*The Annals of Statistics*
1994, Vol. 22, No. 1, 460–473

# UNIVERSALLY CONSISTENT CONDITIONAL $U$-STATISTICS[1]

BY WINFRIED STUTE

*Universität Giessen*

We introduce a general class of conditional $U$-statistics and present sufficient conditions for their universal consistency in $r$th mean. It is shown that under mild assumptions on the smoothing parameters, window and $k_n$–nearest neighbor estimators are universally consistent. An application to a new nonparametric discrimination problem is also included.

**1. Introduction and main results.** This paper constitutes a continuation of work on so-called conditional (local) $U$-statistics as introduced in Stute (1991). These statistics may be viewed as generalizations of regression function estimates, similar to Hoeffding's (1948) extension of a sample mean to what is now called a $U$-statistic. To be specific, assume that $(X_i, Y_i)$, $1 \le i \le n$, are i.i.d. random vectors in some Euclidean space $\mathbb{R}^{d+s}$, defined on some probability space $(\Omega, \mathcal{A}, \mathbb{P})$. As in the regression case the $X$ and $Y$ may be considered as input and output variables, respectively, the stochastic dependence of which being described by the pertaining regression function. In order to measure the impact of a few $X$'s, say $(X_1, \dots, X_k)$, on a function $h(Y_1, \dots, Y_k)$ of the pertaining $Y$'s, set

$$m(\mathbf{x}) \equiv m(x_1, \dots, x_k) := \mathbb{E}\left[h(Y_1, \dots, Y_k) \mid X_1 = x_1, \dots, X_k = x_k\right].$$

Here $h$ is a given real-valued function on $\mathbb{R}^{s \times k}$ (the $U$-kernel) such that, for some $r \ge 1, h(Y_1, \dots, Y_k) \in \mathcal{L}_r$, the space of all random variables $Z$ for which $|Z|^r$ is integrable; $k$ is called the degree of $h$ (resp. $m$). Stute (1991) contains several examples, and another example will be discussed in greater detail in Section 3.

In this paper we study a fairly general class of conditional $U$-statistics of the form

$$m_n(\mathbf{x}) = \sum_{\pi} W_{\pi n}(\mathbf{x}) h(Y_\pi), \tag{1.1}$$

designed to estimate $m(\mathbf{x})$. Summation in (1.1) takes place over all permutations $\pi = (\pi_1, \dots, \pi_k)$ of length $k$ such that $1 \le \pi_i \le n$. Of course, $Y_\pi = (Y_{\pi_1}, \dots, Y_{\pi_k})$. We shall also write $X_\pi = (X_{\pi_1}, \dots, X_{\pi_k})$ whenever convenient.

In order to make $m_n(\mathbf{x})$ a local average, $W_{\pi n}(\mathbf{x})$ has to give larger weights to those $h(Y_\pi)$ for which $X_\pi$ is close to $\mathbf{x}$. For detailed conditions, see (i)–(v) below.

Received March 1992; revised February 1993.

[1] Work done while on leave at the University of California, Davis. Partially supported by the Deutsche Forschungsgemeinschaft.

*AMS* 1991 *subject classifications.* Primary 62G05; secondary 62G07.

*Key words and phrases.* Conditional $U$-statistics, universal consistency, window and NN weights, nonparametric discrimination.

While in our previous paper pointwise consistency and asymptotic normality of $m_n$ were investigated, the present paper deals with consistency in $r$th mean. For the regression case this was first done in a seminal paper by Stone (1977). According to his suggestions we shall call $\{W_{\pi n}\}$ *universally consistent* iff

$$m_n(\mathbf{X}) \to m(\mathbf{X}) \quad \text{in } \mathcal{L}_r$$

under no conditions on $h$ (up to integrability) or the distribution of $(X, Y)$. Here $\mathbf{X} = (X_1^0, \dots, X_k^0)$ is a vector of $X$'s with the same distribution as $(X_1, \dots, X_k)$ and being independent of $(X_i, Y_i), i \geq 1$. In the discrimination setup of Section 3, $(X_i, Y_i), 1 \leq i \leq n$, will be a training sample and $\mathbf{X}$ is a vector to be classified.

Theorem 1.1 presents sufficient conditions for universal consistency. In Theorems 1.2 and 1.3 it is shown that window weights and $k_n$–nearest neighbor (NN) weights are universally consistent under mild assumptions on the smoothing parameters.

In what follows $\|\cdot\|$ will be the maximum norm on $\mathbb{R}^d$. We also set

$$\|X_\pi - \mathbf{x}\| := \max_{1 \leq i \leq k} \|X_{\pi_i} - x_i\|.$$

The fact that $\|\cdot\|$ will be used to measure the distance between $X_\pi$ and $\mathbf{x}$ as well as the distance between points in $\mathbb{R}^d$ will not cause any confusion; $\mu$ denotes the distribution of each $X_i$. Finally, without further mentioning, we shall omit the index $n$ from $W_{\pi n}$.

