Proceedings

Pattern Recognition and Image Processing

MAY 31 - JUNE 2, 1978
CHICAGO, ILLINOIS

IEEE Computer Society
Conference on

Additional copies available from:
IEEE Computer Society
5855 Naples Plaza, Suite 301
Long Beach, California 90803

IEEE Service Center
445 Hoes Lane
Piscataway, NJ 08854

Institute of Electrical and Electronics Engineers, Inc., New York, N.Y.,
A UNIVERSAL K-NEXTNEIGHBOR
PROCEDURE IN DISCRIMINATION

LUC P. DEVROYE

Mc Gill University
P.O. Box 6070
Montreal, Canada H3C 3G1

Abstract
The classical k-nearest neighbor rule's performance is scale dependent. The data dependent partitioning rules whose error probabilities are invariant under monotone transformations of the coordinate axes appear unnatural because of the discontinuities inherent to all partitioning rules. In this note a k-nearest neighbor rule based upon the notion of empirical distance is discussed that (i) is Bayes risk consistent for all possible distributions of the observations, (ii) has the desired invariance property.

Introduction
In several applications one would like, for physical or economical reasons, to feed the freshly collected data to some machine without preprocessing. In discrimination for instance, a mere rescaling of the coordinate axes has a serious impact on the performance of the popular k-nearest neighbor rules. On the other hand, the k-nearest neighbor rule is appealing because of its simple definition, easy implementation and distribution-free properties for small samples and large samples. In this note we discuss a variation of the k-nearest neighbor rule whose performance (that is, the probability of error given the data) is not affected by monotone transformations of the coordinate axes. Such a modification would save the user a tedious preprocessing step and would save the manufacturer the trouble of adding input specifications to his discrimination machine.

The data we have collected can be regarded as a sequence of independent identically distributed random vectors: \( D_i = (X_i, Y_i) \), \( i = 1, ..., n \). Here \( X_i \in \mathbb{R}^d \) is called an observation and \( Y_i \in \{1, 2, ..., M\} \) is called a state (class). If \( (X, Y) \) is distributed as \( (X_1, Y_1) \) and is independent of \( D_n \), then in the next-nearest neighbor rule we estimate \( Y \) from \( X \) and \( D_n \) by \( \hat{Y}_{n} = \text{argmax} \sum \mathbb{1}_{y \neq \hat{Y}_{n}} \),

\[
\hat{Y}_{n} = \arg\max_{1 \leq i < k} \sum_{1 \leq j < M} \mathbb{1}_{y_j = 1} X_j(X) = 1, \tag{1}
\]

(\(X_1(X), Y_1(X), ..., X_n(X), Y_n(X)\) is a reordering of the data according to increasing values of \( |X_1 - X| \) and \( I \) is the indicator function. The scaling of the coordinate axes influences the rule through the norm \(|\cdot|\). We will write \( L_n \) for the prob-

\[ L_n = \Pr[\hat{Y}_{n} \neq Y | D_n] \]

and \( L^* \) for the Bayes probability of error,

\[ L^* = \inf_{g: \mathbb{R}^d = \{1, ..., M\}} \Pr(g(X) \neq Y) \]

Stone showed that \( L_n \geq L^* \) in probability for all possible distributions of \( (X, Y) \) whenever \( k \geq \text{and } k/n \geq 0. \tag{2} \)

Thus, the k-nearest neighbor rule behaves asymptotically well without exceptions. To continue our discussion, let us replace the \( X_i = (X_i^1, ..., X_i^d) \) and \( X = (X^1, ..., X^d) \) by \( \psi(X_i) = (\psi_1(X_i^1), ..., \psi_d(X_i^d)) \)

and \( \psi(X) = (\psi_1(X^1), ..., \psi_d(X^d)) \)

where \( \psi_1, ..., \psi_d \) are one-to-one monotone transformations from \( R \) to \( R \). Examples are a dilation \( \psi_1(u) = au \), a limiter \( \psi_2(u) = u/(1 + |u|) \) and an expansion \( \psi_3(u) = \exp(au) \). Writing \( D_n' \) for \( (\psi(X_1, Y_1), ..., (\psi(X_n, Y_n) \) it is true that the Bayes probability of error \( L^*(\psi) \) is independent of \( \psi := \psi(X) \neq Y | D_n' \)

