The Annals of Statistics
2001, Vol. 29, No. 5, 1424–1442

NEAREST NEIGHBOR CLASSIFICATION WITH DEPENDENT TRAINING SEQUENCES

BY M. HOLST AND A. IRLE

Christian-Albrechts-Universität Kiel

The asymptotic classification risk for nearest neighbor procedures is well understood in the case of i.i.d. training sequences. In this article, we generalize these results to a class of dependent models including hidden Markov models. In the case where the observed patterns have Lebesgue densities, the asymptotic risk takes the same expression as in the i.i.d. case. For discrete distributions, we show that the asymptotic risk depends on the rule used for breaking ties of equal distances.

1. Introduction and model. Pattern recognition considers the following basic situation: a random variable (X, Y) consists of an observed pattern $X \in \mathbb{R}^d$ from which we wish to infer the unobservable class Y. We assume that this class belongs to some known finite set M, fixed as $M = \{1, \dots, m\}$. If the joint distribution of (X, Y) is known, then we may simply choose the class having maximum a posteriori probability, given the observed pattern. The resulting probability of misclassification is usually called the Bayes risk.

In general the joint distribution of (X, Y) will be unknown, and we have a

$$\text{training sequence } Z_n = ((X_1, Y_1), \dots, (X_n, Y_n))$$

at our disposal, where patterns and corresponding classes are observed. Here the random variables $(X_1, Y_1), \dots, (X_n, Y_n)$ are taken to be jointly stochastically independent of (X, Y), while having the same distribution, at least asymptotically, as (X, Y).

1.1. *Nearest neighbor classification.* A well-known classification procedure is the k-NN procedure where NN stands for nearest neighbor.

Having observed $x \in \mathbb{R}^d$ we order $\|x - X_i\|$ according to increasing values with respect to the Euclidean norm for $\mathbb{R}^d$ so that

$$\|x - X_{R_{1:n}(x)}\| \le \|x - X_{R_{2:n}(x)}\| \le \cdots \le \|x - X_{R_{n:n}(x)}\|,$$

where $(R_{1:n}(x), \dots, R_{n:n}(x))$ is a random permutation of $(1, \dots, n)$. The event that different patterns from our training sequence have identical distance from x has probability 0 when we have Lebesgue densities. For discrete distributions, the problem of equal distances has to be treated with some care; see Section 3.

Received November 1998; revised October 2000.

AMS 2000 *subject classifications*. Primary 62H30; secondary 62G20.

Key words and phrases. Nearest neighbor classification, asymptotic risk, dependent training samples.

The k-NN procedure chooses that class which occurs most often among $Y_{R_{1:n}(x)}, \dots, Y_{R_{k:n}(x)}$. If this class is not unique, some tie-breaking rule has to be applied.

Let us denote for $k \in \mathbb{N}$, $j \in M = \{1, \dots, m\}$ the number of occurences of class j among the k-nearest neighbors by

$$N_{k,n}^{(j)}(x, Z_n) = |\{i \in \{1, \dots, k\} \mid Y_{R_{i:n}(x)} = j\}|.$$

We assume that for any $J \subset M$, $J \neq \emptyset$, we have stochastically independent random variables T_J with values in J which are also stochastically independent of (X, Y) and Z_n and give the tie-breaking rule. Then the k-NN procedure is formally given as

$$\delta_{k,n}(x, Z_n) = T_J \text{ with } J = \Big\{ l \in M \mid N_{k,n}^{(l)}(x, Z_n) = \max_{j \in M} N_{k,n}^{(j)}(x, Z_n) \Big\}.$$

In the following, we shall usually omit the dependence on x and Z_n in our notations by writing, for example,

$$R_{i:n}, N_{k,n}^{(j)}, \delta_{k,n}.$$

Since the introduction of k-NN procedures by Fix and Hodges (1951, 1952), substantial research on their theoretical properties and their practical performance has been carried out. This is documented in the tutorial by Dasarathy (1991), which also contains reprints of various key papers in this area, and in the monograph by Devroye, Györfi and Lugosi (1996), a considerable part of which is devoted to nearest neighbor rules. On the theoretical side, two problems in particular have been investigated.

1. *Evaluation of the asymptotic risk, that is, the asymptotic probability of misclassification, for fixed k as n* $\to \infty$. In the i.i.d. situation, where (X, Y), $(X_1, Y_1), (X_2, Y_2) \cdots$ form an i.i.d. sequence, the asymptotic risk was derived by Cover and Hart (1967) under certain continuity assumptions. The extension to the general nonparametric case follows from Stone (1977), and was stated explicitly by Devroye (1981a, b).

Although this is the quantity to be studied in our paper, let us also point out a second major line of research.

2. *Consistency of k-NN procedures as* $k = k(n) \to \infty$ *and* $n \to \infty$. Here a classification procedure is called consistent if the asymptotic risk is equal to the Bayes risk. In the i.i.d. case, nonparametric consistency was established by Stone (1977) as $k(n)/n \to \infty$, and questions of this type were later investigated by various authors [see, e.g., the discussion in Devroye, Györfi and Lugosi (1996), Chapters 5 and 11, and Devroye, Györfi, Krzyzak and Lugosi (1994)]. We remark that consistency for discrimination problems follows from consistency in the corresponding regression problems; hence any result on regression consistency for k-NN procedures yields a result on discrimination consistency.

Let us finally mention a new line of research where the risk of k-NN procedures is expanded for finite sample size n; see Snapp and Venkatesch (1998).

1.2. *The asymptotic risk in the i.i.d. case.* To state the result of Cover and Hart (1967) in the nonparametric situation considered by Stone (1977), Devroye (1981a, b), we introduce several notations which will also be used throughout this paper.

The underlying probability measure will be denoted by P and the distribution of a random variable V by P^V, so that $P^{(X,Y)}$ denotes the joint distribution of (X, Y) with respect to the underlying probability measure.

Let $p_y(x) = P(Y = y \mid X = x)$ for $x \in \mathbb{R}^d$, $y \in M$.

For $J \subset M$, $J = \{y_1, \dots, y_j\}$, $y_1 < \cdots < y_j$, let

$$A_J = \bigcup_{i=1}^{k} \{0, \dots, i-1\}^{y_1 - 1} \times \{i\} \times \{0, \dots, i-1\}^{y_2 - y_1 - 1} \times \{i\} \\ \times \cdots \times \{i\} \times \{0, \dots, i-1\}^{m - y_j}.$$

Then the event, that the classes from J occur most often among $Y_{R_{1:n}}, \dots, Y_{R_{k:n}}$, is given by

$$\left(\sum_{i=1}^{k} 1_{\{Y_{R_{i:n}} = 1\}}, \dots, \sum_{i=1}^{k} 1_{\{Y_{R_{i:n}} = m\}} \right) \in A_J.$$

$M(k, p_1, \dots, p_m)$ denotes the multinomial distribution with parameters k, $p_1, \dots, p_m$, so that

$$M(k, p_1, \dots, p_m)(\{(j_1, \dots, j_m)\}) = \frac{k!}{j_1! \cdot \cdots \cdot j_m!} \prod_{i=1}^{m} p_i^{j_i},$$

with $j_1, \dots, j_m \in \{0, \dots, k\}$, $\sum_{i=1}^{m} j_i = k$.

