functions," *IEEE Trans. Inform. Theory*, vol. IT-15, pp. 525–531, 1969.
[12] A. Crolotte and J. Pearl, "Asymptotic rate-distortion functions for coding precedence relations," *IEEE Trans. Inform. Theory*, vol. IT-25, no. 1, pp. 80–82, Jan. 1979.
[13] A. Crolotte, "Memory versus error tradeoffs in question-answering systems," Ph.D. dissertation, Univ. California, Los Angeles, UCLA-ENG-7753, July 1977.

Distribution-Free Inequalities for the Deleted and Holdout Error Estimates

LUC P. DEVROYE AND TERRY J. WAGNER, MEMBER, IEEE

***Abstract*—In the discrimination problem the random variable θ, known to take values in $\{1,\cdots,M\}$, is estimated from the random vector X taking values in $\mathbb{R}^d$. All that is known about the joint distribution of (X,θ) is that which can be inferred from a sample $(X_1,\theta_1),\cdots,(X_n,\theta_n)$ of size n drawn from that distribution. A discrimination rule is any procedure which determines a decision $\hat{\theta}$ for θ from X and $(X_1,\theta_1),\cdots,(X_n,\theta_n)$. The rule is called k-local if the decision $\hat{\theta}$ depends only on X and the pairs (X_i,θ_i), for which X_i is one of the k closest to X from $X_1,\cdots,X_n$. If L_n denotes the probability of error for a k-local rule given the sample, then estimates $\hat{L}_n$ of L_n are determined for which $P\{|\hat{L}_n-L_n|\geq\epsilon\}\leq A\exp(-Bn)$, where A and B are positive constants depending only on d, M, and ϵ.**

I. Introduction

In the discrimination problem a statistician makes an *observation* X, a random vector with values in $\mathbb{R}^d$, and wishes to estimate its *state* θ, a random variable known to take values in $\{1,\cdots,M\}$. All that he knows about the distribution of (X,θ) is that which can be inferred from a sample $(X_1,\theta_1),\cdots,(X_n,\theta_n)$ of size n drawn from that distribution. The sample, commonly called *data*, is assumed to be independent of (X,θ). Using X and the data, the statistician makes a decision $\hat{\theta}$ for θ where his *rule* is any procedure which determines $\hat{\theta}$ from X and the data.

The rules which we are interested in are called *k-local rules*. Here the estimate $\hat{\theta}$ is a function of X, the k nearest observations to X from $X_1,\cdots,X_n$, and the states of these k nearest observations. Because there may be ties in determining the k nearest observations to X, we use an independent sequence of random variables $Z,Z_1,Z_2,\cdots$ which itself is an independent identically distributed (i.i.d.) sequence with a uniform distribution on [0, 1]. Then X_i is nearer than X_j to X if

a) $\|X-X_i\|<\|X-X_j\|$, or
b) $\|X-X_i\|=\|X-X_j\|$ and $|Z-Z_i|<|Z-Z_j|$, or
c) $\|X-X_i\|=\|X-X_j\|$, $|Z-Z_i|=|Z-Z_j|$, and $i<j$.

The event c) has probability zero and can be ignored. We will think of Z as being attached to X and of Z_i as being attached to X_i for $i=1,\cdots,n$. If a new independent observation X' is to have its state estimated, another random variable Z' is generated, but $Z_1,\cdots,Z_n$ remain the same. If (X^i,θ^i,Z^i) denotes the ith-nearest observation, its state, and its attached random variable, respectively, then a *k-local rule* is any rule for which

$$\hat{\theta}=g(X,Z,(X^1,\theta^1,Z^1),\cdots,(X^k,\theta^k,Z^k))$$

for some measurable function g.

The most familiar example of a k-local rule is the k-nearest neighbor rule [1], in which $\hat{\theta}$ is taken to be the state which occurs most often among the k nearest observations to X. In the event that several states tie in this respect, $\hat{\theta}$ is taken to be the state from those tied with the nearest observation to X.

For a k-local rule and the given data, the probability of error is

$$L_n=P\{\hat{\theta}\neq\theta|V_n\}$$

where

$$V_n=((X_1,\theta_1,Z_1),\cdots,(X_n,\theta_n,Z_n)).$$

The value taken by the random variable L_n is just the limiting frequency of errors made when a large number of independent observations have their states estimated with the given rule and data. Since L_n measures the effectiveness of the rule, and since it cannot be computed, the immediate need of the statistician is to estimate it as accurately as possible. Suppose, for example, n additional observations and their states $(X_{n+1},\theta_{n+1}),\cdots,$

Manuscript received June 28, 1977; revised December 1, 1977. This work was supported in part by the U.S. Air Force under Grant AFOSR 77-3385. This paper was presented at the IEEE International Symposium on Information Theory, Cornell University, Ithaca, NY, October 10–14, 1977.