\[ L^*(\psi) = \inf_{g \in \mathbb{R}^d = \{1, ..., M\}} \Pr(g(\psi(X)) \neq Y) \tag{4} \]

and that the conditional probability of error with rule (1) and \( D_n' \),

\[ L_n(\psi) = \Pr(g_n(\psi(X), D_n') \neq Y | D_n') \]

varies with \( \psi \). Preprocessing of the data is partly concerned with finding a transformation \( \psi \) that tends to yield low values for \( L_n(\psi) \). If such a \( \psi \) is a priori fixed, then Stone's theorem remains valid (that is, \( L_n(\psi) \rightarrow L^* \) in probability if (2) holds). But if \( \psi \) is found after we have peeked at the data, can we still draw the same conclusion?

In essence, we are looking for a discrimination rule with the following properties:

(i) \( L_n(\psi) = L_n \) for all monotone transformations \( \psi \)

(see (3)),

(ii) \( L_n \geq L^* \) in probability for all possible distributions of \( (X, Y) \).

The impact of (1) is that preprocessing via scaling or any other monotone transformation does not reduce the probability of error. While for finite \( n \) this may or may not be an advantage, (ii) assures that at least for large \( n \) the probability of error

CH1318-5/78/0000-0142$00.75 © 1978 IEEE
is close to the Bayes probability of error. In fact, (14) implies that this asymptotically nice behavior is guaranteed for all possible (unknown) distributions of \((X,Y)\). In the next section we develop a discrimination rule featuring (1-11) which in form coincides with the k-nearest neighbor rule. A similar discrimination rule was recently suggested by Oshen in a comment on Stone's paper.\footnote{1}

A Universal K-Nearest Neighbor Discrimination Rule

The k-nearest neighbor rule we are discussing is based upon the notion of empirical distance. Because in general X does not have a density, all sorts of situations involving ties can occur. For instance, the event \(X_1^2 = X_2^2\) can happen with positive probability. It is for this reason that we propose to enlarge the data by generating \(Z_1, \ldots, Z_n\), a sequence of independent random variables, independent of \((X,Y,D_n)\), and all uniformly distributed on \((0,1)\). The data \(D_n = (X_1, Y_1, Z_1, \ldots, X_n, Y_n, Z_n)\) is then reordered according to increasing values for the empirical distances \(\rho(X_i, Z_i, (X,Z))\) between \((X_i, Z_i)\) and \((X,Z)\) defined below. Notice that from here on \(D_n\) denotes the enlarged data sequence. The reordered data sequence is \((X, (X,Z), Y, (X,Z), Z, (X,Z))\), \ldots, \((X, (X,Z), Y, (X,Z), Z, (X,Z))\). Since \(\rho\) is integer-valued, we will use the values of \(|Z_i - Z_j|\) to break ties in this reordering process. This has the same effect as reordering \(D_n\) according to increasing values of the adjusted empirical distance

\[
\rho'(X_i, Z_i, (X,Z)) = \rho(X_i, Z_i, (X,Z)) + |Z_i - Z_j|.
\]

To define \(\rho\) first order the \((X_i, Z_i)\) and \((X,Z)\) according to increasing values of their first components, \(X_i^1, X^2, \ldots, X^d, Z_i^1, X^2, \ldots, X^d\). If ties occur, subsequently break them by considering increasing values for the \(Z_i\) or \(Z_j\). Let \(r_i^1\) be the rank of \((X_i, Z_i)\) and \(r_j^1\) be the rank of \((X,Z)\). Repeating the same procedure for the remaining \(d-1\) components gives us values for \(r_i^1, r_i^2, 1 \leq j \leq d, 1 \leq i \leq n\).

We define

\[
\rho(X_i, Z_i, (X,Z)) = \max_{1 \leq j \leq d} |r_i^j - r_j^j|.
\]

In Figure 1 the empirical distances are given in an example with no componentwise ties and \(d=2\). The physical meaning of \(\rho(X_i, Z_i, (X,Z))\) in that case is very simple: the empirical distance is the maximal number of lines in the \(d\)-dimensional grid generated by the \(X_i\) that one can cross or touch if one wants to go from \(X\) to \(X_i\) along one of the coordinate axes.