We now formally introduce the quantity whose asymptotic behaviour is investigated in this article.

DEFINITION 1.1. The risk of the k-NN procedure from a training sequence of size n is defined as

$$R(\delta_{k,n}) = P(\delta_{k,n}(X, Z_n) \neq Y).$$

Then in the i.i.d. situation, the asymptotic risk is given by the following result.

THEOREM 1.1. *Consider the i.i.d. situation. Then*

$$\lim_{n \to \infty} R(\delta_{k,n}) = \sum_{y \in M} \int_{\mathbb{R}^d} \sum_{\substack{J \subset M \\ y \in J}} M(k, p_1(x), \dots, p_m(x))(A_J) P(T_J = y) \\ \times (1 - p_y(x)) P^X(dx).$$

In the case of only two classes and using the smaller class in the case of ties, this takes the following simpler form.

THEOREM 1.2. *Consider the i.i.d. situation. Let* $M = \{1, 2\}$ *and* $T_J = \min J$ *for* $J \subset M$. *Then with* $p(x) = p_1(x)$

$$\lim_{n\to\infty} R(\delta_{k,n}) = \int p(x) \sum_{j<k/2} \binom{k}{j} p(x)^j (1-p(x))^{k-j} + (1-p(x)) \sum_{j\geq k/2} \binom{k}{j} p(x)^j (1-p(x))^{k-j} P^X(dx).$$

Although the general expression for the asymptotic risk as given in Theorem 1.1 appears to be well known in the field of pattern recognition, see, for example Snapp and Venkatesh (1998), Formula (9), it does not appear to be stated explicitly in this generality. Cover and Hart (1967) formulate it for $k = 1$, general m, and $m = 2$, general k. [Devroye (1981a, b), see also Devroye, Györfi and Lugosi (1996). Chapter 5, treats the latter case.]

Validity of the general expression may easily be derived by the following argument. Asymptotically, $Y_{R_{1:n}}, \dots, Y_{R_{k:n}}$ behave like k i.i.d. random variables which take the values $1, \dots, m$ with probabilities $p_1(x), \dots, p_m(x)$. Using this and taking care of the definition of the k-NN procedure together with the tie-breaking rule, the result follows immediatedly.

Let us point out two approaches which have been used to obtain this result without continuity assumptions. On the one hand, a geometrical cone-covering argument together with an application of the i.i.d. setting was used by Stone (1977); on the other, the general Lebesgue differentiation theorem was used by Devroye (1981a, b). In our later arguments we shall apply the Lebesgue differentiation theorem as formulated in the proof of Theorem 2.1.

We have introduced nearest neighbor procedures with respect to the Euclidean norm. Any other norm for which the Lebesgue differentiation theorem is known to hold, for example the maximum norm, could also be used for the results of this paper. Furthermore, it is stated in Devroye, Györfi and Lugosi [(1996), Chapter 5, Problem 5.1] that Stone's cone-covering argument, hence also Theorem 1.1, remains valid for general norms.

1.3. *Dependent models.* The aim of this paper is to generalize the above results to training sequences with stochastic dependencies. It is readily conjectured that the results for i.i.d. training sequences carry over to suitably dependent training sequences. Looking at the k nearest neighbors out of a large training sequence, the indices of the k nearest neighbors will tend to be far apart, and their classes should tend to independence under reasonable assumptions. The arguments in this article provide the exact reasoning for this line of thought.

Our starting point was the investigation of hidden Markov models which have elicited strong practical and theoretical interest in recent years; see the monographs by MacDonald and Zucchini (1997) and by Huang, Ariki and Jack (1990) from the more practical viewpoint and Bickel, Ritov and Ryden (1998) from the viewpoint of asymptotic statistics. As it turned out, our results could be proved for more general models which we introduce now.

General conditionally independent model (GCIM).

(a) The distribution $\pi = P^Y$ fulfils the condition $0 < \pi(\{i\}) < 1$ for all $i \in M$.

(b) For any sequence $(A_n)_{n\in\mathbb{N}}$ of Borel subsets $A_n \in \mathscr{B}^d$ we have

$$P((X_n)_{n\in\mathbb{N}} \in (A_n)_{n\in\mathbb{N}} | (Y_n)_{n\in\mathbb{N}} = (y_n)_{n\in\mathbb{N}}) = \prod_{n\in\mathbb{N}} Q_{y_n}(A_n),$$

with $Q_y = P(X \in \cdot \mid Y = y)$, so that in particular the $(X_n)_{n\in\mathbb{N}}$ are conditionally independent given $(Y_n)_{n\in\mathbb{N}}$.

(c) For each $i \in M$,

$$\frac{S_n^{(i)}}{n} = \frac{1}{n}\sum_{j=1}^{n} 1_{\{i\}}(Y_j) \to \pi(\{i\}) \quad \text{in probability.}$$

Condition (a) just states that all classes occur with positive probability and may be assumed without loss of generality.

Condition (b) expresses that the distributions of the observed patterns depend only on their classes and not on other patterns or classes. This is an assumption already essential in working with hidden Markov models, and of course, any i.i.d. model incorporates this feature. It is a common condition in pattern recognition problems.

Condition (c) describes the dependence structure for the sequence of classes. It may be viewed as a weak ergodicity assumption. Ergodic Markov chains provide special cases so that hidden Markov models are special cases of GCI models, but of course many other classes of processes fulfil this assumption. Note that we do not assume stationarity and that no quantitative assumptions, such as those on mixing coefficients, are involved.

Recent work in machine intelligence and learning argues in favor of the use of statistical techniques, in particular k-NN procedures, due to the availability of very high-speed computing [see, e.g., Smith, Bourgoin, Sims and Voorhees (1994)]. In this paper, the use of k-NN procedures is investigated for the recognition of handwritten characters. The databases as described in this paper and similarly in Kahan, Pavlidis and Baird (1987) for the recognition of printed characters stem from written text, here in particular from written English. Then the classes, that is, the consecutive true letters, are of course no longer independent. A first approximation would be provided by a simple Markov model, but Shannon (1951) has already argued that such a model is not adequate. Information theoretical investigations point to a model with dependencies up to eight letters; see Cover and Thomas (1991). In this situation, the $Y_1, Y_2, \dots,$ would come from an ergodic, non-Markovian scource providing an example for a GCIM which is not a hidden Markov model.

Let us finally point out work in dependent models which is related to this articles. To our knowledge, the problem of obtaining the asymptotic risk of k-NN procedures for training sequences with dependencies, which is the problem studied in this article, has not been treated in the literature.

On the other hand, the problem of asymptotic regression consistency, hence discrimination consistency under conditions of dependence in the Y_i's has been treated by various authors. The monograph by Györfi, Härdle, Sarda and Vieu (1989) gives an overview of results under mixing conditions, where the main tools are exponential probabilistic inequalities. A different approach to derive consistency under mixing conditions is given in Irle (1997) where, in particular, consistency for k-NN procedures is treated in detail.