L. P. Devroye is with the School of Computer Science, McGill University, P.O. Box 6070, Station A, Montreal, PQ, Canada H3C 3G1.
T. J. Wagner was with the Department of Electrical Engineering, Rice University, Houston, TX. He is now with the Department of Electrical Engineering, Univeristy of Texas, Austin, TX 78712.

0018-9448/79/0300-0202$00.75

(X_{n+m},θ_{n+m}) are available. One then could estimate L_n by the empirical frequency count

$$\hat{L}_n=\frac{1}{m}\sum_1^m I_{\{\hat{\theta}_{n+i}\neq\theta_{n+i}\}}$$

where $\hat{\theta}_{n+i}$ is the estimate of θ_{n+i} from X_{n+i}, Z_{n+i}, and V_n and where $I_{\{\cdot\}}$ denotes the indicator function of the event $\{\cdot\}$. This estimate will be close to L_n if m is large, and since, conditioned on the data, the sequence $I_{\{\hat{\theta}_{n+1}\neq\theta_{n+1}\}},\cdots,I_{\{\hat{\theta}_{n+m}\neq\theta_{n+m}\}}$ is a sequence of Bernoulli trials with expectation L_n, we have, using Hoeffding's inequality [2],

$$P\{|L_n-\hat{L}_n|\geq\epsilon|V_n\}\leq 2e^{-2m\epsilon^2}. \tag{1}$$

Inequality (1) is interesting because it does not depend on the data or a specific knowledge of the distribution of (X,θ) and, at the same time, is reasonably tight. The difficulty with this estimate is that one rarely has m additional observations and states, and even if they were available, they would be included in the data. One wants then an estimate $\tilde{L}_n$ of L_n which depends only on the data and for which $P\{|\tilde{L}_n-L_n|\geq\epsilon\}$ can be upper-bounded by an expression which depends only on known quantities (e.g., n, d, M, ϵ) and tends to zero with n as fast as possible. In short, one wants an estimate $\tilde{L}_n$ which yields a good distribution-free upper bound for $P\{|\tilde{L}_n-L_n|\geq\epsilon\}$. We note here that since L_n and $\tilde{L}_n$ are both functions of the data it is no longer possible to find such bounds for $P\{|\tilde{L}_n-L_n|\geq\epsilon|V_n\}$.

If $\hat{\theta}(V_n,X,Z)$ denotes the function which first finds the k nearest points to X,Z from V_n and then uses g to get the value of $\hat{\theta}$, we can write the *resubstitution estimate* L_n^R, the *deleted estimate* L_n^D, and the *holdout estimate* L_n^H as

$$L_n^R=\frac{1}{n}\sum_1^n I_{\left[\hat{\theta}(V_n,X_i,Z_i)\neq\theta_i\right]},$$

$$L_n^D=\frac{1}{n}\sum_1^n I_{\left[\hat{\theta}(V_{n,i},X_i,Z_i)\neq\theta_i\right]},$$

and

$$L_n^H=\frac{1}{s}\sum_1^s I_{\left[\hat{\theta}(V_n'',X_i,Z_i)\neq\theta_i\right]},$$

respectively, where $V_{n,i}$ denotes the sequence V_n with (X_i,θ_i,Z_i) deleted and where $V_n''=(X_{s+1},\theta_{s+1},Z_{s+1}),\cdots,(X_n,\theta_n,Z_n)$, for $n>s\geq 1$. Notice that we must have $n\geq k+1$ for L_n^D to be defined and $n\geq k+s$ for L_n^H to be defined.

The main objective of this paper is to present distribution-free bounds for $P\{|\hat{L}_n-L_n|\geq\epsilon\}$ when the rule used is a k-local rule and $\hat{L}_n$ is one of the above estimates. The first distribution-free bound for k-local rules was found by Rogers and Wagner [3], who showed that

$$E(L_n^D-L_n)^2\leq((0.25+6k)/n)+(4k/n^2)$$

(see also [4], where the bound $(1+6k)/n$ is obtained) so that

$$P\{|L_n^D-L_n|\geq\epsilon\}\leq((0.25+6k)/n\epsilon^2)+(4k/n^2\epsilon^2)$$

by Chebyshev's inequality. The bounds derived in this paper, by contrast, will be exponential in n with an exponent depending on d, k, and ϵ. Similar bounds will be derived for L_n^H. The resubstitution estimate has been shown to have exponential distribution-free bounds for linear discrimination rules [5]–[7], which are not local, and has been discounted for local rules because, when the distribution of X given θ is absolutely continuous, $L_n^R=0$ with probability one for the single-nearest rule regardless of the value of L_n. In spite of this, we show that L_n^R is close to L_n^D and to L_n for large k.

II. Results

The bounds below use the constant γ_d, the maximum number of distinct points in $\mathbb{R}^d$ which can share the same nearest neighbor. While one can easily see that $\gamma_1=2$ and $\gamma_2=6$, no explicit formula for γ_d is known. It can be shown that $\gamma_d\leq 3^d-1$ for all d while other upper and lower bounds for $d\geq 9$ are given in [8].