It is easy to check that for all transformations (3)

\[
\rho(\psi(X_i), Z_i, (\psi(X), Z)) = \rho(X_i, Z_i, (X,Z))
\]

and that therefore the desired invariance property (1) holds true for all discrimination rules \(g_n\) that are functions of the empirical distances only.

Consider for instance the rules in which \(Y\) is estimated by \(Y' = g_n (X,Z,D_n)\) where

\[
y' = \max_{1 \leq j \leq m} \sum_{1 \leq i \leq n} v_j I(Y_j (X,Z) = i)\]

and the weights \(v_1, \ldots, v_m\) vary with \(n\) such that:

\[
v_1 \geq v_2 \geq \ldots \geq v_m \geq 0 ; \sum v_j = 1
\]

and

\[
\sup v_i \geq 0 ; \sum v_i \geq 0 \text{ for some } k \text{ with } k/n \geq 0.
\]

Picking the weights \(v_1 = v_2 = \ldots = v_k = 1/k\) and \(v_i = 0\) otherwise yields a rule that equally weights the \(k\) nearest neighbors.

In the next section we prove the following theorem.

Theorem. For all rules satisfying (6-9), \(L_n \leq L^*\) in probability. This property is valid for all distributions of \((X,Y)\).

Oshen, in his comment on Stone's paper, announced a similar theorem for this class of rules under the additional restriction that all the marginal distributions of \(X\) are atom-free (which in effect eliminates the possibility of ties). In the proof of our theorem we will follow the train of thought of Stone's proof of his corresponding theorem for the classical k-nearest neighbor rules.
Remark 1. (A generalization.) Instead of $D_n$ one can of course use $D_n'$ which is obtained from $D_n$ by considering $(\phi(x), Y_{1}(x), Z_{1})$ instead of $(x_{1}, Y_{1}, Z_{1})$. Here $\phi$ is an a priori given but otherwise arbitrary mapping from $\mathbb{R}^d$ to $\mathbb{R}^d$. We are thinking of some applications in which some nonlinear functions of $X$ play an important role. If $\phi$ is a one-to-one mapping, the Bayes probability of error remains unchanged. In particular, this is always the case if $\phi(x_1, \ldots, x_{d'})$ is such that $d' > d$, $\phi_1(x)=x_1, \ldots, \phi_{d'}(x)=x_{d'}$ and $\phi_{d'+1}, \ldots, \phi_d$ are arbitrary real-valued functions on $\mathbb{R}^d$. The theorem stated above remains valid for this interesting case.

Remark 2. (Related work on statistically equivalent blocks.) Based on the theory of statistically equivalent blocks (Anderson13) several discrimination rules can be found in the literature that also enjoy the invariance property (13). In all these rules $\mathbb{R}^d$ is partitioned into rectangles on which $Y_1$ is constant. The size and location of the rectangles is data-dependent. In the simple case ($d=1$) the real line is divided into $k$ parts by the $(1/(k-1))$-th, $(2/(k-1))$-th, $\ldots$, $(k/(k-1))$-th quantiles of $x_1, \ldots, x_n$. On each interval $g(x_{1:n})$ is constant and its value is determined by a majority vote among the $Y_i$ for which $x_i$ is in the same interval. Gordon and Olshen13 show that this simple data dependent histogram rule satisfies (ii) if $k = + \infty$ and $k/n \rightarrow 0$ as $n \rightarrow + \infty$. The experimental results reported by Gessman and Gessman11 seem to indicate that these rules are strong both where computer time and performance are concerned. Practical ways of generalizing the histogram rules for $d > 1$ and make the rectangular partition depend upon the $Y_i$ as well as the $X_i$ are proposed by Henrichon and Fu11, Meisel and Michalopoulos11, Friedman12 and Gordon and Olshen13. Return to the case $d=1$. For any new observation $X$ the decision $(Y_1)$ is based upon the approximately $n/k$ states (namely those that correspond to observations in the interval $X$ is in). In general $X$ is not the median of the observations in that interval. In contrast, with the $(n/k)$-nearest neighbor version of (6), $X$ is the median of those $x_i$'s ($X$ is included) whose states $Y_i$ are taken into consideration in (6). One might therefore expect a slightly superior performance of the rules suggested in this note.