It is known that, for general ergodic models such as those treated in this article, nonparametric consistency no longer holds universally [see, e.g., Györfi and Lugosi (1992) for a regression problem, Adams and Nobel (1998) for a problem of density estimation]. A universally consistent procedure for a special estimation problem in ergodic models is given by Morvai, Yakowitz and Györfi (1996). A different approach dealing with nonstochastic sequences, and then transforming the results to a stochastic setting under certain continuity conditions, has been proposed in Kulkarni and Posner (1995) and Nobel, Morvai and Kulkarni (1998). Let us remark that the last article treats nearest neighbor procedures under the assumption that the classes Y_i's are independent given the observables X_i's, which is rather different from the common assumption, which is also used in this article of conditional independence of the observables given the classes.

2. Models with Lebesgue densities.

In this section we shall show that, for models with Lebesgue densities, we have the same behaviour of k-NN procedures as in the i.i.d. case. We start by giving two basic facts, valid in any GCIM.

LEMMA 2.1. *Consider a GCIM. Then for P^X-almost all $x \in \mathbb{R}^d$,*

$$\lim_{n\to\infty} \|x - X_{R_{k:n}}\| = 0 \ P\text{-almost surely.}$$

PROOF. Let $\varepsilon > 0$ and $x \in \text{support}(P^X)$. Then there exists $l \in M$ with $Q_l(\overline{K}(x,\varepsilon)) > 0$, where $\overline{K}(x,\varepsilon)$ denotes the closed ball with center x and radius ε.

For $S_n^{(l)}$ we have

$$\lim_{n\to\infty} P\biggl(\biggl|\frac{S_n^{(l)}}{n} - \pi(\{l\})\biggr| > \delta\biggr) = \lim_{n\to\infty} P(S_n^{(l)} \notin D_{n,\delta}) = 0,$$

where $\delta > 0$ and $D_{n,\delta} = [n(\pi(\{l\}) - \delta), n(\pi(\{l\}) + \delta)] \cap \mathbb{N}$ with $[\cdot]$ denoting integer part.

This implies

$$\begin{aligned}
&P(\|x-X_{R_{k:n}}\|>\varepsilon)\\
&\le P\Bigl(\sum_{i=1}^{n}1_{\{\|x-X_i\|\le\varepsilon\}}\le k-1\Bigr)\\
&\le\sum_{j=0}^{n}\sum_{1\le k_1<\dots<k_j\le n}P(Y_i=l\quad\text{for } i\in\{k_1,\dots,k_j\}\text{ and }Y_i\ne l\text{ otherwise})\\
&\quad\times P\Bigl(\sum_{i=1}^{j}1_{\{X_{k_i}\in\overline{K}(x,\varepsilon)\}}\le k-1\,|\,Y_i=l\text{ for }i\in\{k_1,\dots,k_j\}\text{ and }Y_i\ne l\text{ otherwise}\Bigr)\\
&\le\sum_{j\in D_{n,\delta}}P\Bigl(\sum_{i=1}^{j}1_{\{X_i\in\overline{K}(x,\varepsilon)\}}\le k-1\,|\,Y_1=l,\dots,Y_j=l\Bigr)P(S_n^{(l)}=j)+P(S_n^{(l)}\notin D_{n,\delta})\\
&\le\text{ Bin }([n(\pi(\{l\})-\delta)],Q_l(\overline{K}(x,\varepsilon))(\{0,\dots,k-1\})+P(S_n^{(l)}\notin D_{n,\delta})\to 0,
\end{aligned}$$

$\text{Bin}(\cdot,\cdot)$ denoting the binomial distribution. This shows convergence in probability and also almost sure convergence since $\|x-X_{R_{k:n}}\|$ is decreasing. □

To apply property (b) of a GCIM we shall compute probabilities by conditioning with respect to $Y_1,\dots,Y_n$. For this, the following invariance regarding permutations will be useful.

LEMMA 2.2. *Consider a GCIM. Let $x\in\mathbb{R}^d$ and $r_i\in M$ for $i=1,\dots,n$. Then*

$$\begin{aligned}
&P(N_{k,n}^{(1)}=j_1,\dots,N_{k,n}^{(m)}=j_m\,|\,Y_1=r_1,\dots,Y_n=r_n)\\
&\qquad=P(N_{k,n}^{(1)}=j_1,\dots,N_{k,n}^{(m)}=j_m\,|\,Y_1=r_{\tau(1)},\dots,Y_n=r_{\tau(n)})
\end{aligned}$$

for any permutation τ of $\{1,\dots,n\}$.

The proof is immediate from the second assumption for our GCIM.

We now turn to a GCIM such that the probability measures $Q_1,\dots,Q_m$ have densities $f_1,\dots,f_m$ with respect to d-dimensional Lebesgue measure λ^d, that is,

$$f_i=\frac{dQ_i}{d\lambda^d}.$$

For $x\in\mathbb{R}^d$ we define the mappings

$$d_x(z)\colon\mathbb{R}^d\to[0,\infty),\,d_x(z)=\|z-x\|,$$

so that the distributions $Q_1^{d_x},\dots,Q_m^{d_x}$ of these mappings are probability measures on $[0,\infty)$ with continuous distribution functions $F_1^x,\dots,F_m^x$. Hence

$$Q_i^{d_x}\text{ is the conditional distribution of }\|X-x\|\text{ given }Y=i$$

and

$$F_i^x(t) = P(\|X - x\| \le t | Y = i).$$

The asymptotic risk of a k-NN procedure is determined by the asymptotic distribution of the random variables

$$(N_{k,n}^{(1)}, \dots, N_{k,n}^{(m)}).$$

Our main result shows that this asymptotic distribution, hence the asymptotic risk, is the same for a GCIM with Lebesgue densities as in the i.i.d. case.

THEOREM 2.1. *Consider a GCIM with Lebesgue densities. Then*:

(i) *For P^X-almost all $x \in \mathbb{R}^d$*,

$$\lim_{n\to\infty} P(N_{k,n}^{(1)} = j_1, \dots, N_{k,n}^{(m)} = j_m) = M(k, p_1(x), \dots, p_m(x))(\{(j_1, \dots, j_m)\})$$

for any $k \in \mathbb{N}$ and $j_1, \dots, j_m \in \{0, \dots, k\}$ with $\sum_{i=1}^m j_i = k$.

(ii)

$$\lim_{n\to\infty} R(\delta_{k,n})$$
$$= \sum_{y\in M} \int_{\mathbb{R}^d} \sum_{\substack{J\subset M \\ y\in J}} M(k, p_1(x), \dots, p_m(x))(A_J) P(T_J = y)(1 - p_y(x)) P^X(dx).$$

PROOF. It is immediate that (ii) follows from (i):

$$\lim_{n\to\infty} R(\delta_{k,n})$$
$$= \sum_{y\in M} \int_{\mathbb{R}^d} \lim_{n\to\infty} P(\delta_{k,n}(x, Z_n) = y)(1 - p_y(x)) P^X(dx)$$
$$= \sum_{y\in M} \int_{\mathbb{R}^d} \lim_{n\to\infty} \sum_{\substack{J\subset M \\ y\in J}} P((N_{k,n}^{(1)}, \dots, N_{k,n}^{(m)}) \in A_J, T_J = y)(1 - p_y(x)) P^X(dx)$$
$$= \sum_{y\in M} \int_{\mathbb{R}^d} \sum_{\substack{J\subset M \\ y\in J}} M(k, p_1(x), \dots, p_m(x))(A_J) P(T_J = y)(1 - p_y(x)) P^X(dx).$$

So in the remaining part of the proof we shall derive (i). We use the fact that the result is valid for the i.i.d. case, as discussed in Section 1, and shall show that the GCIM and the i.i.d. model have the same asymptotic behavior with regard to a fixed number of nearest neighbors.