Consider the following set of assumptions [cf. conditions (1)–(5) in Stone (1977)]:

(i) There exists some finite $C \geq 1$ such that, for each $f \geq 0$ and every $n \geq n_0$, say,

$$\mathbb{E}\left[\sum_\pi |W_\pi(\mathbf{X})| f(X_\pi)\right] \leq C\mathbb{E}f(\mathbf{X}).$$

(ii) For some $D \geq 1$,

$$\mathbb{P}\left(\sum_\pi |W_\pi(\mathbf{X})| \leq D\right) = 1.$$

As $n \to \infty$,

(iii) $$\sum_\pi |W_\pi(\mathbf{X})| 1_{\{\|X_\pi - \mathbf{x}\| > \varepsilon\}} \to 0 \quad \text{in probability for each } \varepsilon > 0,$$

(iv) $$\sum_\pi W_\pi(\mathbf{X}) \to 1 \quad \text{in probability},$$

(v) $$\sum_{\pi, \sigma}{}^{(q)} |W_\pi(\mathbf{X}) W_\sigma(\mathbf{X})| \to 0 \quad \text{in probability},$$

for each $1 \leq q \leq k$, where $\Sigma^{(q)}$ denotes summation over all permutations $\pi, \sigma$ which have $q$ indices in common (not necessarily at the same position).

The constants $C$ and $D$ need not be known and may depend on the underlying distributions.

We now state our main result. It constitutes an extension to the $U$-statistic setup of Stone [(1977), Theorem 1].

THEOREM 1.1. *Assume that* $h(Y_1, \ldots, Y_k) \in \mathcal{L}_r$. *Then, under* (i)–(v), *we have*

$$m_n(\mathbf{X}) \to m(\mathbf{X}) \quad \text{in } \mathcal{L}_r,$$

*that is,*

$$\mathbb{E}\left[\int |m_n(\mathbf{x}) - m(\mathbf{x})|^r \mu(dx_1) \cdots \mu(dx_k)\right] \to 0.$$

Theorems 1.2 and 1.3 deal with two special cases: window weights and NN-weights. Consistency of window estimates for the regression function has been obtained by Devroye and Wagner (1980) and Spiegelman and Sacks (1980). NN-weights for the regression function have been studied in Stone [(1977), Theorem 2].

To define the window weights, put

$$W_\pi(\mathbf{x}) = \begin{cases} 1_{\{\|X_\pi - \mathbf{x}\| \leq h_n\}} \Big/ \sum_\sigma 1_{\{\|X_\sigma - \mathbf{x}\| \leq h_n\}}, & \text{if well-defined,} \\ 0, & \text{otherwise.} \end{cases}$$

Here $h_n > 0$ is a given window size to be chosen by the statistician.

THEOREM 1.2. *Assume* $h_n \to 0$ *and* $nh_n^d \to \infty$ *as* $n \to \infty$. *Then*

$$m_n(\mathbf{X}) \to m(\mathbf{X}) \quad \text{in } \mathcal{L}_r.$$

For the NN-weights recall that $X_j$ is among the $k_n$-NN of $x \in \mathbb{R}^d$ iff $d_j(x) := \|X_j - x\|$ is among the $k_n$-smallest ordered values $d_{1:n}(x) \leq \cdots \leq d_{n:n}(x)$ of the $d$'s. Ties may be broken by randomization.

For a given $1 \leq k_n \leq n$, set

$$W_\pi(\mathbf{x}) = \begin{cases} k_n^{-d}, & \text{if } X_{\pi_i} \text{ is among the } k_n\text{-NN of } x_i \text{ for } 1 \leq i \leq k, \\ 0, & \text{otherwise.} \end{cases}$$

THEOREM 1.3. *Assume* $k_n \to \infty$ *and* $k_n/n \to 0$ *as* $n \to \infty$. *Then*

$$m_n(\mathbf{X}) \to m(\mathbf{X}) \quad \text{in } \mathcal{L}_r.$$

Theorems 1.2 and 1.3 will be proved by verifying conditions (i)–(v) in Theorem 1.1.

**2. Proofs.** To prove Theorem 1.1, a few lemmas are needed.

LEMMA 2.1. *Under* (i)–(iii), *we have, for each $f$ such that $f \circ \mathbf{X} \in \mathcal{L}_r$,*

$$\mathbb{E}\left[\sum_{\pi} |W_\pi(\mathbf{X})|\, |f(X_\pi) - f(\mathbf{X})|^r\right] \to 0.$$

PROOF. For $\varepsilon > 0$ given, choose a continuous $g$ on $\mathbb{R}^{d\times k}$ with compact support such that

$$\mathbb{E}|f(\mathbf{X}) - g(\mathbf{X})|^r \le \varepsilon.$$

By (i), assuming $n_0 = 1$ w.l.o.g.,

$$\mathbb{E}\left[\sum_{\pi} |W_\pi(\mathbf{X})|\, |f(X_\pi) - g(X_\pi)|^r\right] \le C\,\mathbb{E}|f(\mathbf{X}) - g(\mathbf{X})|^r \le C\varepsilon.$$

