asymptotically for sufficiently large (but finite) m. First, however, it is instructive to evaluate (2) in the infinite sample limit.

2.2 THE INFINITE-SAMPLE LIMIT

Under fairly general conditions, Cover and Hart [CH67] showed that

$$\lim_{m\to\infty} f_m(\mathbf{x}' \,|\, \mathbf{x}) = \delta(\mathbf{x}' - \mathbf{x}),$$

an n-dimensional Dirac delta function, for a.e. $\mathbf{x} \in \mathcal{S}$. Applying the dominated convergence theorem to (2) yields

$$R_\infty \equiv \lim_{m\to\infty} R_m$$

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

$$
\begin{aligned}
&= 2\int_{\mathcal{S}} \widehat{P_1}(\mathbf{x})\widehat{P_2}(\mathbf{x}) f(\mathbf{x})\,d\mathbf{x} \\
&= 2\mathbf{E}\left[\widehat{P_1}(\mathbf{x})\widehat{P_2}(\mathbf{x})\right]. \qquad (3)
\end{aligned}
$$

This quantity can be quite close to the Bayes risk R_B. By definition, a Bayes classifier minimizes the expected risk. Hence,

$$R_B = \mathbf{E}\left[r_B(\mathbf{x})\right],$$

where

$$
\begin{aligned}
r_B(\mathbf{x}) &= \min\left[\widehat{P_1}(\mathbf{x}), \widehat{P_2}(\mathbf{x})\right] \\
&= \min\left[\widehat{P_1}(\mathbf{x}), 1 - \widehat{P_1}(\mathbf{x})\right]
\end{aligned}
$$

is the conditional Bayes risk. By symmetry, $\widehat{P_1}(\mathbf{x})\widehat{P_2}(\mathbf{x}) = \widehat{P_1}(\mathbf{x})\left(1 - \widehat{P_1}(\mathbf{x})\right) = r_B(\mathbf{x})\left(1 - r_B(\mathbf{x})\right)$, hence, from (3)

$$
\begin{aligned}
R_\infty &= 2\left(\mathbf{E}\left[r_B(\mathbf{x})\right] - \mathbf{E}\left[r_B(\mathbf{x})^2\right]\right) \\
&= 2\left(R_B - \mathrm{Var}\left[r_B(\mathbf{x})\right] - R_B^2\right) \\
&\leq 2R_B(1 - R_B)
\end{aligned}
$$

Furthermore, as the nearest neighbor classifier can never out perform a Bayes classifier, R_∞ is bounded by the inequalities $R_B \leq R_\infty \leq 2R_B$. Thus, if $R_B \ll 1$, then a nearest neighbor classifier with a sufficiently large reference sample is nearly optimal. This optimistic assessment, however, is of practical benefit only if R_m converges to R_∞ at a moderately rapid rate. We now examine some nontrivial conditions under which this may occur.

3 THE RATE OF CONVERGENCE

Before stating our main theorem, we describe the four hypotheses under which we prove the results. In principle, the theorem will hold under weaker conditions, as some of these hypotheses are introduced solely to avoid certain analytic complications.

Hypotheses

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

H2. The mixture density, $f = P_1 f_1 + P_2 f_2$ is bounded away from zero a.e.[2] over its probability one support $\mathcal{S} \subset \mathbb{R}^n$. (We can thus assume, without loss of generality, that $\mathcal{S}$ is compact.)

H3. The class-conditional densities f_k have uniformly bounded derivatives up to order $N+1$ for almost every $\mathbf{x} \in \mathcal{S}$. Explicitly, this condition allows the class-conditional densities to be expressed as a truncated Taylor series with a bounded remainder. For example, with $N = 2$ one obtains,

$$
\begin{aligned}
f_k(\mathbf{x}') = f_k(\mathbf{x}) &+ \sum_{i=1}^{n} \frac{\partial f_k(\mathbf{x})}{\partial x_i}(x_i' - x_i) \\
&+ \frac{1}{2!}\sum_{i=1}^{n}\sum_{j=1}^{n} \frac{\partial^2 f_k(\mathbf{x})}{\partial x_i \partial x_j}(x_i' - x_i)(x_j' - x_j) \\
&+ O(|\mathbf{x}' - \mathbf{x}|^3). \qquad (4)
\end{aligned}
$$

for almost every $\mathbf{x}, \mathbf{x}' \in \mathcal{S}$ and $k = 1, 2$,

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

$$\bar{\mathcal{S}}_t = \{\mathbf{x} \in \mathcal{S} : |\mathbf{x} - \bar{\mathcal{S}}| \leq t\}.$$

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

REMARK: 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.