As $f_1, \dots, f_m$ are Lebesgue densities for $Q_1, \dots, Q_m$,

$$f = \sum_{i=1}^m \pi(\{i\}) f_i$$

is a Lebesgue density for P^X.

Since $P^X(\{x\in\mathbb{R}^d \mid f(x)=0\})=0$ we only have to consider $x\in\mathbb{R}^d$ with $f(x)>0$. Furthermore the general Lebesgue differentiation theorem states that, for any $i\in M$,

$$\lim_{h\to 0}\frac{Q_i(\overline{K}(x,h))}{\lambda^d(\overline{K}(x,h))}=f_i(x) \quad \text{for}\lambda^d\text{-almost all } x$$

[see, e.g., Wheeden and Zygmund (1977), page 100]. So let us for the following consider x with these properties.

We now introduce the corresponding i.i.d. model. In addition to the GCIM let us consider a sequence $(X'_n, Y'_n)_n$ of i.i.d. random variables, where the conditional distribution of $(X'_n)_n$ given $(Y'_n)_n$ is the same as in the GCIM and each Y'_n has distribution π.

By property (b) of the GCIM and by the law of large numbers there exists a sequence $(\varepsilon_n)_n$, $\varepsilon_n\downarrow 0$, such that for all $i\in M$,

$$P\left(\left|\sum_{j=1}^n 1_{\{i\}}(Y_j)-\pi(\{i\})\right|\geq\varepsilon_n n\right)\to 0$$

and

$$P\left(\left|\sum_{j=1}^n 1_{\{i\}}(Y'_j)-\pi(\{i\})\right|\geq\varepsilon_n n\right)\to 0.$$

Let

$$D_n=\left\{(y_1,\dots,y_n)\in M^n : \left|\sum_{j=1}^n 1_{\{i\}}(y_j)-\pi(\{i\})\right|<\varepsilon_n n\right\}.$$

Then

$$P((Y_1,\dots,Y_n)\in D_n)\to 1, \qquad P((Y'_1,\dots,Y'_n)\in D_n)\to 1.$$

Take $(y_1,\dots,y_n)\in D_n$ and $(z_1,\dots,z_n)\in D_n$, with ordered vectors $(y_{1:n},\dots,y_{n:n})$ and $(z_{1:n},\dots,z_{n:n})$. Denoting the number of i's in these vectors by k_i and l_i, respectively, the definition of D_n shows

$$|k_i-l_i|\leq 2n\varepsilon_n \quad \text{for any } i.$$

In the ordered tuple $(y_{1:n},\dots,y_{n:n})$, the i's are at positions $\sum_{r=1}^{i-1}k_r+1,\dots,$ $\sum_{r=1}^{i}k_r$, and in the ordered vector $(z_{1:n},\dots,z_{n:n})$ at positions $\sum_{r=1}^{i-1}l_r+1,\dots,$ $\sum_{r=1}^{i}l_r$. Since

$$\left|\sum_{r=1}^{i}k_r-\sum_{r=1}^{i}l_r\right|\leq 2in\varepsilon_n,$$

there is a nonoverlap for at most $4in\varepsilon_n$ positions. Hence $(y_{1:n},\dots,y_{n:n})$ and $(z_{1:n},\dots,z_{n:n})$ agree for at least

$$n(1-\delta_n) \text{ positions with } \delta_n=2m(m-1)\varepsilon_n\to 0.$$

We may assume that $n\delta_n$ is an integer and set

$$a(n)=n(1-\delta_n), b(n)=n\delta_n \quad \text{hence } \frac{a(n)}{b(n)}\to\infty.$$

Choose a subset K_n of $a(n)$ positions such that $y_{j:n}$ and $z_{j:n}$ agree at these positions. Then for any $i\in M$, the number of positions j in K_n with value $y_{j:n}=z_{j:n}=i$ is at least $n(\pi(\{i\})-\delta_n-\varepsilon_n)$.

Now let us look at

$$|P(N_{k,n}^{(1)}=j_1,\dots,N_{k,n}^{(m)}=j_m)-P(N_{k,n}^{(1')}=j_1,\dots,N_{k,n}^{(m')}=j_m|,$$

the quantities $N_{k,n}^{(j')}$ being defined as the $N_{k,n}^{(j)}$ but with respect to $(X'_n,Y'_n)_n$. We then obtain

$$\begin{aligned}
&|P(N_{k,n}^{(1)}=j_1,\dots,N_{k,n}^{(m)}=j_m)-P(N_{k,n}^{(1')}=j_1,\dots,N_{k,n}^{(m')}=j_m)|\\
&\quad\le\bigg|\int_{D_n}P(N_{k,n}^{(1)}=j_1,\dots,N_{k,n}^{(m)}=j_m\,|\,Y_1=y_1,\dots Y_n=y_n)P^{(Y_1,\dots,Y_n)}(dy_1,\dots,dy_n)\\
&\qquad-\int_{D_n}P(N_{k,n}^{(1)}=j_1,\dots,N_{k,n}^{(m)}=j_m|Y_1=y_1,\dots Y_n=y_n)\\
&\qquad\times P^{(Y'_1,\dots,Y'_n)}(dy_1,\dots,dy_n)\bigg|+o(1)\\
&\quad\le\sup_{y,z\in D_n}|P(N_{k,n}^{(1)}=j_1,\dots,N_{k,n}^{(m)}=j_m\,|\,Y_1=y_{1:n},\dots,Y_n=y_{n:n})\\
&\qquad-P(N_{k,n}^{(1)}=j_1,\dots,N_{k,n}^{(m)}=j_m\,|\,Y_1=z_{1:n},\dots,Y_n=z_{n:n})|+o(1).
\end{aligned}$$

Note that we have used equality of conditional distributions for the GCIM and the i.i.d. model and permutational invariance according to Lemma 2.2. We thus have to look at two vectors $(y_1,\dots,y_n)$ and $(z_1,\dots,z_n)$ which agree at all positions in some subset K_n with $a(n)$ elements, and we want to give a suitable bound for

$$\begin{aligned}
&|P(N_{k,n}^{(1)}=j_1,\dots,N_{k,n}^{(m)}=j_m\,|\,Y_1=y_1,\dots,Y_n=y_n)\\
&\qquad-P(N_{k,n}^{(1)}=j_1,\dots,N_{k,n}^{(m)}=j_m\,|\,Y_1=z_1,\dots,Y_n=z_n)|.
\end{aligned}$$

For our estimate, this bound has to be independent of the particular $(y_1,\dots,y_n)$ and $(z_1,\dots,z_n)$. We introduce the following probabilistic model where all random variables are defined on the same probability space.