To economize on the use of parentheses in the formulas that follow, $(abc)/(def)$ will be written abc/def (e.g., in an expression involving only multiplications and one division, the multiplications are done first).

Theorem 1: For k-local rules with $k\leq n-1$,

$$P\{|L_n^D-L_n|\geq\epsilon\}\leq 2e^{-n\epsilon^2/18}+6e^{-n\epsilon^3/108k(2+\gamma_d)}. \tag{2}$$

Theorem 2: For k-local rules with $k\leq n-s$,

$$E(L_n-L_n^H)^2\leq(1/2s)+(2sk/n) \tag{3}$$

and

$$P\{|L_n-L_n^H|\geq\epsilon\}\leq 2e^{-s\epsilon^2/2}+(2sk/n\epsilon). \tag{4}$$

By using an argument similar to the one used for Theorem 1, an exponential bound for L_n^H can be obtained which depends on d.

Theorem 3: For k-local rules with $k\leq n-s$ and $s<n\epsilon/12k(\gamma_d+2)$,

$$P\{|L_n^H-L_n|\geq\epsilon\}\leq 2e^{-2s\epsilon^2/9}+4e^{-n\epsilon^3/216k(\gamma_d+2)}. \tag{5}$$

The holdout estimate always poses the problem of the selection of s. From Theorem 2 one might be tempted to conclude that $s=\sqrt{n/4k}$ would be a good choice since it minimizes the bound (3) for $E(L_n^H-L_n)^2$. However, such a choice will not yield an exponential bound in (5). If one lets $s=\rho n/k$, for some $0<\rho<\epsilon/12(\gamma_d+2)<1$, then the bound (5) is exponential in n. This is somewhat surprising since $E(L_n^H-L_n)^2$ can go to zero at an algebraically slow rate in this case (see [3]). One might still wonder, however, if the dependence on d is necessary. The following example shows that it is.

Example: Put $M=2$, and consider the nearest neighbor rule with n fixed and $d=2n$. In $\mathbb{R}^{2n}$ let the distribution of (X,θ) put weight $1/n$ at $((0,\cdots,0),1)$ and weight $(1/2n)\cdot(1-1/n)$ at each of the points $(e_i,2)$, $1\leq i\leq 2n$, where e_i is the ith unit vector in $\mathbb{R}^{2n}$. If A is the event that exactly one of the (X_i,θ_i) equals $((0,\cdots,0),1)$ for $1\leq i\leq s$ and

none of the (X_i,θ_i) do for $s+1\leqslant i\leqslant n$, then, on A,

$$L_n^H=\frac{1}{s}$$

and

$$L_n\geqslant\frac{1}{2n}\left(1-\frac{1}{n}\right)(1+n)=\frac{1}{2}-\frac{1}{2n^2}$$

so that

$$L_n-L_n^H\geqslant\frac{1}{2}-\frac{1}{2n^2}-\frac{1}{s}.$$

Thus, whenever $1/2-1/2n^2-1/s\geqslant\epsilon$,

$$P\{|L_n-L_n^H|\geqslant\epsilon\}\geqslant P\{A\}=\frac{s}{n}\left(1-\frac{1}{n}\right)^{n-1}\geqslant\frac{s}{en}\quad(6)$$

since $\exp((x-1)/x)\leqslant x$, $0\leqslant x\leqslant 1$, implies $(1-1/n)^{n-1}\geqslant e^{-1}$. Thus if one picks $s=\rho n$ for some $\rho=\rho(\epsilon)$, where $0<\rho<1$, then (6) shows that $P\{|L_n-L_n^H|\geqslant\epsilon\}$ cannot go to zero exponentially fast in n uniformly in d and the distribution of (X,θ). In fact, when $s=\rho n$, (6) shows that it cannot even go to zero in n uniformly in d and the distribution of (X,θ).

We do not know if the dependence on d in (2) is necessary for the deleted estimate to have a distribution-free exponential bound.

If one considers the specific case of the k-nearest neighbor rule, then the above bounds can be improved somewhat by replacing k/n with $\sqrt{k}\,/n$. In both Theorem 4 and Theorem 5 below, one should be cautioned that k is fixed and not a function of n.

Theorem 4: For the k-nearest neighbor rule with $M=2$ and $k\leqslant n-1$,

$$E\left(L_n^D-L_n\right)^2\leqslant(1/n)+\left(24\sqrt{k}\,/\sqrt{2\pi}\,n\right).\quad(7)$$

If $M=2$ and $k\leqslant n-s$, then

$$E\left(L_n^H-L_n\right)^2\leqslant(1/2s)+\left(8s\sqrt{k}\,/\sqrt{2\pi}\,n\right)\quad(8)$$

and

$$P\{|L_n^H-L_n|\geqslant\epsilon\}\leqslant 2e^{-s\epsilon^2/2}+\left(8s\sqrt{k}\,/\sqrt{2\pi}\,n\epsilon\right).\quad(9)$$

In fact, by using a large k, the resubstitution estimate becomes a reasonable estimate of L_n.