Proof of the Theorem

Consider the regression functions of the indicators $1(Y_{1}=1)$ on $X$,

$$p_1(x) = \mathbb{E}[\mathbb{I}(Y_{1}=1)|X=x] = \mathbb{P}(Y_{1}=1|X=x), \; 1 \leq i \leq M.$$  

These functions are not uniquely defined, but it is always possible to find $p_1, \ldots, p_M$ that are Borel measurable and that satisfy

$$0 \leq p_i(x) \leq 1, \text{ all } x, \text{ and } \sum_{1 \leq i \leq M} p_i(x) = 1, \text{ all } x.$$  

Clearly, all discrimination rules of the form $Y=g(X)$ where

$$g(x) \equiv \begin{cases} 1, & \text{whenever } \quad p_1(x) < \max_{1 \leq i \leq M} p_i(x) \\ 0, & \text{otherwise} \end{cases}$$

are Bayes (i.e., they achieve $L^*$). If the unknown $p_i$ are estimated from the data by $p_{n,i}$, and if the condition (10) is replaced by

$$g_{n}(x,Z,D_n) \equiv \begin{cases} 1, & \text{whenever } \quad p_{n,1}(x,Z) < \max_{1 \leq i \leq M} p_{n,i}(x,Z) \\ 0, & \text{otherwise} \end{cases}$$

then, regardless of the estimation procedure of the $p_{n,i}$, the following inequality is valid.

Inequality 1. When $g_n$ satisfies (11) then

$$\begin{align*}
L_{n} - L^{*} & = 2 \left( \sum_{1 \leq i \leq M} \mathbb{E}[|p_1(X) - p_{n,i}(X,Z)| |D_n] \right)^2 \\
& = 2 \sum_{1 \leq i \leq M} \int |p_1(x) - p_{n,i}(x,Z)| \mu(dx) dx.
\end{align*}$$

Here $\mu$ is the probability measure of $X$.

Proof. Since $L^{*} = \mathbb{E}[\max_{1 \leq i \leq M} p_i(X)]$ and

$$L_n = \mathbb{E}[\mathbb{E}[g_n(X,Z,D_n)|D_n] |D_n] = \mathbb{E}[1-g_n(X,Z,D_n)|D_n] |D_n]$$

we have

$$\begin{align*}
L_{n} - L^{*} & = \mathbb{E}[\max_{1 \leq i \leq M} p_1(X) - \max_{1 \leq i \leq M} p_{n,i}(X,Z,D_n)(X)|D_n] \\
& = \mathbb{E}[\max_{1 \leq i \leq M} p_1(X) - \max_{1 \leq i \leq M} p_{n,i}(X,Z,D_n)(X)|D_n] \\
& = \mathbb{E}[\max_{1 \leq i \leq M} p_1(X) - \max_{1 \leq i \leq M} p_{n,i}(X,Z,D_n)(X)|D_n] \\
& = \mathbb{E}[\max_{1 \leq i \leq M} p_1(X) - \max_{1 \leq i \leq M} p_{n,i}(X,Z,D_n)(X)|D_n] \\
& \leq \sum_{1 \leq i \leq M} \mathbb{E}[|p_1(X) - p_{n,i}(X,Z)| |D_n] + 0 \\
& \leq \sum_{1 \leq i \leq M} \mathbb{E}[|p_1(X) - p_{n,i}(X,Z)| |D_n] + 0 \\
& \leq Q.E.D.
\end{align*}$$

We recall that a discrimination rule (that is, a sequence of mappings $g_n$) is weakly Bayes risk consistent if $L_{n} - L^{*} \rightarrow 0$ in probability as $n \rightarrow \infty$. From Inequality 1 we conclude that this is true whenever for all $1$

$$\mathbb{E}[|p_1(X) - p_{n,1}(X,Z)| |D_n] \rightarrow 0.$$  

By Jensen’s inequality it suffices that for all $1$

$$\mathbb{E}[|p_1(X) - p_{n,1}(X,Z)|^2 |D_n] \rightarrow 0.$$  

Now, the rules of the form (6) are obtained from (11) if one replaces $p_{n,1}(x,Z)$ by

$$\begin{align*}
\hat{Y}_{j} & = \mathbb{E}[\mathbb{I}(Y_{j}=1) |X=x] \\
& = \sum_{j:Y_j(x)=1} Y_j \mathbb{I}(Y_j(x)=1).
\end{align*}$$