Hypothesis H4 is introduced to avoid technical complications arising from boundary effects. Basically, if both $\mathbf{x}$ and its nearest neighbor $\mathbf{x}' \in X_m$ happen to fall within $\bar{\mathcal{S}}_{t_0}$, 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 $\bar{\mathcal{S}}$, and would also result in more awkward asymptotic expressions for R_m.

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

Theorem 3.1 *Under Hypotheses H1 - H4, there exists a unique set of constants* $\{c_{2k}\}$ *such that*

$$R_m = R_\infty + \sum_{k=1}^{N/2} c_{2k} m^{-2k/n} + o\left(m^{-N/n}\right)$$

as $m \to \infty$.

Remark: The leading coefficient in the expansion is just the infinite-sample risk defined in (3). For c_2, we obtain [SPV91]

$$c_2 = P_1 P_2 \frac{\Gamma(1+\frac{2}{n})\,\Gamma(1+\frac{n}{2})^{2/n}}{2n\pi} \times \sum_{j=1}^{n} \int_S \frac{1}{[f(\mathbf{x})]^{1+2/n}} \left[f_1(\mathbf{x}) \frac{\partial^2 f_2(\mathbf{x})}{\partial x_j^2} + f_2(\mathbf{x}) \frac{\partial^2 f_1(\mathbf{x})}{\partial x_j^2} - 2 \frac{f_1(\mathbf{x}) f_2(\mathbf{x})}{f(\mathbf{x})} \frac{\partial^2 f(\mathbf{x})}{\partial x_j^2} \right] d\mathbf{x}. \tag{5}$$

For $N = 2$ and $n = 1$ this result is in accordance with Cover's results for the convergence rate of a one-dimensional, nearest neighbor classifier [Cov68].

4 ASYMPTOTIC ANALYSIS

The proof of Theorem 3.1 involves an asymptotic evaluation of the integral appearing in (2) for sufficiently large values of m. First we obtain an explicit expression for the conditional density $f_m(\mathbf{x}' \mid \mathbf{x})$ defined in Section 2.

The event $\mathbf{X}' = \mathbf{x}'$ occurs if one of the reference feature vectors $\mathbf{X}^{(j)}$ assumes the value $\mathbf{x}'$ *and* every other feature vector $\mathbf{X}^{(k)}$, $k \neq j$, assumes a value outside $B(\rho, \mathbf{x})$, the (closed) ball of radius $\rho = |\mathbf{x}' - \mathbf{x}|$ centered at $\mathbf{x}$. (By Hypothesis H1, ties occur with zero probability.) Because of the independent nature of the training set, the latter may occur with $j = 1, 2, \ldots, m$. We thus obtain,

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

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

$$\psi(\rho, \mathbf{x}) = \mathbf{P}\{\mathbf{Y} \in B(\rho, \mathbf{x})\} = \int_{B(\rho,\mathbf{x})} f(\mathbf{y})\, d\mathbf{y}. \tag{6}$$