If $j\in K_n$, we consider random variables $W_j=W'_j$ with distribution $Q_{y_j}^{d_x}$, the conditional distribution of $||X-x||$ given $Y=y_j$. If $j\notin K_n$, we let W_j have distribution $Q_{y_j}^{d_x}$ and let W'_j have distribution $Q_{z_j}^{d_x}$. Of course all random variables are chosen to be independent.

The probability $P(N_{k,n}^{(1)}=j_1,\dots,N_{k,n}^{(m)}=j_m\,|\,Y_1=y_1,\dots,Y_n=y_n)$ may obviously be expressed in terms of the W_j and $P(N_{k,n}^{(1)}=j_1,\dots,N_{k,n}^{(m)}=j_m\,|\,Y_1=z_1,\dots,Y_n=z_n)$ in terms of the W'_j.

Using $|P(A)-P(B)|\leq P(A\cap B^c)+P(B\cap A^c)$ we are led to the following bound: on $(A\cap B^c)\cup(B\cap A^c)$ the k smallest members of the W_j and of the W'_j cannot agree. Defining

$$W_{k:a(n)} \text{ as the } k\text{-smallest of } W_j,\ j\in K_n$$

and

$$W_{1:b(n)} \text{ as the smallest of } W_j,\ j\notin K_n,$$

similarly $W'_{k:a(n)}, W'_{1:b(n)}$, we have

$$W_{k:a(n)}=W'_{k:a(n)}$$

and

$$\begin{aligned}&\big|P(N_{k,n}^{(1)}=j_1,\dots,N_{k,n}^{(m)}=j_m\,|\,Y_1=y_1,\dots,Y_n=y_n)\\&\quad -P(N_{k,n}^{(1)}=j_1,\dots,N_{k,n}^{(m)}=j_m\,|\,Y_1=z_1,\dots,Y_n=z_n)\big|\\&\leq P(W_{k:a(n)}\geq W_{1:b(n)})+P(W'_{k:a(n)}\geq W'_{1:b(n)}).\end{aligned}$$

We finally have to find a bound for $P(W_{k:a(n)}\geq W_{1:b(n)})$, the second probability being of exactly the same type.

Since $f(x)>0$ we may choose l such that $f_l(x)>0$. Let $c(n)=[n(\pi(\{l\})-\delta_n-\varepsilon_n)]$. We choose additional independent random variables V_j, V'_j on the same probability space such that all the V_j have distribution $Q_l^{d_x}$, hence distribution function F_l^x, and the V'_j have distribution function F' given by

$$F'=\max_i F_i^x.$$

Among the random variables which contribute to $W_{k:a(n)}$, there are at least $c(n)$ with distribution function F_l^x. Furthermore, the V'_j are stochastically smaller than any of the random variables contributing to $W_{1:b(n)}$. This implies

$$P(W_{k:a(n)}\geq W_{1:b(n)})\leq P(V_{k:c(n)}\geq V'_{1:b(n)}).$$

Since $c(n)/b(n)\to\infty$, it follows easily from standard results of extreme value theory that

$$P(V_{k:c(n)}\geq V'_{1:b(n)})\to 0,$$

completing the proof. □

The precise argument for this convergence to 0 is given in the concluding Lemma 2.3. To see that its assumptions hold in our situation we use that

$$\lim_{h\to 0}\frac{Q_i(\overline{K}(x,h))}{h^d}=f_i(x)\lambda^d(\overline{K}(x,1)),\qquad i\in M,$$

hence

$$nF_l\left(\frac{z}{n^{1/d}}\right) \to z^d f_l(x)\lambda^d(\overline{K}(0,1)) > 0$$

and

$$nF'\left(\frac{z}{n^{1/d}}\right) \to z^d \max_i f_i(x)\lambda^d(\overline{K}(0,1)) > 0.$$

Then we may apply Lemma 2.3 with $e_n = d_n = n^{1/d}$ and $\alpha(n) = c(n), \beta(n) = b(n)$.

LEMMA 2.3. *Let $(V_n)_n$ and $(V'_n)_n$ be i.i.d. sequences of real-valued random variables with distribution functions F and F', respectively. Assume $F(0) = F'(0) = 0$ and that there exist sequences $(d_n)_n$ and $(e_n)_n$ with $d_n \to \infty$, $e_n \to \infty$, such that*

$$nF\left(\frac{z}{d_n}\right) \text{ and } nF'\left(\frac{z}{e_n}\right) \text{ converge in } (0,\infty)$$

for all $z > 0$.

Then for any sequences of integers $(\alpha(n))_n$, $(\beta(n))_n$ with $\alpha(n) \to \infty$, $\beta(n) \to \infty$,

$$\frac{d_{\alpha(n)}}{e_{\beta(n)}} \to \infty \text{ implies } P(V_{j:\alpha(n)} \geq V'_{k:\beta(n)}) \to 0$$

for all fixed j, k.

PROOF. It is well known from extreme value theory [see, e.g., Leadbetter, Lindgren and Rootzen (1983)] that the conditions on F and F', respectively, imply the convergence in distribution of $d_n V_{j:n}$ and $e_n V'_{k:n}$ to random variables with support on $(0,\infty)$. For $\varepsilon > 0$ we may thus choose $\delta > 0$ such that

$$P(e_{\beta(n)} V'_{k:\beta(n)} < \delta) \leq \varepsilon$$

for all sufficiently large n, and consequently,

$$\begin{aligned} &P(V_{j:\alpha(n)} \geq V'_{k:\beta(n)}) \\ &\quad \leq P(e_{\beta(n)} V'_{k:\beta(n)} < \delta) + P\left(e_{\beta(n)} V_{k:\beta(n)} \geq \delta, \frac{e_{\beta(n)}}{d_{\alpha(n)}} d_{\alpha(n)} V_{j:\alpha(n)}) \geq \delta\right) \\ &\quad \leq \varepsilon + P\left(d_{\alpha(n)} V_{j:\alpha(n)} \geq \delta \frac{d_{\alpha(n)}}{e_{\beta(n)}}\right). \end{aligned}$$

The last probability tends to 0 which proves the result. □

Looking at the proof, we can give the following local formulation of Theorem 2.1(i) which will be used in Section 3.

Let U be an open subset of $\mathbb{R}^d$ and $\mathscr{U} = \{A \cap U \mid A \in \mathscr{B}^d\}$ the Borel σ-algebra restricted to U. If we assume that the restricted measures $Q_1|_{\mathscr{U}}, \ldots, Q_m|_{\mathscr{U}}$ have densities with respect to $\lambda^d|_{\mathscr{U}}$ then the conclusion (i) of Theorem 2.1 holds for P^X-almost all $x \in U$.

3. Discrete models.

We continue to consider a GCIM but shall now turn to models where $P(X=x)>0$ occurs.