Thus, if we define $Y_j = \mathbb{I}(Y_j=1)$ (where $j$ is fixed ) and if

$$m(x) = p_1(x) = \mathbb{E}[\hat{Y}_j |X=x]$$

and

$$m_n(x,z) = p_{n,1}(x,z) = \sum_{1 \leq j \leq n} \mathbb{I}(Y_j(x)=1),$$

then we need only show that

$$\mathbb{E}[|p_1(X) - p_{n,1}(X,Z)| |D_n] \rightarrow 0.$$  

\text{Q.E.D.}$$

144
\[ E(m(X) - m(X, Z))^2 \leq 0. \]  

The crucial results needed to obtain (13) for all possible distributions of \((X, Y)\) are gathered in the inequalities 2 and 3 below. Inequality 3 is essentially due to Stone7. If \(a_1, \ldots, a_n\) are elements of \(K \times [0, 1]\) then let
\[
\sum_{1 \leq i \leq n} c_i(a_1, \ldots, a_n) = v_i
\]
where \(a_i\) is the \(i\)-th nearest neighbor (in the sense explained above) to a \(a_1, \ldots, a_n\).

Inequality 2. If \(a_1, \ldots, a_n\) are different elements from \(K \times [0, 1]\) and if \(v_1 \geq \cdots \geq v_n \geq 0\) then
\[
\sum_{1 \leq i \leq n} c_i(a_1, \ldots, a_{i-1}, a_i-1, a_i+1, \ldots, a_n) \leq 2^d \sum_{1 \leq i \leq n} v_i.
\]  

Proof. Let \(a_j = a(j, 1), \ldots, a(d+1)\). Then \(a = a(j, 1), \ldots, a(d+1)\). Partition \(K \times [0, 1]\) into \(2^d\) quadrants centered at \(a\) in the following way. The quadrants are indexed by the elements of \(\{+, -\}^d\), then \(a_j\) is in the quadrant \((+, +, +, \ldots)\) if for all \(1 \leq j \leq d\)
\[
a_j(a_j) > a(j), \text{ or } a_j(a_j) = a(j) \text{ and } a_1(d+1) > a(d+1).
\]
(Thus, the least components are used to take care of the ties.) If \(Q_j\) is the \(j\)-th quadrant, we find the points \(b_1, \ldots, b_N\) from among \(a_1, \ldots, a_n\) that are in \(Q_j\). The crucial observation is that if \(b_1\) is the \(k\)-th nearest neighbor to \(a\) (from \(b_1, \ldots, b_N\)) then \(a\) is at least the \(k\)-th nearest neighbor to \(b_1\) (among \(a, b_1, \ldots, b_N\)). By view of the monotonicity condition on \(v_i\), we have
\[
\sum_{1 \leq i \leq n} c_i(a_1, a_j, \ldots, a_{i-1}, a_{i+1}, \ldots, a_n) \leq \sum_{1 \leq i \leq n} c_i(b_1, b_j, \ldots, b_{i-1}, b_{i+1}, \ldots, b_n) \leq \sum_{1 \leq i \leq n} c_i(a_1, b_j, \ldots, b_n) = \sum_{1 \leq i \leq n} v_i \leq \sum_{1 \leq i \leq n} v_i.
\]
This bound does not depend upon \(j, a_1, \ldots, a_n\). Since there are \(2^d\) quadrants, (14) obtains.

Q.E.D.