From (ii),

$$\mathbb{E}\left[\sum_{\pi} |W_\pi(\mathbf{X})|\, |f(\mathbf{X}) - g(\mathbf{X})|^r\right] \le D\varepsilon.$$

Altogether this shows that we need to prove the lemma only for continuous $f$ with compact support. Set $M = \|f\|_\infty$, the sup-norm of $f$. Since $f$ is uniformly continuous, for a given $\varepsilon > 0$, we may find some $\delta > 0$ such that

$$\|\mathbf{x} - \mathbf{x}_1\| \le \delta \quad \text{implies} \quad |f(\mathbf{x}) - f(\mathbf{x}_1)|^r \le \varepsilon.$$

By (ii)

$$\mathbb{E}\left[\sum_{\pi} |W_\pi(\mathbf{X})|\, |f(X_\pi) - f(\mathbf{X})|^r\right] \le (2M)^r\,\mathbb{E}\left[\sum_{\pi} |W_\pi(\mathbf{X})|\, 1_{\{\|X_\pi - \mathbf{X}\| > \delta\}}\right] + \varepsilon D.$$

Use (ii) and (iii) to conclude that the last expectation converges to zero. This completes the proof of the lemma. □

LEMMA 2.2. *Under* (i)–(iv), *for each $f$ such that $f \circ \mathbf{X} \in \mathcal{L}_r$, we have*

$$\sum_{\pi} W_\pi(\mathbf{X}) f(X_\pi) \to f(\mathbf{X}) \quad \textit{in r-th mean.}$$

PROOF. By (ii),

$$\left|\sum_{\pi} W_\pi(\mathbf{X}) - 1\right|^r \le (1 + D)^r \quad \text{with probability 1.}$$

Thus, by (iv),

$$\mathbb{E}\left|\left[\sum_{\pi} W_{\pi}(\mathbf{X}) - 1\right] f(\mathbf{X})\right|^{r} \to 0. \tag{2.1}$$

Further, apply (ii), Lemma 2.1 and Hölder's inequality to get

$$\mathbb{E}\left|\sum_{\pi} W_{\pi}(\mathbf{X})\left[f(X_{\pi}) - f(\mathbf{X})\right]\right|^{r} \to 0. \tag{2.2}$$

Clearly, (2.1) and (2.2) imply the assertion of the lemma. □

Lemmas 2.1 and 2.2 now readily yield a proof of Theorem 1.1.

PROOF OF THEOREM 1.1. For this note that since $\mathbb{E}|h(Y_1, \dots, Y_k)|^r < \infty$ by assumption, Jensen's inequality implies

$$\begin{aligned}\mathbb{E}\left[|m(\mathbf{X})|^{r}\right] &= \mathbb{E}\left[\left|\mathbb{E}\left[h(Y_1, \dots, Y_k) \mid X_1, \dots, X_k\right]\right|^{r}\right] \\ &\leq \mathbb{E}\left|h(Y_1, \dots, Y_k)\right|^{r} < \infty.\end{aligned}$$

Now,

$$\begin{aligned}&\sum_{\pi} W_{\pi}(\mathbf{X}) h(Y_{\pi_1}, \dots, Y_{\pi_k}) - m(\mathbf{X}) \\ &\quad = \sum_{\pi} W_{\pi}(\mathbf{X}) m(X_{\pi}) - m(\mathbf{X}) + \sum_{\pi} W_{\pi}(\mathbf{X}) Z_{\pi},\end{aligned} \tag{2.3}$$

where

$$\begin{aligned}Z_{\pi} &= h(Y_{\pi_1}, \dots, Y_{\pi_k}) - \mathbb{E}\left[h(Y_{\pi_1}, \dots, Y_{\pi_k}) | X_{\pi}\right] \\ &= h(Y_{\pi_1}, \dots, Y_{\pi_k}) - m(X_{\pi}).\end{aligned}$$

The first term in (2.3) converges to zero in $r$th mean according to Lemma 2.2. For the second sum we first treat the case $r = 2$. We then have

$$\mathbb{E}\left[\sum_{\pi} W_{\pi}(\mathbf{X}) Z_{\pi}\right]^{2} = \sum_{\pi, \sigma} \mathbb{E}\left[W_{\pi}(\mathbf{X}) W_{\sigma}(\mathbf{X}) Z_{\pi} Z_{\sigma}\right]. \tag{2.4}$$

It is easily seen that the expectation vanishes if $\pi$ and $\sigma$ do not have any indices in common. So, fix $1 \leq q \leq k$, with $q$ indicating the number of indices appearing in both $\pi$ and $\sigma$. Then

$$\begin{aligned}\mathbb{E}\left[W_{\pi}(\mathbf{X}) W_{\sigma}(\mathbf{X}) Z_{\pi} Z_{\sigma}\right] &= \mathbb{E}\left[W_{\pi}(\mathbf{X}) W_{\sigma}(\mathbf{X}) \mathbb{E}\left[Z_{\pi} Z_{\sigma} \mid X_{\pi}, X_{\sigma}, \mathbf{X}\right]\right] \\ &= \mathbb{E}\left[W_{\pi}(\mathbf{X}) W_{\sigma}(\mathbf{X}) \mathbb{E}\left[Z_{\pi} Z_{\sigma} \mid X_{\pi}, X_{\sigma}\right]\right],\end{aligned}$$

by independence. Furthermore, by Cauchy–Schwarz and independence,

$$\begin{aligned}\big|\mathbb{E}[Z_\pi Z_\sigma|X_\pi, X_\sigma]\big| &\le \sqrt{\mathbb{E}[Z_\pi^2|X_\pi]\mathbb{E}[Z_\sigma^2|X_\sigma]}\\ &\equiv \gamma(X_\pi)\gamma(X_\sigma), \quad \text{say.}\end{aligned}$$