Ranking the training sequence. In the case of such models we of course have to specify how to handle equal distances in the allocation of nearest neighbors. It is natural to distribute the possible ranks with equal probability among those who have the same distance from the x at hand. We shall call this fair allocation. This can be done by the following randomization procedure.

Let $U_1,\dots,U_n$ be i.i.d. with continuous distribution, also stochastically independent of $(X,Y), Z_n$. Then the $R_{i:n}$ are almost surely uniquely determined by the requirement that for $i<j$,

$$\|x-X_{R_{i:n}}\|<\|x-X_{R_{j:n}}\| \text{ or } \|x-X_{R_{i:n}}\|=\|x-X_{R_{j:n}}\|, \qquad U_{R_{i:n}}\le U_{R_{j:n}},$$

where

$$R_{i:n}=R_{i:n}(x;X_1,U_1,\dots,X_n,U_n).$$

Of course, any continuous distribution used in this way leads to fair allocation.

We also note that, as in the previous parts, the dependence on x is suppressed in our notations so that, for example, $Y_{R_{i:n}}=Y_{R_{i:n}(x)}=Y_{R_{i:n}(x;X_1,U_1,\dots,X_n,U_n)}$.

We shall now prove that, in the case of fair allocation, the same result as in the case of Lebesgue densities holds.

THEOREM 3.1. *Consider a GCIM with fair allocation. Let $x\in\mathbb{R}^d$ such that $P(X=x)>0$. Then for all $k\in\mathbb{N}$, and $j_1,\dots,j_m\in\{0,\dots,k\}$ with $\sum_{i=1}^m j_i=k$,*

$$\lim_{n\to\infty} P(N_{k,n}^{(1)}=j_1,\dots,N_{k,n}^{(m)}=j_n)=M(k,p_1(x),\dots,p_m(x))(\{(j_1,\dots,j_m)\}).$$

PROOF. We start by specifying fair allocation. Choose $r>0$. Consider i.i.d. random variables V_i taking values in $\mathbb{R}^d$ with distribution $\widetilde{Q}$ given by

$$\widetilde{Q}(A)=\frac{\lambda^d(A\cap\overline{K}(0,r))}{\lambda^d(\overline{K}(0,r))}.$$

For fair allocation we use the sequence of random variables

$$U_1=\|V_1\|, U_2=\|V_2\|,\dots.$$

An equivalent problem is obtained in the following way: we consider the mapping

$$\phi_x\colon \mathbb{R}^d\times\mathbb{R}^d\mapsto\mathbb{R}^d,\ \phi_x(z,v)=\begin{cases} x+v, & \text{if } z=x,\\ z+r\dfrac{z-x}{\|z-x\|}, & \text{if } z\ne x\end{cases}$$

and the random variables $\phi_x(X_i,V_i)$ $i=1,\dots,n$. Given $X_{R_{k:n}}=x$ we have for $i<j\le k$,

$$U_{R_{i:n}}\le U_{R_{j:n}} \quad \text{if and only if } \|x-\phi_x(X_{R_{i:n}},V_{R_{i:n}})\|\le\|x-\phi_x(X_{R_{j:n}},V_{R_{j:n}})\|.$$

Of course as in Lemma 2.1,

$$\lim_{n\to\infty} P(X_{R_{k:n}}=x)=1,$$

hence asymptotically the two problems

(I) Classify x using $((X_1,U_1,Y_1),\dots,(X_n,U_n,Y_n))$

and

(II) Classify x using $((\phi_x(X_1,V_1),Y_1),\dots,(\phi_x(X_n,V_n),Y_n))$

are equivalent.

The distribution

$$P(\phi_x(X_i,V_i)\in\cdot\,|\,Y_i=y)=(Q_y\otimes\widetilde{Q})^{\phi_x}$$

is given by

$$(Q_y\otimes\widetilde{Q})^{\phi_x}(A\cap\overline{K}(x,r))=\int_{A\cap\overline{K}(x,r)}\frac{Q_y(\{x\})}{\lambda^d(\overline{K}(0,r))}\,d\lambda^d,\qquad A\in\mathscr{B}^d.$$

We may thus apply the local version of Theorem 2.1(i) to $(\phi_x(X_n,V_n),Y_n)_n$ and obtain

$$\begin{aligned}\lim_{n\to\infty} P\Big(N_{k,n}^{(1)}=j_1,\dots,N_{k,n}^{(m)}=j_m\Big)\\ &=\lim_{n\to\infty}P\Big(N_{k,n}^{(1)}=j_1,\dots,N_{k,n}^{(m)}=j_m,X_{R_{k:n}}=x\Big)\\ &=M(k,p_1(x),\dots,p_m(x))(\{(j_1,\dots,j_m)\}).\qquad\square\end{aligned}$$

Let us call a distribution discrete if the support is finite or countably infinite without an accumulation point. Then we immediatedly obtain the asymptotic risk for discrete distributions.

THEOREM 3.2. *Consider a GCIM with fair allocation. Let P^X be a discrete distribution. Then*:

(i) *For P^X-almost all $x\in\mathbb{R}^d$,*

$$\lim_{n\to\infty}P\Big(N_{k,n}^{(1)}=j_1,\dots,N_{k,n}^{(m)}=j_m\Big)=M(k,p_1(x),\dots,p_m(x))(\{(j_1,\dots,j_m)\})$$

for any $k\in\mathbb{N}$ and $j_1,\dots,j_m\in\{0,\dots,k\}$ with $\sum_{i=1}^m j_i=k$.

(ii) $\lim_{n\to\infty}R(\delta_{k,n},P)$

$$=\sum_{y\in M}\int_{\mathbb{R}^d}\sum_{\substack{J\subset M\\ y\in J}}M(k,p_1(x),\dots,p_m(x))(A_J)P(T_J=y)(1-p_y(x))P^X(dx).$$

The proof follows from Theorem 3.1, with (ii) an immediate consequence of (i) as in Theorem 2.1.

3.1. *Nearest neighbor classification for different allocation procedures.* In the case of discrete distributions it may be suspected that different types of allocation also lead to different expressions for the risk. We shall look into this problem for the special case of hidden Markov models.

We assume that $M=\{1,2\}$ and that we have a GCIM such that $(Y_n)_{n\in\mathbb{N}}$ is an ergodic Markov chain with stationary distribution $\pi=P^Y$.

We denote the transition probabilities by $p_{yz}=P(Y_n=z\,|\,Y_{n-1}=y)$ and the initial distribution by μ.

Types of allocation.

1. In the case of equal distances $\|x-X_{R_{i:n}}\|=\|x-X_{R_{j:n}}\|$ for $i<j$ we prescribe $R_{i:n}<R_{j:n}$.
   Setting $\tau=\tau(x)=\inf\{k\in\mathbb{N}\,|\,X_k=x\}$ it follows that, on the event $X_{R_{1:n}}=x$,
   $$R_{1:n}=\tau,$$
   hence
   $$\lim_{n\to\infty}P_\mu(Y_{R_{1:n}}=y)=P_\mu(Y_\tau=y)$$
   if $P(X=x)>0$.
2. For the second type of allocation, we use in the case of equal distances $\|x-X_{R_{i:n}}\|=\|x-X_{R_{j:n}}\|$ for $i<j$ the prescription $R_{i:n}>R_{j:n}$.
   Setting $\tilde{\tau}_n=\tilde{\tau}_n(x)=\sup\{k\in\{1,\dots,n\}\,|\,X_k=x\}$ we obtain that, on the event $X_{R_{1:n}}=x$,
   $$R_{1:n}=\tilde{\tau}_n,$$
   hence
   $$\lim_{n\to\infty}P_\mu(Y_{R_{1:n}}=y)=\lim_{n\to\infty}P_\mu(Y_{\tilde{\tau}_n}=y)$$
   if $P(X=x)>0$.