Inequality 3. Let \((X, Z, Y, (X_1, Z_1, Y_1), \ldots, (X_n, Z_n, Y_n))\) be iid random vectors from \(K \times [0, 1] \times K\) and let \(Z\) be independent of \((X, Y)\) and uniformly distributed on \([0, 1]\). If \((v_1, \ldots, v_n)\) satisfies (7) then
\[
E(|c_i(X, Z); (X_1, Z_1), \ldots, (X_n, Z_n)) Y_1 P) \leq E\sum_{1 \leq i \leq n} c_i(X, Z); (X_1, Z_1), \ldots, (X_n, Z_n)) Y_1 P) \leq 2^d E(Y_1 P), p \geq 1
\]  

Proof. The first inequality follows from Jensen's inequality. Inequality 3 now follows from Inequality 2 and the fact that \((X, Z, Y)\) and the \((X_1, Z_1, Y_1)\) are independent and identically distributed:
\[
E\left(\sum_{1 \leq i \leq n} c_i(X, Z); (X_1, Z_1), \ldots, (X_n, Z_n)) Y_1 P) \leq \sum_{1 \leq i \leq n} E(c_i(X, Z); (X_1, Z_1), \ldots, (X_n, Z_n)) Y_1 P) \leq \sum_{1 \leq i \leq n} E(c_i(X_1, Z_1); (X_1, Z_1), \ldots, (X_n, Z_n)) Y_1 P) \leq 2^d \sum_{1 \leq i \leq n} v_1 E(Y_1 P) = 2^d E(Y_1 P).
\]

Q.E.D.

Return to the proof of the Theorem. The regression function \(m\) (which is essentially bounded) can be any \(\epsilon > 0\) be approximated by a bounded continuous function \(h\) such that
\[
E|b(X) - m(x)|P_0(dx) < \epsilon.
\]
See Dunford and Schwartz16, pp. 298. By the \(c_x -\)

inequality (Loève17, pp. 155) we have
\[
\int \mid m(x,z) - m(x) \mid P_0(dx)dz \leq 4\epsilon (U_1 + U_2 + U_3 + U_4) \leq 4\epsilon (U_1 + U_2 + U_3 + U_4)
\]
where
\[
U_1 = \int \mid \sum_{1 \leq i \leq n} c_i(x, z; (X_1, Z_1), \ldots, (X_n, Z_n)) \mid P_0(dx)dz, \quad U_2 = \int \mid \sum_{1 \leq i \leq n} c_i(x, z; (X_1, Z_1), \ldots, (X_n, Z_n)) \mid P_0(dx)dz, \quad U_3 = \int \mid h(x) - m(x) \mid P_0(dx), \quad U_4 = E(\mid \sum_{1 \leq i \leq n} c_i (x, z; (Y_1 - m(x))) \mid D_n).
\]
\(U_3\) is small by choice of \(h\). Next, by Inequality 3,
\[
E(U_4) \leq 2^d E(|b(X) - h(X)|P_0),
\]
which again can be made arbitrarily small. For \(p > 2\), the term \(E(U_4)\) is estimated as follows:
\[
E(U_4) \leq \sum_{1 \leq i \leq n} E(\mid \sum_{1 \leq i \leq n} c_i (x, z; (Y_1 - m(x)))^2 \mid D_n) \leq K(p) E(\mid \sum_{1 \leq i \leq n} c_i^2 (x, z; (Y_1 - m(x)))^2 \mid D_n),
\]
\(K(p) > 0\) by the inequality of Marcinkiewicz and Zygmund18; see also Petrov19, pp. 59-60.}

\[
\leq K(p) \sum_{1 \leq i \leq n} P(\mid \sum_{1 \leq i \leq n} c_i (x, z; (Y_1 - m(x)))^2 \mid D_n) \quad \leq K(p) \sum_{1 \leq i \leq n} P(\mid \sum_{1 \leq i \leq n} c_i (x, z; (Y_1 - m(x)))^2 \mid D_n)
\]
\[ \sum_{i=1}^{d} K(p) (\sup_{i=1}^{n} |Y_{i1} - m(X_{1i})|^p) \leq 2^d K(p) (\sup_{i=1}^{n} |Y_{i1}|^p/2) \]

Inequality 3)

\[ \leq 2^d K(p) (\sup_{i=1}^{n} |Y_{i1}|^p/2) \]

(since $Y_{i1}$ is $\{0,1\}$-valued, and thus $|Y_{i1}|^p \leq 1$ for almost all $x_i$.)