Theorem 5: For the k-nearest neighbor rule with $M=2$ and $k\leqslant n-1$,

$$E\left(L_n^R-L_n\right)^2\leqslant 2E\left(L_n^D-L_n\right)^2+8/\sqrt{2\pi k}\,.\quad(10)$$

III. Proofs

If $x_1,\cdots,x_n$ is a sequence of distinct points in $\mathbb{R}^d$ and if the nearest point or neighbor to x_j is found from $x_1,\cdots,x_{j-1},x_{j+1},\cdots,x_n$ for $1\leqslant j\leqslant n$, then as noted earlier, no point can be the nearest neighbor to more than γ_d of the remaining points. If the points are now the values assumed by the observations $X_1,\cdots,X_n$, we can no longer make this statement because the distribution of X may have atoms so that $X_1,\cdots,X_n$ are not necessarily distinct. Nevertheless, if we use the sequence $Z_1,\cdots,Z_n$ and the notion of "nearest" in Section I, we have the following easy lemma.

Lemma 1: Suppose $(X_1,Z_1),\cdots,(X_n,Z_n)$ is the sequence obtained from the data by omitting the states of each observation. If, for each j, the nearest neighbor to (X_j,Z_j) is found from $(X_1,Z_1),\cdots,(X_{j-1},Z_{j-1})$, $(X_{j+1},Z_{j+1}),\cdots,(X_n,Z_n)$, then no point (X_i,Z_i) can be the nearest neighbor to more than γ_d+2 of the remaining points.

Lemma 2: Suppose the probability distribution of the binary-valued sequence $Y_1,\cdots,Y_n$ is the same as that of $Y_{\sigma(1)},\cdots,Y_{\sigma(n)}$ for every permutation σ of $1,\cdots,n$. Then

$$P\left\{\left|\frac{1}{l}\sum_1^l Y_i-\frac{1}{n}\sum_1^n Y_i\right|\geqslant\epsilon\right\}\leqslant 2e^{-2l\epsilon^2},\qquad 1\leqslant l\leqslant n.$$

Proof: If $Q(y_1,\cdots,y_n)$ is the probability distribution of $Y_1,\cdots,Y_n$, then

$$Q(y_{\sigma(1)},\cdots,y_{\sigma(n)})=Q(y_1,\cdots,y_n)$$

for all σ, and

$$\begin{aligned}
&P\left\{\left|\frac{1}{l}\sum_1^l Y_i-\frac{1}{n}\sum_1^n Y_i\right|\geqslant\epsilon\right\}\\
&=\frac{1}{n!}\sum_\sigma P\left\{\left|\frac{1}{l}\sum_1^l Y_{\sigma(i)}-\frac{1}{n}\sum_1^n Y_i\right|\geqslant\epsilon\right\}\\
&=\frac{1}{n!}\sum_\sigma\sum_{y_1,\cdots,y_n} I_{\{|(1/l)\Sigma_1^l y_{\sigma(i)}-(1/n)\Sigma_1^n y_i|\geqslant\epsilon\}}\\
&\quad\cdot Q(y_1,\cdots,y_n)\\
&=\sum_{y_1,\cdots,y_n} Q(y_1,\cdots,y_n)\\
&\quad\cdot\left(\frac{1}{n!}\sum_\sigma I_{\{|(1/l)\Sigma_1^l y_{\sigma(i)}-(1/n)\Sigma_1^n y_i|\geqslant\epsilon\}}\right)\\
&\leqslant\sum_{y_1,\cdots,y_n} Q(y_1,\cdots,y_n)2e^{-2l\epsilon^2}=2e^{-2l\epsilon^2}
\end{aligned}$$

where the inequality follows from Hoeffding [2, sec. 6, theorem 4].

Proof of Theorem 1: We consider one-local rules first. Suppose

$$(X,\theta,Z),(X_1,\theta_1,Z_1),\cdots,(X_{n+m},\theta_{n+m},Z_{n+m})$$

are i.i.d. with

$$V_n=(X_1,\theta_1,Z_1),\cdots,(X_n,\theta_n,Z_n)$$

$$U_n=(X_{n+1},\theta_{n+1},Z_{n+1}),\cdots,(X_{n+m},\theta_{n+m},Z_{n+m})$$

$$T_n=(X_1,\theta_1,Z_1),\cdots,(X_{n+m},\theta_{n+m},Z_{n+m}).$$

We shall write $V_{n,i}$, $U_{n,i}$, $T_{n,i}$ to denote the corresponding sequences with (X_i,θ_i,Z_i) deleted. As before, V_n denotes the data used with X,Z to estimate θ. The sequence U_n and its concatenation with V_n, denoted by T_n, are used