Assume for a moment that $\gamma$ is bounded. Condition (v) then guarantees that

$$\sum_{\pi,\sigma}{}^{(q)}|W_\pi(\mathbf{X})W_\sigma(\mathbf{X})|\gamma(X_\pi)\gamma(X_\sigma) \to 0 \quad \text{in probability}$$

and therefore, by (ii), also in $\mathcal{L}_2$. The case of a general (integrable) $\gamma$ is easily traced back to a bounded $\gamma$. For this, write

$$\gamma = \gamma 1_{\{|\gamma|\le M\}} + \gamma 1_{\{|\gamma|>M\}} \equiv \gamma_1 + \gamma_2.$$

Hence

$$\begin{aligned}\gamma(X_\pi)\gamma(X_\sigma) = {}&\gamma_1(X_\pi)\gamma_1(X_\sigma) + \gamma_1(X_\pi)\gamma_2(X_\sigma)\\ &+ \gamma_2(X_\pi)\gamma_1(X_\sigma) + \gamma_2(X_\pi)\gamma_2(X_\sigma).\end{aligned}$$

Since $\gamma_1$ is bounded, the sum over the first terms need not be dealt with again. As to the second group, say, we have

$$\begin{aligned}&\sum_{\pi,\sigma}{}^{(q)}\ \mathbb{E}\Big[|W_\pi(\mathbf{X})W_\sigma(\mathbf{X}) \mid \gamma_1(X_\pi)\gamma_2(X_\sigma)\Big]\\ &\quad\le \mathbb{E}\left\{\left[\sum_\pi |W_\pi(\mathbf{X})|\gamma_1(X_\pi)\right]\left[\sum_\sigma |W_\sigma(\mathbf{X})|\gamma_2(X_\sigma)\right]\right\}\\ &\quad\le \mathbb{E}^{1/2}\left[\sum_\pi |W_\pi(\mathbf{X})|\gamma_1(X_\pi)\right]^2 \mathbb{E}^{1/2}\left[\sum_\pi |W_\pi(\mathbf{X})|\gamma_2(X_\pi)\right]^2\\ &\quad\le D\mathbb{E}^{1/2}\left[\sum_\pi |W_\pi(\mathbf{X})|\gamma_1^2(X_\pi)\right]\mathbb{E}^{1/2}\left[\sum_\pi |W_\pi(\mathbf{X})|\gamma_2^2(X_\pi)\right].\end{aligned}$$

By (i), the last term may be bounded from above by $DC\mathbb{E}^{1/2}\gamma_1^2(\mathbf{X})\mathbb{E}^{1/2}\gamma_2^2(\mathbf{X})$. However, $|\gamma_1| \le \gamma$ and

$$\mathbb{E}\gamma^2(\mathbf{X}) = \mathbb{E}\big[h(Y_1, \ldots, Y_k) - m(X_1, \ldots, X_k)\big]^2 < \infty.$$

So we may choose $M$ so large that $\mathbb{E}\gamma_2^2(\mathbf{X}) \le \varepsilon$, where $\varepsilon > 0$ is any given positive number. In summary, we have found an upper bound for (2.4) which tends to zero. However, (2.4) is nonnegative, and the theorem is proved for $r = 2$. For a general $r \ge 1$, let $M \ge 0$ and write

$$h = 1_{\{|h|\le M\}}h + 1_{\{|h|>M\}}h \equiv h_1 + h_2.$$

Denote with $m_1$ and $m_2$ the corresponding regression function, so that

$$m = m_1 + m_2.$$

Then

$$\mathbb{E}|m_2(\mathbf{X})|^r \le \mathbb{E}|h_2(Y_1,\dots,Y_k)|^r \to 0 \quad \text{as } M \to \infty,$$

as well as, by (i) and (ii), upon applying Hölder's inequality,

$$\mathbb{E}\left|\sum_\pi W_\pi(\mathbf{X})h_2(Y_\pi)\right|^r \to 0 \quad \text{as } M \to \infty,$$

uniformly in $n$. So it suffices to deal with bounded $h$'s. For such an $h$, we have already shown convergence in $\mathcal{L}_2$. Consequently, we obtain convergence in probability and therefore, by boundedness, convergence in $\mathcal{L}_r$, for any $r \ge 1$. □

The following lemma provides a useful sufficient criterion for (i) to the effect that (i) only needs to be verified for tensor products of $d$-variate functions.