It follows that

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

In terms of these quantities, (2) can be written as

$$R_m = m \int_{S \times S} g(\mathbf{x}', \mathbf{x}) e^{-mh(\mathbf{x}',\mathbf{x})}\, d\mathbf{x}'\, d\mathbf{x}, \tag{7}$$

where

$$g(\mathbf{x}', \mathbf{x}) = \frac{f(\mathbf{x}') f(\mathbf{x})}{1 - \psi(\rho, \mathbf{x})} \times \left(\widehat{P_1}(\mathbf{x}) \widehat{P_2}(\mathbf{x}') + \widehat{P_1}(\mathbf{x}') \widehat{P_2}(\mathbf{x}) \right), \tag{8}$$

and

$$h(\mathbf{x}', \mathbf{x}) = -\log[1 - \psi(\rho, \mathbf{x})]. \tag{9}$$

Observe the useful result that $h(\mathbf{x}', \mathbf{x})$ depends only on $\mathbf{x}$ and $|\mathbf{x}' - \mathbf{x}|$. Furthermore, Hypothesis H2 indicates the h will have a minimum if and only if $|\mathbf{x}' - \mathbf{x}| = 0$. Thus, for large m, the integral in (7) appears to be in a form amenable to Laplace's asymptotic method, which asserts that the dominant contribution arises from a neighborhood of the point where h has a discrete minimum [Erd56]. 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 method can be extended to multidimensional integrals where, for instance, h has an interior minimum in the domain of integration (cf. Fulks and Sather [FS61]).

The evaluation of (7) by this asymptotic technique, however, is complicated by the fact that the minima of h (defined over $\mathrm{I\!R}^{2n}$) occur on a *continuum* of points in the n-dimensional, linear manifold $\{(\mathbf{x}', \mathbf{x}) \in S \times S \mid |\mathbf{x}' - \mathbf{x}| = 0\}$. Consequently: (*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 has its minima at and near the boundary of the domain of integration pose particular difficulties in evaluation, and depend, in general, on

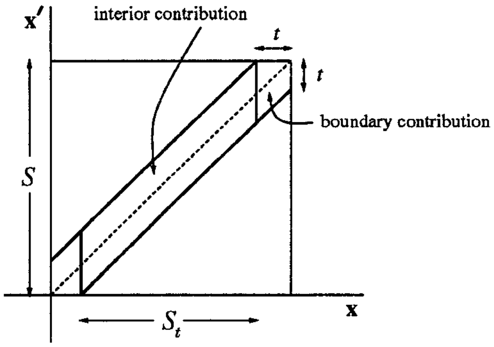

Figure 2: A schematic of the cylinder set C_t, and the interior and boundary contributions to the dominant integral.

the boundary shape. The first difficulty is resolved by adopting a generalized cylindrical coordinate system in the $2n$-dimensional domain $\mathcal{S} \times \mathcal{S}$. Then a technical result of Fulks and Sather [FS61] permits the replacement of (7) by essentially a single Stieltjes integral. The second difficulty is finessed by Hypothesis H4 which essentially eliminates any contribution to the classifier's error rate in the boundary of the domain of integration.

4.1 LAPLACE'S METHOD

Now let us return to a consideration of (7). For $t > 0$, define the family of "cylinder" sets

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

This is schematically the cylindrical area along the diagonal in Fig. 2. We can then partition the integral contribution to R_m into two parts:

$$R_m = m \int_{C_t} g e^{-mh} + m \int_{\mathcal{S}^2 \setminus C_t} g e^{-mh}.$$

For any fixed $t > 0$, the integral contribution from outside C_t is asymptotically subdominant. 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. point in a set $\bar{\mathcal{S}}_{t_0}$ of points in $\mathcal{S}$ whose distance from the boundary $\bar{\mathcal{S}}$ of $\mathcal{S}$ is no more than t_0. Now choose $0 < t \leq t_0/2$, and define the set

$$\mathcal{S}_t = \mathcal{S} \setminus \bar{\mathcal{S}}_t = \{\mathbf{x} \in \mathcal{S} : |\mathbf{x} - \bar{\mathcal{S}}| > t\}.$$

We now partition C_t into the sets

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

We now further partition the dominant integral contribution to R_m according to whether $\mathbf{x}$ takes values in $\mathcal{S}_t$ or $\mathbf{x}$ takes values in $\bar{\mathcal{S}}_t$:

$$\begin{aligned} m \int_{C_t} g e^{-mh} &= m \int_{Q_t} g e^{-mh} + m \int_{\bar{Q}_t} g e^{-mh} \\ &\equiv I_m + J_m, \end{aligned}$$