only in the proof. Let

$$L_n = P\{\hat{\theta}(V_n, X, Z) \neq \theta | V_n\}$$

$$L_{n1} = \frac{1}{m}\sum_{i=1}^{m} I_{\{\hat{\theta}(V_n, X_{n+i}, Z_{n+i}) \neq \theta_{n+i}\}}$$

$$L_{n2} = \frac{1}{m}\sum_{i=1}^{m} I_{\{\hat{\theta}(T_{n,n+i}, X_{n+i}, Z_{n+i}) \neq \theta_{n+i}\}}$$

$$L_{n3} = \frac{1}{n+m}\sum_{i=1}^{n+m} I_{\{\hat{\theta}(T_{n,i}, X_i, Z_i) \neq \theta_i\}}$$

$$L_{n4} = \frac{1}{n}\sum_{i=1}^{n} I_{\{\hat{\theta}(T_{n,i}, X_i, Z_i) \neq \theta_i\}}$$

$$L_n^D = \frac{1}{n}\sum_{i=1}^{n} I_{\{\hat{\theta}(V_{n,i}, X_i, Z_i) \neq \theta_i\}}.$$

Our proof consists of noting that

$$\begin{aligned} P\{|L_n - L_n^D| \geq \epsilon\} \leq\ & P\{|L_n - L_{n1}| \geq \epsilon/6\} \\ & + P\{|L_{n1} - L_{n2}| \geq 2\epsilon/6\} \\ & + P\{|L_{n2} - L_{n3}| \geq \epsilon/6\} \\ & + P\{|L_{n3} - L_{n4}| \geq \epsilon/6\} \\ & + P\{|L_{n4} - L_n^D| \geq \epsilon/6\} \end{aligned} \tag{11}$$

and showing that each term can be bounded in a distribution-free way by picking m properly.

1) Using Hoeffding's inequality for sums of independent [0, 1]-valued random variables [2], we have

$$P\{|L_n - L_{n1}| \geq \epsilon\} \leq 2e^{-2m\epsilon^2}.$$

2) We have

$$\begin{aligned} & P\{|L_{n1} - L_{n2}| \geq \epsilon\} \\ & \leq P\left\{\frac{1}{m}\sum_{i=1}^{m} I_{\left[\hat{\theta}(V_n, X_{n+i}, Z_{n+i}) \neq \hat{\theta}(T_{n,n+i}, X_{n+i}, Z_{n+i})\right]} \geq \epsilon\right\} \\ & \leq P\left\{\frac{1}{m}\sum_{i=1}^{m} I_{A(n+i)} \geq \epsilon\right\} \end{aligned}$$

where, for $l = 1, \cdots, n+m$, $A(l)$ is the event that (X_{n+j}, Z_{n+j}) is the nearest neighbor to (X_l, Z_l) from $T_{n,l}$ for some $j = 1, \cdots, m$, excluding $j = l - n$ if $l > n$. To use Lemma 2, we symmetrize the sequence $I_{A(1)}, \cdots, I_{A(n+m)}$ as follows. Let $Y_1, \cdots, Y_{n+m}$ be a random permutation of $1, \cdots, n+m$. Then $I_{A(Y_1)}, \cdots, I_{A(Y_{n+m})}$ satisfies the conditions of Lemma 2, and

$$\begin{aligned} & P\left\{\frac{1}{m}\sum_{i=1}^{m} I_{A(n+i)} \geq \epsilon\right\} \\ & \leq P\left\{\frac{1}{m}\sum_{i=1}^{m} I_{A(Y_i)} \geq \epsilon\right\} \\ & = P\left\{\frac{1}{m}\sum_{1}^{m} I_{A(Y_i)} - \frac{1}{n+m}\sum_{1}^{n+m} I_{A(i)} \geq \epsilon - \frac{1}{n+m}\sum_{1}^{n+m} I_{A(i)}\right\} \\ & \leq 2e^{-2m(\epsilon/2)^2} \end{aligned}$$

whenever

$$\frac{1}{n+m}\sum_{1}^{n+m} I_{A(i)} \leq \epsilon/2. \tag{12}$$

However, since each (X_{n+j}, Z_{n+j}) can be the nearest neighbor of at most $(\gamma_d + 2)$ other points from T_n, we see that $\sum_1^{n+m} I_{A(i)} \leq (\gamma_d + 2)m$, and thus (12) is valid whenever $(\gamma_d + 2)m < (n+m)\epsilon/2$.