Thus, $E(Y_{i1}) \leq 0$ from (8). We need only show that $E(U_{i1})$ tends to 0 for all bounded continuous functions $h$. If we define

\[ W_n = \sum_{1 \leq i \leq n} C_{ni}(x,z) h(X_{1i} - h(x)) \]

then we need only show that $\int W_n(x,z) \mu(dx) dz \leq 0$ in probability. Actually, we will show more:

Inequality 4. Let $h$ be a bounded continuous function on $\mathbb{R}^d$ and let $\{v_1,\ldots,v_n\}$ be a probability vector satisfying (9). For every $\epsilon > 0$, $p > 1$ there exist $a_1, a_2, \ldots, a_n > 0$ such that

\[ P(\sum_{1 \leq i \leq n} |v_i| \leq \epsilon) \leq e^{-a_1 \epsilon} + e^{-a_2 \epsilon}, \quad a_1 > a_2 > 0. \]

Proof. Define the following ordering on $Rx[0,1]^d$:

\[ (x_1, z_1) \prec (x_2, z_2) \text{ if } x_1 < x_2 \text{ or } x_1 = x_2, z_2 < z_1. \]

The relation $(x_1, z_1) \prec (x_2, z_2)$ also includes the case $x_1 = x_2, z_1 = z_2$. For any sequence from $Rx[0,1]^d$, say $(z_0, z_1, \ldots, z_n)$, the set $B_j$ from $B_X[0,1]$ is defined as the set of $(x_1, z_1, \ldots, x_d, z_d)$ with the property that

\[ (x_{j-1}, z_{j-1}) \prec (x_j, z_j). \]

For a given number $c \in (0,1)$ and for each $j = 1, \ldots, d$, we find numbers $(x_0, z_0, \ldots, x_n, z_n)$ with the property that

\[ P((X,Z) \in B_j) = c/n, \quad \text{all } i=1, \ldots, n. \]

Thus, if

\[ B_j = \bigcap_{i=1}^{n} B_{j_i}, \quad B = \bigcap_{j=1}^{d} B_j, \]

then it follows that $P((X,Z) \in B) = c$.

In the proof we will group all "bad" cylinder sets $B_{j_1}$ together in a set $B^*$. To start with, we include the border strips $B_{j_1}, B_{j_2}, \ldots, B_{j_N}$ for all $j$. We also include all the $B_{j_1}$ whose length (that is, $x_{j-1} - x_{j-1}$) exceeds $\delta > 0$. Notice that for fixed $j$ the number of strips $B_{j_1}$ with length greater than or equal to $\delta$ can not exceed $T/\delta + 1$ where

\[ T = \max_{1 \leq j \leq d} (x_{j} - x_0). \]

In addition we include $B_{j_1}, B_{j_2}, B_{j_3}$ and $B_{j_4}$ whenever $2 \leq k \leq d$ and $B_{j_4}$ already is in $B^*$, that is, for every strip whose length exceeds $\delta$, we include in $B^*$ the four neighbouring strips as well. Repeating this for all $j$, we obtain

\[ P((X,Z) \in B^*) \leq d(4^d + 5^d + 5^d/2)(c/n). \]

A rectangle $C$ is defined as a set of the form

\[ \bigcup_{j=1}^{d} B_{j_1}, \quad (1,1,2, \ldots, 1) \subseteq (1, \ldots, d). \]

Pick $k$ as in (9) and construct a set $B^*_k$ by taking the union of all the rectangles $C$ for which $P((X,Z) \in C) \leq 0 + k/n$ where $\delta > 0$ is a number to be determined further on. Clearly,

\[ P((X,Z) \in B^*_k) \leq \sum_{C \in \text{rectangles } C} P((X,Z) \in B^*_k) \leq n^d (\delta + k/n). \]

If $B^* \supseteq \sup h(x)$ and $(C)$ denotes the complement of a set, then

\[ \int W(x) \mu(dx) dz \leq (2H)^d P((X,Z) \in B^*_k) \leq 0 + \sum_{C \in \text{rectangles } C} P((X,Z) \in B^*_k) \]

The first term on the right hand side of (19) can be estimated by

\[ (2H)^d d(1-c) + d(9^d + 5^d + 5^d/2)(c/n + 2N^d(6^d + k/n)) \]

where we used the fact that $P((X,Z) \in B) \leq d(1-c)$.