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 $(\mathbf{x}', \mathbf{x}) \in \bar{Q}_t$, then clearly $\mathbf{x} \in \bar{\mathcal{S}}_t$ by definition of $\bar{Q}_t = C_t \cap (\mathcal{S} \times \bar{\mathcal{S}}_t)$. Furthermore, in C_t we have $|\mathbf{x}' - \mathbf{x}| \leq t$ so that by the triangle inequality we have $|\mathbf{x}' - \bar{\mathcal{S}}| \leq |\mathbf{x}' - \mathbf{x}| + |\mathbf{x} - \bar{\mathcal{S}}| \leq 2t \leq t_0$ by choice of t. Consequently, both $\mathbf{x}$ and $\mathbf{x}'$ will lie in $\bar{\mathcal{S}}_{t_0}$. It follows from Hypothesis H4 that for a.e. $(\mathbf{x}', \mathbf{x}) \in \bar{Q}_t$, $f_j(\mathbf{x}') = f_k(\mathbf{x}) = 0$ for some j, $k \in \{1, 2\}$. Now consider representation (8) for g. It follows that g is identically zero over $\bar{Q}_t$, and hence $J_m = 0$.

Interior Contribution At this point, it is convenient to adopt n-dimensional polar coordinates to represent $\mathbf{x}'$ relative to $\mathbf{x}$. $\rho = |\mathbf{x}' - \mathbf{x}|$ and $\Omega = (\phi_1, \phi_2, \ldots, \phi_{n-1})$, where

$$x_i' - x_i = \rho \sin \phi_{n-i+1} \prod_{l=i}^{n-1} \cos \phi_{n-l},$$

for $i = 1, \ldots, n$. (For $i = 1$, define $\sin \phi_n \equiv 1$.) Here, ϕ_i assumes values in the range $-\pi/2 \leq \phi_i \leq \pi/2$ if $1 \leq i \leq n-2$, or $0 \leq \phi_{n-1} \leq 2\pi$. We will also let S_{n-1} denote the set of admissible values of Ω, i.e. the solid angle of an n-dimensional sphere.

From Hypothesis H3, we can obtain an asymptotic expansion (as $\rho \to 0$) for the mixture density f. Then, after expanding (6), (8), and (9) in powers of ρ, one obtains

$$\begin{aligned} h(\mathbf{x}', \mathbf{x}) &= \rho^n \sum_{k=0}^{N} h_k(\mathbf{x}) \rho^k + o(\rho^{N+n}) \\ g(\mathbf{x}', \mathbf{x}) &= \sum_{k=0}^{N} g_k(\Omega, \mathbf{x}) \rho^k + o(\rho^N) \end{aligned}$$

The integral I_m can now be evaluated by a procedure that parallels reference [FS61]. After a lengthy calculation [PSV92], we obtain

$$I_m = \sum_{k=0}^{N} c_k m^{-k/n} + o\left(m^{-N/n}\right), \qquad (10)$$

where

$$c_k = \Gamma\left(2 + \frac{k}{n}\right)$$
$$\times \int_{S} d\mathbf{x} \int_{S_{n-1}} d\Omega \sum_{j=0}^{k} \frac{g_k(\Omega, \mathbf{x})}{j+n} B_{k-j}^{j+n}(\mathbf{x}).$$

The coefficients B_j^k are generated from the expression

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

where $\Upsilon_l(\mathbf{x})$ represent the exact expansion coefficients of

$$\rho = \sum_{k=0}^{N} \Upsilon_k(\mathbf{x}) s^{(k+1)/n} + o\left(s^{(N+1)/n}\right),$$

so that

$$s = \rho^n \left[\sum_{k=0}^{N} h_k(\mathbf{x})\rho^k\right]$$

for all $\mathbf{x} \in \mathcal{S}$. Note that c_0 evaluates to R_∞. If k is odd, the parity of the integrand over S_{n-1} then ensures that $c_k = 0$, proving the theorem.