3) Lemma 2 can be applied immediately to $L_{n2} - L_{n3}$ and $L_{n3} - L_{n4}$ to yield

$$P\{|L_{n2} - L_{n3}| \geq \epsilon\} \leq 2e^{-2m\epsilon^2}$$

$$P\{|L_{n3} - L_{n4}| \geq \epsilon\} \leq 2e^{-2n\epsilon^2}.$$

4) Finally,

$$\begin{aligned} & P\{|L_{n4} - L_n^D| \geq \epsilon\} \\ & \leq P\left\{\frac{1}{n}\sum_{1}^{n} I_{\left[\hat{\theta}(T_{n,i}, X_i, Z_i) \neq \hat{\theta}(V_{n,i}, X_i, Z_i)\right]} \geq \epsilon\right\} \\ & \leq P\left\{\frac{1}{n}\sum_{i=1}^{n} I_{A(i)} \geq \epsilon\right\} \\ & = 0, \qquad \text{if } m(\gamma_d + 2) < n\epsilon. \end{aligned}$$

Taking $m = n\epsilon/6(\gamma_d + 2)$ and using inequality (11), we see that

$$P\{|L_n - L_n^D| \geq \epsilon\} \leq 2e^{-n\epsilon^2/18} + 6e^{-me^2/18}$$

which yields (2). For an arbitrary k, it suffices to replace $(\gamma_d + 2)$ by $k\gamma_d + 2 \leq k(\gamma_d + 2)$ since no (X_i, Z_i) can be one of the k nearest neighbors to more than $k\gamma_d + 2$ of the points in $V_{n,i}$.

Proof of Theorem 2: Letting

$$V_n' = (X_1, \theta_1, Z_1), \cdots, (X_s, \theta_s, Z_s)$$

and

$$V_n'' = (X_{s+1}, \theta_{s+1}, Z_{s+1}), \cdots, (X_n, \theta_n, Z_n),$$

we have

$$\begin{aligned} |L_n - L_n^H| & \leq |L_n - E(L_n^H | V_n'')| + |E(L_n^H | V_n'') - L_n^H| \\ & \leq P\{\hat{\theta}(V_n, X, Z) \neq \hat{\theta}(V_n'', X, Z) | V_n\} \\ & \quad + |E(L_n^H | V_n'') - L_n^H| \end{aligned}$$

since

$$\begin{aligned} E(L_n^H | V_n'') & = P\{\hat{\theta}(V_n'', X, Z) \neq \theta | V_n''\} \\ & = P\{\hat{\theta}(V_n'', X, Z) \neq \theta | V_n\} \end{aligned}$$

and

$$\begin{aligned} & |P\{\hat{\theta}(V_n, X, Z) \neq \theta | V_n\} - P\{\hat{\theta}(V_n'', X, Z) \neq \theta | V_n\}| \\ & \qquad \leq P\{\hat{\theta}(V_n, X, Z) \neq \hat{\theta}(V_n'', X, Z) | V_n\}. \end{aligned}$$

Using $|a+b|^r \leq 2^{r-1}(|a|^r+|b|^r)$ for $r \geq 1$, we have

$$\begin{aligned} E(L_n - L_n^H)^2 &\leq 2P\{\hat{\theta}(V_n,X,Z) \neq \hat{\theta}(V_n'',X,Z)\} \\ &\quad + 2E(L_n^H - E(L_n^H|V_n''))^2 \\ &\leq 2P\left\{ \bigcup_{i=1}^{s} \{(X_i,Z_i) \text{ is among the } k \right. \\ &\qquad \left. \text{nearest to } (X,Z)\} \right\} \\ &\quad + 2\frac{1}{4s} \\ &\leq \frac{2sk}{n} + \frac{1}{2s} \end{aligned} \tag{13}$$

which proves (3). Also,

$$\begin{aligned} &P\{|L_n - L_n^H| \geq \epsilon\} \\ &\leq P\{|L_n^H - E(L_n^H|V_n'')| \geq \epsilon/2\} \\ &\quad + P\{P\{\hat{\theta}(V_n,X,Z) \neq \hat{\theta}(V_n'',X,Z)|V_n\} \geq \epsilon/2\} \\ &\leq E\{P\{|L_n^H - E(L_n^H|V_n'')| \geq \epsilon/2|V_n''\}\} \\ &\quad + \frac{2}{\epsilon}P\{\hat{\theta}(V_n,X,Z) \neq \hat{\theta}(V_n'',X,Z)\} \\ &\leq 2e^{-2s(\epsilon/2)^2} + 2sk/n\epsilon \end{aligned}$$

using Hoeffding's inequality and Markov's inequality. This proves Theorem 2.