Now, (20) is smaller than $\epsilon/2$ for all $n$ large enough. First pick $c$ close enough to 1 to take care of $d(1-c)$. This fixes the value of $\delta$. If $\delta$ is given, then we can pick $N$ large enough to make the second term as small as desired. Finally, choose $\delta$ so small that $N^d \delta$ is small, and use (9) to conclude that $N^d \delta c$ as $n \to \infty$. The only parameter that can still be assigned an arbitrary value is $\delta$.

The measure induced by the Borel sets of $\mathbb{R}^d$ is called $\mu$. The empirical measure $v_n$ of a Borel set $S \subseteq \mathbb{R}^d$ is defined by

\[ v_n(S) = \frac{1}{n} \sum_{1 \leq i \leq n} I((X_i, Z_i) \in S). \]

Let $E_1$ be the event that for all $(n^d)$ rectangles $C$, $|\nu(C) - v_n(C)| < \epsilon$, and that for all $(n)$ cylinder sets $B_j$ $|\nu(B_j) - v_n(B_j)| < c/3N$.

Let us fix $(x,z) \in C \cap (x_{1j_1}, z_{1j_1}) \subseteq (x_{1j_2}, z_{1j_2})$. In that case the empirical distance between $(x,z)$ and any of the $(1,1,2, \ldots, 1) \in C$ is less than $n (c/N + c/3N)$ because by definition $\nu(B_{j_1}) = c/n$ for all $j_1, j_2$. Next, if $C = B_{1j_1} B_{2j_2} \cdots B_{dj_d}$ and $c = 4^d + 5^d + 5^d/2$, then its empirical distance to $(x,z)$
is at least equal to
\[ 2n \left( \frac{c}{N} - \frac{c}{3N} \right) = n \left( \frac{c}{N} + \frac{c}{3N} \right). \]

Therefore, all the \((X_i, Z_i)\) in \(C\) are nearer neighbors of \((x, z) \in \mathbb{C}\) than any \((X_j, Z_j) \in \mathbb{C}'\). But if \(E_1\) holds, then
\[ v_n(C) > 0 \geq k/n + \theta = k/n, \]
that is, more than \(k\) of the \((X_i, Z_i)\) belong to \(C\).

This in turn implies that all the \(k\) nearest neighbors to \((x, z)\) are in
\[ \bigcap_{j=1}^{d} \left( B_{ij} \cup B_{ij+1} \right). \]

By the construction of \(B^*\) it is true that under the maximum component norm \(\|\cdot\|\),
\[ |X_i(x, z) - x| < 3\delta, 1 \leq i \leq k. \]

Since we can still pick \(\delta\), we pick it so small that
\[ \sup_{(x, z) \in B; (y, z) \in B} \frac{|h(x) - h(y)|}{|x - y|} < \varepsilon/4, \]
which is possible by the uniform continuity of \(h\) on the closure of the projection of \(B\) on \(\mathbb{R}^2\). Thus, if \(E_2\) holds,
\[ \sup_{(x, z) \in B_n \cap \mathbb{N} \cap B^*} \sum W(x, z) \leq (\varepsilon/4) \sum \nu_i + p(2h)^p \sum \nu_i \leq \varepsilon/2 \text{ for all } n \text{ large enough.} \]

From (19) and (21),
\[ P \left( \sum W(x, z) u(dx) dz \geq \varepsilon \right) \leq 0 \]
\[ \leq P \left( \sum_{B \in B_n \cap \mathbb{N} \cap B^*} \sum W(x, z) u(dx) dz \geq \varepsilon/2 \right) \]
\[ \leq P(\varepsilon^2) \leq \sum_{\text{rectangles } C} P (\nu(C) - \nu(C) > 0) \]
\[ + \sum_{j=1}^{\infty} P (\nu(B_{ij}) > n(j))^2 \leq \frac{c}{3N} \]
\[ \leq 2N^2 e^{-2n\theta^2} + 2N d e^{-2nc^2/3N^2} \]
where we used Hoeffding's inequality for the sum of independent \(0,1\)-valued random variables. This concludes the proof of Inequality 4.

Q. E. D.

References


*Research of author partially sponsored by AFSOR Grant 77-3385.