LEMMA 2.3. *Assume that* (i) *is satisfied for all nonnegative $f$ of the form*

$$f(x_1,\dots,x_k) = \prod_{i=1}^k f_i(x_i).$$

*Then* (i) *is fulfilled in general.*

PROOF. Since both sides in (i) are linear in $f$, (i) also holds for (nonnegative) finite linear combinations of tensor products. Further, to each (integrable) $f \ge 0$, there exists a sequence $(f_r)_r$ of such functions with

$$f_r \to f \quad \mu\otimes\cdots\otimes\mu\text{-almost surely and in the mean.}$$

Now apply Fatou's lemma to get

$$\begin{aligned}\mathbb{E}\left[\sum_\pi |W_\pi(\mathbf{X})| f(X_\pi)\right] &= \mathbb{E}\left[\liminf_{r\to\infty}\sum_\pi |W_\pi(\mathbf{X})| f_r(X_\pi)\right]\\ &\le \liminf_{r\to\infty}\mathbb{E}\left[\sum_\pi |W_\pi(\mathbf{X})| f_r(X_\pi)\right]\\ &\le \liminf_{r\to\infty} C\,\mathbb{E}f_r(\mathbf{X}) = C\,\mathbb{E}f(\mathbf{X}).\end{aligned}$$

□

Next we show that conditions (i)–(v) are satisfied for the window estimator, thus proving Theorem 1.2. According to Lemma 2.3, for (i), only $f$ of the type

$$f(x_1,\dots,x_k) = f_1(x_1)\cdots f_k(x_k)$$

need to be considered. For $x \in \mathbb{R}^d$ and $h > 0$ we denote with $B(x;h)$ the closed ball with center $x$ and radius $h$. $A^c$ is the complement of a set $A$.

LEMMA 2.4. *There exists some finite constant $C = C(k,d,\mu)$ such that, for each nonnegative tensor product $f = \otimes_{j=1}^{k} f_j$,*

$$\mathbb{E}\left[\sum_{\pi} W_{\pi}(\mathbf{X}) f(X_{\pi})\right] \leq C\,\mathbb{E}[f(\mathbf{X})],$$

*for all $n \geq n_0$, say.*

PROOF. Fix some $\pi = (\pi_1, \dots, \pi_k)$. Recall that $\mathbf{X} = (X_1^0, \dots, X_k^0)$. By independence, the conditional expectation

$$\mathbb{E}\left[W_{\pi}(\mathbf{X}) f(X_{\pi}) \mid \mathbf{X}, X_i, 1 \leq i \leq n, i \neq \pi_1, \dots, \pi_k, 1_{\{\|X_{\pi_l} - X_j^0\| \leq h_n\}}, \qquad 1 \leq j, l \leq k\right]$$

equals

$$W_{\pi}(\mathbf{X}) \mathbb{E}\left[f(X_{\pi}) \mid \mathbf{X}, 1_{\{\|X_{\pi_l} - X_j^0\| \leq h_n\}}, \qquad 1 \leq j, l \leq k\right]$$

Given $\mathbf{X} = (x_1, \dots, x_k)$, the last conditional expectation is an elementary function which is constant on each set $\{X_{\pi_1} \in A(x_1), \dots, X_{\pi_k} \in A(x_k)\}$, where $A(x_j)$ is either $B(x_j; h_n)$ or $B^c(x_j; h_n)$. The attained value

$$\prod_{j=1}^{k} \frac{1}{\mathbb{P}(X_1 \in A(x_j))} \int_{\{X_1 \in A(x_j)\}} f_j(X_1)\, d\mathbb{P}$$

does not depend on $\pi$. Since $\sum_{\pi} W_{\pi} \leq 1$ and $W_{\pi}(\mathbf{X})$ vanishes if $\|X_{\pi_j} - x_j\| > h_n$ for at least one $1 \leq j \leq k$, we obtain

$$\begin{aligned}
&\mathbb{E}\left[\sum_{\pi} W_{\pi}(\mathbf{X}) f(X_{\pi})\right] \\
&\quad \leq \sum_{A = B, B^c} \int \prod_{j=1}^{k} \frac{1}{\mathbb{P}(X_1 \in A(x_j))} \int_{\{X_1 \in A(x_j)\}} f_j(X_1)\, d\mathbb{P}\mu(dx_1) \cdots \mu(dx_k).
\end{aligned}$$

The conclusion follows from Lemma 2.5. □

LEMMA 2.5. *There exists a finite constant $C$ such that, for each $x \in \mathbb{R}^d$ and all small enough $h > 0$,*

(i) $\displaystyle\int_{B(x;h)} \frac{\mu(dy)}{\mu(B(y;h))} \leq C,$

(ii) $\displaystyle\int_{B^c(x;h)} \frac{\mu(dy)}{\mu(B^c(y;h))} \leq C.$

PROOF. Part (i) follows from Lemma 2 in Spiegelman and Sacks (1980). Part (ii) is trivial if $\mu$ has continuous marginals or, more generally, at least one marginal is not purely discrete. If $\mu$ is purely discrete with masses $\alpha_i$ at $z_i$, then each integral in (ii) is bounded from above by

$$\sum_i \frac{\alpha_i}{1-\alpha_i} < \infty. \qquad \square$$

As already noted,

$$\sum_\pi |W_\pi(\mathbf{X})| \leq 1,$$

whence (ii). Also, for $h_n < \varepsilon$,

$$1_{\{\|X_\pi - \mathbf{x}\| \leq h_n\}} 1_{\{\|X_\pi - \mathbf{x}\| > \varepsilon\}} = 0,$$

proving (iii). Conditions (iv) and (v) will follow from the next lemma.