In the following we shall consider the case of only two classes for explicitly obtaining the asymptotic distribution $\lim_{n\to\infty}P_\mu(Y_{R_{1:n}}=y)$. Of course this immediatedly gives the asymptotic risk for 1-NN classification.

We start with allocation rule (1).

THEOREM 3.3. *Consider a hidden Markov model with $M=\{1,2\}$ using allocation rule* (1). *Then for any $x\in\mathbb{R}^d$ with $P^X(\{x\})>0$ and $i=1,2$,*

$$\begin{aligned}\lim_{n\to\infty}P_\mu(Y_{R_{1:n}}=i) = {}& \frac{P(Y=i,X=x)}{P(X=x)+\Delta_x}+\Delta_x\frac{P(Y=i)}{P(X=x)+\Delta_x}\\ &+\frac{(\mu(\{i\})-\pi(\{i\}))/(p_{12}+p_{21})Q_1(\{x\})Q_2(\{x\})}{P(X=x)+\Delta_x},\end{aligned}$$

where $\Delta_x=\frac{p_{11}-p_{21}}{p_{12}+p_{21}}Q_1(\{x\})Q_2(\{x\})$.

PROOF. We have

$$P_\mu(\tau<\infty)$$

and

$$\lim_{n\to\infty} P_\mu(Y_{R_{1:n}}=i)=P_\mu(Y_\tau=i), \qquad i=1,2.$$

Obviously,

$$\begin{aligned}
&P_1(Y_\tau=1)\\
&\quad=\sum_{n=1}^{\infty} P_1(Y_n=1,\tau=n)\\
&\quad=P_1(\tau=1)+\sum_{n=2}^{\infty}\bigl[P(X_1\neq x|Y_1=1)P_1(Y_2=1)P_1(Y_{n-1}=1,\tau=n-1)\\
&\qquad\quad+P(X_1\neq x|Y_1=1)P_1(Y_2=2)P_2(Y_{n-1}=1,\tau=n-1)\bigr]\\
&\quad=Q_1(\{x\})+p_{11}(1-Q_1(\{x\}))P_1(Y_\tau=1)+p_{12}(1-Q_1(\{x\}))P_2(Y_\tau=1)
\end{aligned}$$

and

$$P_2(Y_\tau=1)=p_{12}(1-Q_2(\{x\}))P_1(Y_\tau=1)+p_{22}(1-Q_2(\{x\}))P_2(Y_\tau=1).$$

We may solve this to obtain

$$\begin{aligned}
P_1(Y_\tau=1)&=\frac{p_{12}Q_1(\{x\})(1-Q_2(\{x\})}{p_{12}Q_2(\{x\})+p_{21}Q_1(\{x\})+(p_{11}-p_{21})Q_1(\{x\})Q_2(\{x\})}\\
P_2(Y_\tau=1)&=\frac{Q_1(\{x\})(p_{21}+Q_2(\{x\})-p_{21}Q_2(\{x\})}{p_{12}Q_2(\{x\})+p_{21}Q_1(\{x\})+(p_{11}-p_{21})Q_1(\{x\})Q_2(\{x\})},
\end{aligned}$$

which implies

$$\begin{aligned}
&P_\mu(Y_\tau=1)\\
&\quad=\mu(\{1\})P_1(Y_\tau=1)+\mu(\{2\})P_2(Y_\tau=1)\\
&\quad=\frac{p_{21}Q_1(\{x\})+(\mu(\{1\})-p_{21})Q_1(\{x\})Q_2(\{x\})}{p_{12}Q_2(\{x\})+p_{21}Q_1(\{x\})+(p_{11}-p_{21})Q_1(\{x\})Q_2(\{x\})}\\
&\quad=\frac{P(Y=1,X=x)}{P(X=x)+\Delta_x}+\Delta_x\frac{P(Y=1)}{P(X=x)+\Delta_x}\\
&\qquad+\frac{\bigl(\mu(\{1\})-\pi(\{1\})\bigr)/\bigl(p_{12}+p_{21}\bigr)Q_1(\{x\})Q_2(\{x\})}{P(X=x)+\Delta_x}
\end{aligned}$$

using

$$\pi(\{1\})=\frac{p_{21}}{p_{12}+p_{21}} \quad\text{and}\quad \pi(\{2\})=\frac{p_{12}}{p_{12}+p_{21}}. \qquad \square$$

EXAMPLE. Obviously this leads to results which differ from the independent case. To illustrate this we consider the following example: let the transition matrix be given as

$$\begin{bmatrix} 0.99 & 0.01 \\ 0.1 & 0.9 \end{bmatrix}$$

and let

$$Q_1(\{x\})=0.03, \qquad Q_2(\{x\})=0.99.$$

The stationary distribution is $\pi(\{1\})=\frac{10}{11}$, $\pi(\{2\})=\frac{1}{11}$ and we have

$$P_1(Y_\tau=1)=0.756, \qquad P_2(Y_\tau=1)=7.6\cdot 10^{-4}.$$

So we have convergence of $P_\pi(Y_{R_{1:n}}=1)$ to

$$P_\pi(Y_\tau=1)=0.687,$$

whereas in the case of fair allocation $P_\pi(Y_{R_{1:n}}=1)$ converges to

$$P_\pi(Y=1|X=x)=0.233.$$

We now consider allocation procedure (2).

THEOREM 3.4. *Consider a hidden Markov model with* $M=\{1,2\}$ *using allocation rule* (2). *Then for any* $x\in\mathbb{R}^d$ *with* $P^X(\{x\})>0$ *and* $i=1,2$,

$$\lim_{n\to\infty} P_\mu(Y_{R_{1:n}}=i)=\frac{P(Y=i,X=x)}{P(X=x)+\Delta_x}+\Delta_x\frac{P(Y=i)}{P(X=x)+\Delta_x},$$

where $\Delta_x=\frac{p_{11}-p_{21}}{p_{12}+p_{21}}Q_1(\{x\})Q_2(\{x\})$.

PROOF. We may assume $0<Q_i(\{x\})<1$ for $i=1,2$, the result being obvious otherwise. Setting $d_n=d_x(X_{R_{1:n}})$ we consider the Markov process $((Y_n,Y_{R_{1:n}}, d_n))_{n\in\mathbb{N}}$.