Proof of Theorem 3: Using the notation of the proofs of Theorems 1 and 2, consider k equal to one. From the argument of Theorem 2 we have

$$\begin{aligned} P\{|L_n - L_n^H| \geq \epsilon\} &\leq 2e^{-2s\epsilon^2/9} \\ &\quad + P\{P\{\hat{\theta}(V_n,X,Z) \neq \hat{\theta}(V_n'',X,Z)|V_n\} \geq 2\epsilon/3\}. \end{aligned}$$

Letting

$$L_n^* = \frac{1}{m}\sum_1^m I_{[\hat{\theta}(V_n,X_{n+i},Z_{n+i}) \neq \hat{\theta}(V_n'',X_{n+i},Z_{n+i})]}$$

we have

$$\begin{aligned} &P\{P\{\hat{\theta}(V_n,X,Z) \neq \hat{\theta}(V_n'',X,Z)|V_n\} \geq 2\epsilon/3\} \\ &\leq P\{|L_n^* - P\{\hat{\theta}(V_n,X,Z) \neq \hat{\theta}(V_n'',X,Z)|V_n\}| \geq \epsilon/3\} \\ &\quad + P\{L_n^* \geq \epsilon/3\} \\ &\leq 2e^{-2m\epsilon^2/9} + P\{L_n^* \geq \epsilon/3\} \end{aligned}$$

where, for the first term, we first condition the probability on V_n and use Hoeffding's inequality. Also,

$$P\{L_n^* \geq \epsilon/3\} \leq P\left\{\frac{1}{m}\sum_{i=1}^m I_{C(n+i)} \geq \epsilon/3\right\}$$

where $C(n+j)$ is the event that the closest point to (X_{n+j},Z_{n+j}) from V_n is (X_i,Z_i) for some $1 \leq i \leq s$. However, if $D(j)$ is the event that the closest point to (X_j,Z_j) from $T_{n,j}$ is (X_i,Z_i) for some $1 \leq i \leq s$ or $n < i \leq n+m$, then

$$C(n+j) \subseteq D(n+j)$$

and

$$\begin{aligned} P\{L_n^* \geq \epsilon/3\} &\leq P\left\{\frac{1}{m}\sum_{i=1}^m I_{D(n+i)} \geq \epsilon/3\right\} \\ &\leq P\left\{\frac{1}{m}\sum_{i=1}^m I_{D(Y_i)} \geq \epsilon/3\right\} \end{aligned}$$

where $Y_1,\cdots,Y_{n+m}$ is an independent random permutation of $1,\cdots,n+m$. Using the same arguments as Theorem 1, we see that this last term is bounded by $2e^{-2m(\epsilon/6)^2}$ if $1/(n+m)\sum_1^{n+m} I_{D(i)} \leq \epsilon/6$. This occurs if $(\gamma_d+2)(s+m) < (\epsilon/6)(n+m)$. Taking $m = n\epsilon/12(\gamma_d+2)$ and $s < n\epsilon/12(\gamma_d+2)$ yields the theorem for $k=1$ after collecting terms. For arbitrary k, we need only replace (γ_d+2) by $k(\gamma_d+2)$.

Lemma 3: If $P\{Y=j\} = \binom{n}{j}(1/2)^n$, $0 \leq j \leq n$, then

$$P\{|Y-n/2| \leq a/2\} < 4a/\sqrt{2\pi n}$$

for all positive integers a.

Proof: We make repeated use of Feller's [9] approximation for $n!$. If n is even, the maximal term of the binomial expansion is

$$\frac{1}{2^n}\binom{n}{n/2} \leq \frac{2}{\sqrt{2\pi n}} \exp\left(\frac{1}{12n} - \frac{2}{6n+1}\right) < \frac{2}{\sqrt{2\pi n}}.$$

Hence $P\{|Y-n/2| \leq a/2\}$ is upper-bounded by

$$(a+1)2/\sqrt{2\pi n} < 4a/\sqrt{2\pi n}.$$

For n odd, $n \geq 3$, the maximal term is

$$\begin{aligned} \frac{1}{2^n}\binom{n}{(n-1)/2} &< \frac{1}{2^{n-1}}\binom{n-1}{(n-1)/2} \\ &< \frac{1}{\sqrt{2\pi(n-1)}} < \frac{2}{\sqrt{\pi n}}. \end{aligned}$$

Thus

$$P\{|Y-n/2| \leq a/2\} \leq 2a/\sqrt{\pi n} < 4a/\sqrt{2\pi n}.$$

Lemma 4: For k-local estimates with $k \leq n-1$,

$$\begin{aligned} E(L_n - L_n^D)^2 &\leq 1/n + 6P\{\hat{\theta}(V_n,X,Z) \neq \hat{\theta}(V_{n,1},X,Z)\} \\ &\leq 1/n + 6k/n. \end{aligned}$$

Lemma 4 is proved in [4]. It can, in implicit form, also be found in [3].

Proof of Theorem 4: We will show that

$$P\{\hat{\theta}(V_n,X,Z) \neq \hat{\theta}(V_n'',X,Z)\} \leq 4s\sqrt{k}/\sqrt{2\pi}\, n. \tag{14}$$

A combination of (14) and (13) yields (8); (7) follows from Lemma 4 and (14) upon noting that $V_n'' = V_{n,1}$ if $s=1$.