LEMMA 2.6. *Assuming $h_n \to 0$ and $nh_n^d \to \infty$, we have*

$$n\,\mathbb{P}(\|X_1 - X_1^0\| \leq h_n) \to \infty.$$

PROOF. From Corollary (10.50) in Wheeden and Zygmund (1977),

$$\lim_{n\to\infty} h_n^{-d}\mathbb{P}(\|X_1 - x\| \leq h_n) = :\gamma(x)$$

exists for $\mu$-almost all $x \in \mathbb{R}^d$ and is positive, possibly infinite. Hence

$$\lim_{n\to\infty} n\mathbb{P}(\|X_1 - x\| \leq h_n) = \infty \quad \mu\text{-almost surely.} \tag{2.5}$$

Integrate out to obtain the assertion of the lemma. □

LEMMA 2.7. *Under $h_n \to 0$ and $nh_n^d \to \infty$, conditions* (iv) *and* (v) *are satisfied.*

PROOF. Write, for $\mathbf{x} = (x_1, \dots, x_k) \in \mathbb{R}^{d\times k}$,

$$S_n(\mathbf{x}) = \frac{(n-k)!}{n!} \sum_\pi \frac{1_{\{\|X_\pi - \mathbf{x}\| \leq h_n\}}}{\mathbb{P}(\|X_\pi - \mathbf{x}\| \leq h_n)}.$$

Note that the denominator does not depend on $\pi$ and equals

$$\prod_{i=1}^k \mathbb{P}(\|X_1 - x_i\| \leq h_n) = :V_n.$$

A straightforward extension to dimension $d$ of the arguments utilized in the proof of Theorem 2 in Stute (1991) to handle $B_{n2}$ there yields, for $\mu \otimes \cdots \otimes \mu$-almost all $\mathbf{x}$,

$$S_n(\mathbf{x}) \to 1 \quad \text{in probability as } n \to \infty. \tag{2.6}$$

Together with Lemma 2.6 this implies

$$\sum_{\pi} 1_{\{\|X_\pi - \mathbf{x}\| \le h_n\}} \to \infty \quad \text{in probability.}$$

Integrate out to obtain

$$\sum_{\pi} 1_{\{\|X_\pi - \mathbf{X}\| \le h_n\}} \to \infty \quad \text{in probability}$$

and consequently (iv). To verify (v), due to (2.6), it suffices to show that in probability

$$(2.7) \qquad V_n^{-2}\left[\frac{(n-k)!}{n!}\right]^2 \sum_{\pi,\sigma}^{(q)} 1_{\{\|X_\pi - \mathbf{x}\| \le h_n, \|X_\sigma - \mathbf{x}\| \le h_n\}} \to 0.$$

For this, assume first that $x_1, \dots, x_k$ are pairwise distinct. Hence

$$\varepsilon \colon = \min_{i \ne r} \|x_i - x_r\| > 0.$$

From some $n_0$ on, we have $h_n < \varepsilon/2$. It follows that $\|X_j - x_i\| \le h_n$ for at most one of the $x$'s, $1 \le j \le n$. Denote with $R_n(i)$ the number of data points $X_j$ satisfying $\|X_j - x_i\| \le h_n$. As for (2.6), we obtain in probability

$$\frac{R_n(i)}{n\mathbb{P}(\|X_1 - x_i\| \le h_n)} \to 1, \qquad 1 \le i \le k.$$

Since $q \ge 1$, the sum in (2.7) is therefore, for $n \ge n_0$, bounded from above by

$$O_{\mathbb{P}}(1) \max_{1 \le i \le k} \left[n\mathbb{P}(\|X_1 - x_i\| \le h_n)\right]^{-1}$$

Apply (2.5) to get (2.7).

If $\mu$ has no atoms, the above reasoning applies to $\mu \otimes \cdots \otimes \mu$-almost all $\mathbf{x}$, proving (v). Ties among the $x_i$ can only occur (up to a null set) if they are atoms. The proof is similar to before. Actually, it simplifies a bit since the corresponding factors in $V_n$ are bounded away from zero so that, for these $x_i$, application of the SLLN rather than a differentiation argument suffices. □

Lemmas 2.4–2.7 establish the proof of Theorem 1.2.

The following lemma is well known and is stated here just for the sake of reference. It will be needed for the proof of Theorem 1.3. Denote by supp($\mu$) the support of $\mu$. Recall $d_{k:n}(x)$.