5 DISCUSSION

The preceding theorem was proved under a restrictive set of hypotheses so that expressions for the coefficients, R_∞ and c_{2k}, 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 in successive powers of $m^{-2/n}$, the expansion coefficients will be more complex. (However, for some problems, they 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. 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. First, we present an example where the theorem provides 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 the class-conditional densities are given by

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

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 DeMoivre-Laplace limit theorem. Because of computational limitations, large error bars are unavoidable once R_m becomes sufficiently close to R_∞.

This data is compared to the asymptotic expansion obtained from the theorem, which we truncate to second order,

$$R_m \approx R_\infty + c_2 m^{-2/n}. \qquad (11)$$

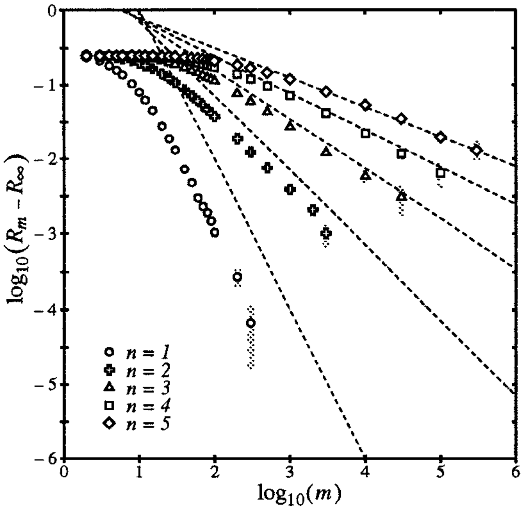

Figure 3: Numerical evidence supporting the nearest neighbor scaling hypothesis for two trigonometric distributions. Here the broken lines describe the leading asymptotic behavior predicted by the asymptotic expansion in the theorem. The convergence thus occurs at rates of order $m^{-2/n}$.

Using the explicit expressions for the coefficients obtained under Hypotheses H1 through H4, a broken curve is plotted for each dimensionality. As Hypothesis H4 is clearly violated, the close agreement between the theoretical and experimental curves is surprising. In this case it suggests that the net boundary contributions to the finite sample risk are very small, if not exponentially subdominant. ∎

The next example suggests that Hypothesis H2 is not necessary to obtain an asymptotic scaling law similar to the one predicted by our convergence theorem.

Example: *Normal Distributions*

Consider the classification problem described in $\mathbb{R}^n$ by the two normal class-conditional densities,

$$f_1(\mathbf{x}) \quad = \quad N_n \exp\left[-\frac{(x_1-\mu)^2+\sum_{j=2}^n x_j^2}{2\sigma^2}\right],$$

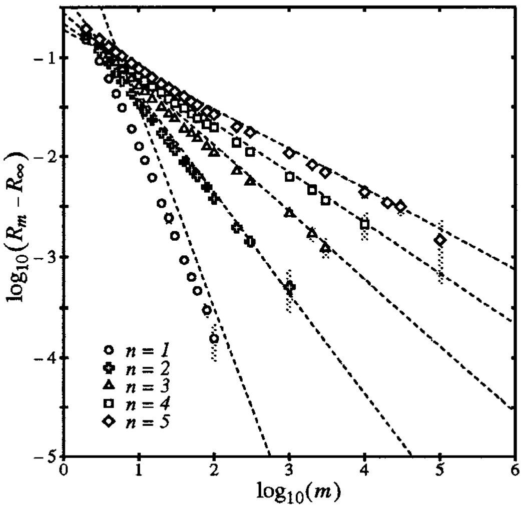

Figure 4: Numerical evidence supporting the nearest neighbor scaling hypothesis for two normally distributed classes in $\mathbb{R}^n$. The data suggests that the risk converges at rates of order $m^{-2/n}$.

$$f_2(\mathbf{x}) \quad = \quad N_n \exp\left[-\frac{(x_1+\mu)^2+\sum_{j=2}^n x_j^2}{2\sigma^2}\right],$$

where $N_n = (2\pi\sigma^2)^{-n/2}$, and prior probabilities, $P_1 = P_2 = 1/2$. Using (3), the risk of the nearest neighbor classifier tends to

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

For $\mu = \sigma = 1$, a numerical integration yields $R_\infty \approx 0.22480$, which is consistent with the Bayes risk, $R_B = (1/2)\operatorname{erfc}(1/\sqrt{2}) \approx 0.15865$.