Let N, N_1, and N_2 be the number of (X_i,Z_i) that are among the k nearest neighbors to (X,Z) and for which, respectively, $1 \leq i \leq s$, $\theta_i = 1$, and $\theta_i = 2$. Conditioned on (X^{k+1},Z^{k+1},X,Z), the random variables N and $|N_1 - N_2|$

are independent. Thus

$$\begin{aligned}
&P\{\hat{\theta}(V_n,X,Z)\neq\hat{\theta}(V_n'',X,Z)\}\\
&\quad\leqslant\sum_{j=1}^{s}P\{N=j,\,|N_1-N_2|\leqslant j\}\\
&\quad=E\left\{\sum_{j=1}^{s}P\{N=j|X,Z\}\right.\\
&\qquad\left.\cdot P\{|N_1-N_2|\leqslant j|X,Z,X^{k+1},Z^{k+1}\}\right\}.
\end{aligned}$$

If $\binom{k}{j}=0$ for $j>k$,

$$P\{N=j|X,Z\}=\frac{\binom{k}{j}\binom{n-k}{s-j}}{\binom{n}{s}}.$$

Also, by Lemma 3,

$$\begin{aligned}
&P\{|N_1-N_2|\leqslant j|X,Z,X^{k+1},Z^{k+1}\}\\
&\quad=P\{|N_1-k/2|\leqslant j/2|X,Z,X^{k+1},Z^{k+1}\}\\
&\quad\leqslant 4j/\sqrt{2\pi k}\,.
\end{aligned}$$

Collecting bounds and using a property of the hypergeometric distribution [10] yields

$$\begin{aligned}
P\{\hat{\theta}(V_n,X,Z)\neq\hat{\theta}(V_n'',X,Z)\}&\leqslant\sum_{j=1}^{s}j\frac{\binom{k}{j}\binom{n-k}{s-j}}{\binom{n}{s}}\frac{4}{\sqrt{2\pi k}}\\
&=4ks/n\sqrt{2\pi k}\\
&=4s\sqrt{k}\,/\sqrt{2\pi}\,n.
\end{aligned}$$

Proof of Theorem 5: From $(a+b)^2\leqslant 2(a^2+b^2)$ and

$$|L_n-L_n^R|\leqslant|L_n-L_n^D|+\left|\frac{1}{n}\sum_{i=1}^{n}I_{\{\hat{\theta}(V_n,X_i,Z_i)\neq\hat{\theta}(V_{n,i},X_i,Z_i)\}}\right|,$$

we deduce

$$\begin{aligned}
E(L_n-L_n^R)^2&\leqslant 2E(L_n-L_n^D)^2\\
&\quad+2E\left(\frac{1}{n}\sum_{i=1}^{n}I_{\{\hat{\theta}(V_n,X_i,Z_i)\neq\hat{\theta}(V_{n,i},X_i,Z_i)\}}\right)^2\\
&\leqslant 2E(L_n-L_n^D)^2\\
&\quad+2P\{\hat{\theta}(V_n,X_1,Z_1)\neq\hat{\theta}(V_{n,1},X_1,Z_1)\}\\
&\leqslant 2E(L_n-L_n^D)^2\\
&\quad+2\sup_{x,z}P\{|N_1(x,z)-N_2(x,z)|\leqslant 1\}\\
&\leqslant 2E(L_n-L_n^D)^2+8/\sqrt{2\pi k}
\end{aligned}$$

where we use Lemma 3 and where $N_1(x,z)$ and $N_2(x,z)$ are as in the proof of Theorem 4 after replacement of (X,Z) by (x,z).

References

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

[2] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. Amer. Stat. Ass.*, vol. 58, pp. 13–30, 1963.

[3] W. H. Rogers and T. J. Wagner, "A finite sample distribution-free performance bound for local discrimination rules," *Ann. Stat.*, vol. 6, pp. 506–514, 1978.

[4] L. P. Devroye and T. J. Wagner, "Nonparametric discrimination and density estimation," Information Systems Research Laboratory, Univ. Texas, Austin, Tech. Rep. 183, 1976.

[5] V. N. Vapnik and A. Ya. Chervonenkis, "Theory of uniform convergence of frequencies of events to their probabilities and problems of search for an optimal solution from empirical data," *Automation and Remote Control*, vol. 32, pp. 207–217, 1971.

[6] L. P. Devroye and T. J. Wagner, "A distribution-free performance bound in error estimation," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 586–587, Sept. 1976.

[7] ——, "Distribution-free performance bounds with the resubstitution error estimate," in *Proc. 1977 Computer Society Conf. Pattern Recognition and Image Processing*, Troy, NY, 1977, pp. 323–326.

[8] C. Rogers, "Covering a sphere with spheres," *Mathematika*, vol. 10, pp. 157–164, 1963.

[9] W. Feller, *An Introduction to Probability Theory and Its Applications*, vol. 1. New York: Wiley, 1968.

[10] G. Roussas, *A First Course in Mathematical Statistics*. Reading, MA: Addison-Wesley, 1973.