LEMMA 2.8. *For each $x \in$ supp($\mu$) we have, under $k_n/n \to 0$,*

$$d_{k_n:n}(x) \to 0 \quad \text{as } n \to \infty.$$

As a consequence, Fubini's theorem yields the following lemma.

LEMMA 2.9. *Assume $X_0 \sim \mu$. Then, under $k_n/n \to 0$,*

$$d_{k_n:n}(X_0) \to 0 \quad \text{as } n \to \infty \text{ with probability } 1.$$

We are now in a position to verify conditions (i)–(v) from Theorem 1.1. Recall $\mathbf{X} = (X_1^0, \dots, X_k^0)$ and set, for $\pi = (\pi_1, \dots, \pi_k)$,

$$A_{\pi_i} = \left\{X_{\pi_i} \text{ is among the } k_n\text{-NN of } X_i^0 \text{ in the sample } X_1, \dots, X_n, X_i^0\right\},$$

$$B_{\pi_i} = \left\{X_i^0 \text{ is among the } k_n\text{-NN of } X_{\pi_i} \text{ in the sample } X_1^0, \dots, X_k^0, X_j, \right.$$
$$\left. j \neq \pi_1, \dots, \pi_{i-1}, \pi_{i+1}, \dots, \pi_k\right\}.$$

Note that, for any $f \geq 0$,

$$\int_{\cap A_{\pi_i}} f(X_\pi)\, d\mathbb{P} = \int_{\cap B_{\pi_i}} f(\mathbf{X})\, d\mathbb{P}.$$

However,

$$B_{\pi_i} \subset \left\{X_i^0 \text{ is among the } (k_n + k - 1)\text{-NN of } X_{\pi_i}\right.$$
$$\left. \text{in the sample } X_1^0, \dots, X_k^0, X_1, \dots, X_n\right\}.$$

According to Bickel and Breiman [(1983); Corollary S1] there exists some finite $C(d)$ such $X_i^0$ can be among the $(k_n + k - 1)$-NN of at most $C(d)(k_n + k - 1)$ points. Conclude that

$$\mathbb{E}\left[\sum_\pi |W_\pi(\mathbf{X})| f(X_\pi)\right] \leq C^k(d)\left[\frac{k_n + k - 1}{k_n}\right]^k \mathbb{E}f(\mathbf{X}),$$

proving (i). Since

$$\sum_\pi W_\pi(\mathbf{X}) = 1$$

and

$$\sum_{\pi, \sigma}{}^{(q)} |W_\pi(\mathbf{X}) W_\sigma(\mathbf{X})| = O(k_n^{-1}) = o(1)$$

for $1 \leq q \leq k$, only (iii) is left. But this follows from Lemma 2.9 and

$$\sum_\pi |W_\pi(\mathbf{X})| 1_{\{\|X_\pi - \mathbf{X}\| > \varepsilon\}} \leq 1_{\{\max_{1 \leq i \leq k} d_{k_n:n}(\mathbf{X}_i) > \varepsilon\}}.$$

The proof of Theorem 1.3 is complete. □

**3. Nonparametric discrimination.** The classical problem of discrimination is one of estimating the value of an unobservable random variable $Y^0$ taking values from a finite set $\{1, 2, \dots, M\}$, say. This estimation takes into account an observable vector $X^0$ which is (hopefully) correlated with $Y^0$. The optimal predictor $g(X^0)$ minimizing the probability or error $\mathbb{P}(g(X^0) \neq Y^0)$

among all $g$'s is called the Bayes rule. It is well-known that this $g$ satisfies

$$g(x) = \arg \max_{1 \le j \le M} p_j(x),$$

where

$$p_j(x) = \mathbb{P}(Y^0 = j \mid X^0 = x).$$

Since in practice the $p_j$ are seldom known, they need to be estimated from a training sample $(X_i, Y_i)$, $1 \le i \le n$. Universally consistent regression function estimators then guarantee consistent estimation of the $p_j$ and hence of $g$, resulting in asymptotically Bayes-risk consistent discrimination rules. See Devroye and Wagner (1980) for a nice treatment.

The results of the present paper enable us to extend the classical discrimination problem to a more complex one. Suppose that not just one $X^0$ but $k \ge 2$ random vectors $X_1^0, \dots, X_k^0$ are available for which the corresponding $\mathbf{Y} = (Y_1^0, \dots, Y_k^0)$ is not observable. Take, for example, $k = 2$ and let the $Y$'s be real-valued. We may then wonder whether $Y_1^0 \le Y_2^0$ or not. Setting

$$h(y_1, y_2) = \begin{cases} 1, & \text{if } y_1 \le y_2, \\ 0, & \text{if } y_1 > y_2, \end{cases}$$

we arrive at a discrimination problem for $h(Y_1^0, Y_2^0)$. Given $(X_1^0, X_2^0)$ and a training sample $(X_i, Y_i)$, $1 \le i \le n$, a decision has to be made as to whether $h = 1$ or 0. For an arbitrary $k \ge 2$, we might be interested in that $j$, $1 \le j \le k$, for which $Y_j^0$ is the $r$-largest among the $Y$'s. As another example, we may wonder if $Y_1^0$ and $Y_2^0$ are within a given distance $\varepsilon$. Finally, for discrete $Y$'s, $Y_1^0$ and $Y_2^0$ may coincide or not.