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 (11), where the coefficient c_2 was chosen to obtain a convincing fit. (For this example, violation of Hypothesis H2 causes the expression for c_2 in (5) to diverge.) The close agreement suggests that an asymptotic expression for R_m, in a form similar to (1), exists. ∎

Finally, we consider what can happen when

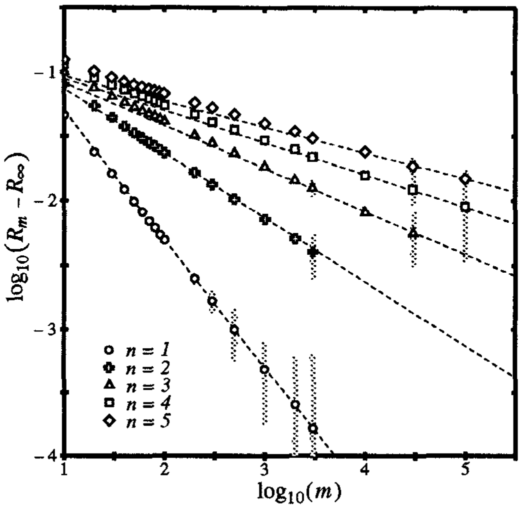

Figure 5: Numerical evidence supporting the nearest neighbor scaling hypothesis for two, nonoverlapping, uniformly distributed classes in $\mathbb{R}^n$. The broken lines depict curves of the form $R_m = am^{-1/n}$, where the coefficient a was chosen to obtain a convincing fit.

the smoothness condition, Hypothesis H3, is relaxed.

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

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

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

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 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 $m^{-1/n}$. This suggests that problems described by densities with piecewise discontinuities are more difficult to learn. The necessity of some degree of smoothness for rapid convergence is underscored by Cover's construction of a discrete classification problem for which the nearest neighbor classifier converges at an arbitrarily slow rate, even in one dimension [Cov68]. ∎

6 CONCLUSION

The examples developed in the previous section indicate that the conditions under which Theorem 3.1 was proved are not strictly necessary. While some degree of smoothness (such as Hypotheses H1 and H3) appears to be necessary if reasonable performance is to be attained, it may be possible to relax the constraints imposed by Hypotheses H2 and H4. The difficulties here appear to be purely technical involving boundary effects that may not be resolvable in a general way.

In fine, the complete asymptotic series expansions for the finite-sample risk developed in the theorem allow us to not just investigate the large sample behavior of the classifier, but the small sample behavior as well. Indeed, our numerical simulations indicate that the leading asymptotic terms may dominate the convergence of R_m for practical sample sizes. 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 $m^{-2/n}$ (or, even slower if the distributions are not smooth enough). 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.

References

[CH67] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, IT–13:21–27, 1967.

[Cov68] T. M. Cover. Rates of convergence of nearest neighbor decision procedures. In *Proceedings of the First Annual Hawaii Conference on Systems Theory*, pages 413–415, 1968.

[DH73] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* John Wiley & Sons, New York, 1973.

[Erd56] A. Erdélyi. *Asymptotic Expansions.* Dover, New York, 1956.

[FS61] W. Fulks and J. O. Sather. Asymptotics II: Laplace's method for multiple integrals. *Pacific Journal of Mathematics*, 11:185–192, 1961.

[PSV92] D. Psaltis, R. R. Snapp, and S. S. Venkatesh. On the finite sample performance of the nearest neighbor classifier. Submitted, 1992.

[SPV91] R. R. Snapp, D. Psaltis, and S. S. Venkatesh. Asymptotic slowing down of the nearest-neighbor classifier. In R. Lippman, J. Moody, and D. Touretzky, editors, *Advances in Neural Information Processing Systems, 3*, pages 932–938, San Mateo, CA, 1991. Morgan Kaufmann.