For any initial distribution μ, the process reaches the absorbing set

$$\mathscr{R}=\{(1,1,0),(2,1,0),(1,2,0),(2,2,0)\}$$

with probability 1. The transition matrix for the Markov chain induced on $\mathscr{R}$ is given by

$$P_{\mathscr{R}}=\begin{bmatrix} p_{11} & p_{12}(1-Q_2(\{x\})) & 0 & p_{12}Q_2(\{x\}) \\ p_{21} & p_{22}(1-Q_2(\{x\})) & 0 & p_{22}Q_2(\{x\}) \\ p_{11}Q_1(\{x\}) & 0 & p_{11}(1-Q_1(\{x\})) & p_{12} \\ p_{21}Q_1(\{x\}) & 0 & p_{21}(1-Q_1(\{x\})) & p_{22} \end{bmatrix}.$$

We thus have an ergodic Markov chain with stationary distribution $\tilde{\pi}$ on $\mathscr{R}$, determined by

$$\tilde{\pi}P_{\mathscr{R}}=\tilde{\pi}.$$

The quantities we are looking for are given by

$$\lim_{n\to\infty} P_\mu(Y_{R_{1:n}}=2) = \tilde{\pi}(\{(1,2,0),(2,2,0)\}), \lim_{n\to\infty} P_\mu(Y_{R_{1:n}}=1) = 1-\tilde{\pi}(\{(1,2,0),(2,2,0)\}).$$

We may compute explicitly

$$\begin{aligned}\tilde{\pi}(\{(1,2,0)\}) &= \frac{\pi(\{2\})Q_2(\{x\})p_{21}(1-Q_1(\{x\}))}{p_{12}Q_2(\{x\})+p_{21}Q_1(\{x\})+(p_{11}-p_{21})Q_1(\{x\})Q_2(\{x\})}\\ &= \pi(\{2\})Q_2(\{x\})\frac{P(Y=1)-p_{21}Q_1(\{x\})/\big(p_{21}+p_{12}\big)}{P(X=x)+\Delta_x}\end{aligned}$$

and

$$\begin{aligned}\tilde{\pi}(\{(2,2,0)\}) &= \frac{\pi(\{2\})Q_2(\{x\})(p_{12}+p_{11}Q_1(\{x\}))}{p_{12}Q_2(\{x\})+p_{21}Q_1(\{x\})+(p_{11}-p_{21})Q_1(\{x\})Q_2(\{x\})}\\ &= \pi(\{2\})Q_2(\{x\})\frac{P(Y=2)+p_{11}Q_1(\{x\})/p_{21}+p_{12}}{P(X=x)+\Delta_x}.\end{aligned}$$

This implies

$$\begin{aligned}\lim_{n\to\infty} P_\mu(Y_{R_{1:n}}=2) &= \pi(\{2\})\frac{Q_1(\{x\})+\Delta_x}{P(X=x)+\Delta_x}\\ &= \frac{P(Y=2,X=x)}{P(X=x)+\Delta_x}+\Delta_x\frac{P(Y=2)}{P(X=x)+\Delta_x},\end{aligned}$$

hence the result. □

Acknowledgments. The authors thank both referees and the Associate Editor for their helpful comments, which led to a highly improved presentation of the results. In particular, the present proof of Theorem 2.1 was suggested by one of the referees. The original proof by the authors was much more complicated.

REFERENCES

ADAMS, T. M. and NOBEL, A. B. (1998). On density estimation from ergodic processes. *Ann. Probab.* **26** 794–804.

BICKEL, P. J., RITOV, Y. and RYDEN, T. (1998). Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Ann. Statist.* **26** 1614–1635.

COVER, T. M. and THOMAS, J. A. (1991). *Elements of Information Theory*. Wiley, New York.

COVER, T. M. and HART, P. E. (1967). Nearest neighbor pattern classification. *IEEE Trans. Inform. Theory* **13** 21–27.

DASARATHY, B. (1991). *Nearest Neighbor Classification Techniques*. IEEE, Los Alamitos, CA.

Devroye, L. (1981a). On the inequality of Cover and Hart in nearest neighbor discrimination. *IEEE Trans. Pattern Anal. Mach. Intelligence* **3** 75–78.

DEVROYE, L. (1981b). On the asymptotic probability of error in nonparametric discrimination. *Ann. Statist.* **9** 1320–1327.

DEVROYE, L., GYÖRFI, L., KRZYZAK, A. and LUGOSI, G. (1994). On the strong universal consistency of nearest neighbor regression function estimates. *Ann. Statist.* **22** 1371–1385.

DEVROYE, L., GYÖRFI, L. and LUGOSI G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer, New York.

FIX, E. and HODGES, J. (1951). Discriminatory analysis. Nonparametric discrimination: consistency properties. Technical report, USAF School of Aviation Medicine, Randolph Field.

FIX, E. and HODGES, J. (1952). Discriminatory analysis: small sample performance. Technical report, USAF School of Aviation Medicine, Randolph Field.

GYÖRFI, L., HÄRDLE, W., SARDA, P. and VIEU, V. (1989). *Nonparametric Curve Estimation from Time Series. Lecture Notes Statistics* **60** Springer, Berlin.

GYÖRFI, L. and LUGOSI, G. (1992). Kernel density estimation from ergodic samples is not universally consistent. *Comp. Statist. Data Anal.* **24** 437–442.

HUANG, X. D., ARIKI, Y. and JACK, M. A. (1990). *Hidden Markov Models for Speech Recognition*. Edinburgh Univ. Press.

IRLE, A. (1997). On consistency in nonparametric estimation under mixing assumptions. *J. Multivariate Anal.* **60** 123–147.

KAHAN, S., PAVLIDES, T. and BAIRD, H. S. (1987). On the recognition of printed characters of any font and size. *IEEE Trans. Pattern Anal. Mach. Intelligence* **9** 274–288.

KULKARNI, S. R. and POSNER, S. E. (1995). Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Trans. Inform. Theory* **41** 1028–1039.

LEADBETTER, M. R., LINDGREN, G. and ROOTZEN, H. (1983). *Extremes and Related Properties of Random Sequences and Processes*. Springer, New York.

MACDONALD, I. and ZUCCHINI, W. (1997). *Hidden Markov and Other Models for Discrete-valued Time-series*. Chapman and Hall, London.

MORVAI, G., YAKOWITZ, S. and GYÖRFI, L. (1996). Nonparametric inference for ergodic stationary time series. *Ann. Statist.* **24** 370–379.

NOBEL, A. B., MORVAI, G. and KULKARNI, S. R. (1998). Density estimation from an individual numerical sequence. *IEEE Trans. Inform. Theory* **44** 537–541.

SHANNON, C. E. (1951). Prediction and entropy of handwritten English. *Bell Systems Tech. J.* **30** 50–64.

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. Mach. Intelligence* **3** 75–78.

SNAPP, R. S. and VENKATESH, S. S. (1998). Asymptotic expansion of the k nearest neighbor risk. *Ann. Statist.* **26** 850–878.

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

WHEEDEN, R. L. and ZYGMUND A. (1977). *Measure and Integral*. Marcel Dekker, New York.

MATHEMATISCHES SEMINAR
DER UNIVERISTÄT KIEL
LUDEWIG-MEYN STR. 4
D-24908 KIEL
GERMANY
E-MAIL: irle@math.uni-kiel.de