More generally, let $h$ be any function taking on at most finitely many values, say, $1, \dots, M$. The sets

$$A_j = \{(y_1, \dots, y_k) \colon h(y_1, \dots, y_k) = j\}, \qquad 1 \le j \le M,$$

then yield a partition of the feature space. Predicting the value of $h(Y_1^0, \dots, Y_k^0)$ is tantamount to predicting the set in the partition to which $(Y_1^0, \dots, Y_k^0)$ belongs. Now, it is easily seen that, for any discrimination rule $g$,

$$\begin{aligned} \mathbb{P}(g(\mathbf{X}) = h(\mathbf{Y})) &= \sum_{j=1}^{M} \mathbb{P}(g(\mathbf{X}) = j, h(\mathbf{Y}) = j) \\ &= \sum_{j=1}^{M} \int_{\{g(\mathbf{X}) = j\}} \mathbb{P}(h(\mathbf{Y}) = j \mid \mathbf{X})\, d\mathbb{P} \\ &= \sum_{j=1}^{M} \int_{\{\mathbf{x} : g(\mathbf{x}) = j\}} m^j(\mathbf{x}) \mu(dx_1) \cdots \mu(dx_k) \\ &\le \sum_{j=1}^{M} \int_{\{\mathbf{x} : g(\mathbf{x}) = j\}} \max m^j(\mathbf{x}) \mu(dx_1) \cdots \mu(dx_k), \end{aligned}$$

with

$$m^j(\mathbf{x}) = \mathbb{P}(h(\mathbf{Y}) = j \mid \mathbf{X} = \mathbf{x}).$$

The above inequality becomes an equality for

$$g_0(\mathbf{x}) = \arg\max_{1 \le j \le M} m^j(\mathbf{x}); \tag{3.1}$$

$g_0$ defined by (3.1) is again called the Bayes rule, and the pertaining probability or error

$$L^* = 1 - \mathbb{P}(g_0(\mathbf{X}) = h(\mathbf{Y})) = 1 - \mathbb{E}\left[\max_{1 \le j \le M} m^j(\mathbf{X})\right]$$

is called the Bayes risk. Each of the above $m^j$'s can be consistently estimated by one of the two methods discussed in Section 1. Let

$$m_n^j(\mathbf{x}) = \sum_{\pi} W_\pi(\mathbf{x}) 1_{\{h(Y_\pi)=j\}}, \qquad 1 \le j \le M,$$

and set

$$g_{n0}(\mathbf{x}) = \arg\max_{1 \le j \le M} m_n^j(\mathbf{x}).$$

Write

$$L_n \colon = \mathbb{P}(g_{n0}(\mathbf{X}) \ne h(\mathbf{Y})).$$

The following theorem shows that the discrimination rule $g_{n0}$ is asymptotically Bayes-risk consistent, under no assumptions whatsoever.

THEOREM 3.1. *Assume the weights $\{W_{\pi n}\}$ are universally consistent. Then*

$$L_n \to L^* \quad \textit{as } n \to \infty.$$

PROOF. Follows from Theorem 1.1 and the obvious relation

$$0 \le L_n - L^* \le 2\mathbb{E}\left[\max_{1 \le j \le M} \left|m_n^j(\mathbf{X}) - m^j(\mathbf{X})\right|\right]. \qquad \square$$

Theorem 1.1 may also be utilized to study the conditional error given the data

$$\hat{L}_n = \mathbb{P}(g_{n0}(\mathbf{X}) \ne h(\mathbf{Y}) \mid X_1, Y_1, \ldots, X_n, Y_n).$$

Similar to before, one obtains

$$\left|\hat{L}_n - L^*\right| \le 2 \int \max_{1 \le j \le M} \left|m_n^j(\mathbf{x}) - m^j(\mathbf{x})\right| \mu(dx_1) \cdots \mu(dx_k),$$

and therefore

$$\widehat{L}_n \to L^* \quad \text{in the mean}$$

(and hence in probability).

## REFERENCES

BICKEL, P. J. and BREIMAN, L. (1983). Sums of functions of nearest neighbor distances, moment bounds, limit theorems and a goodness of fit test. *Ann. Probab.* **11** 185–214.

DEVROYE, L. P. and WAGNER, T. J. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Ann. Statist.* **8** 231–239.

HOEFFDING, W. (1948). A class of statistics with asymptotically normal distribution. *Ann. Math. Statist.* **19** 293–325.

SPIEGELMAN, C. and SACKS, J. (1980). Consistent window estimation in nonparametric regression. *Ann. Statist.* **8** 240–246.

STONE, C. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–620.

STUTE, W. (1991). Conditional $U$-statistics. *Ann. Probab.* **19** 812–825.

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

MATHEMATISCHES INSTITUT
JUSTUS-LIEBIG-UNIVERSITÄT
ARNDTSTRASSE 2
D-35392 GIESSEN
GERMANY