from (56) we get

$$\frac{1}{\eta} = \frac{1+\sum_{m=1}^{\infty} p(m)}{K\,R} \le \frac{1+\left\lceil \frac{K\,R}{\eta_{\mathrm{erg}}(K)} \right\rceil + \frac{\mathrm{e}^{-n_2(K)\phi_1}}{1-\mathrm{e}^{-\phi_1}}}{K\,R} + \frac{\sum_{\ell=1}^{K-1} \binom{K}{\ell} \left[\frac{1-\mathrm{e}^{-n_2(\ell)\phi_2}}{1-\mathrm{e}^{-\phi_2}} + \frac{\mathrm{e}^{-n_2(\ell)\phi_3}}{1-\mathrm{e}^{-\phi_3}} \right]}{K\,R}. \quad (63)$$

By taking the limit for $R \to \infty$ and by using the fact that $\eta \le \eta_{\mathrm{erg}}(K)$, we obtain that $\lim_{R\to\infty} \eta = \eta_{\mathrm{erg}}(K)$ for all K. This shows that the "all-or-none" INR strategy coupled with JMUD is throughput-wise optimal for every finite K. Finally, by letting $K \to \infty$ we can achieve the unfaded single-user upper bound in (55) with equality.

REFERENCES

[1] G. Caire and D. Tuninetti, "ARQ protocols for the Gaussian collision channel," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1971–1988, July 2001.

[2] S. Shamai (Shitz) and S. Verdú, "The impact of frequency-flat fading on the spectral efficiency of CDMA," *IEEE Trans. Inform. Theory*, vol. 47, pp. 1302–1327, May 2001.

[3] I. Bettesh and S. Shamai (Shitz), "Outages, expected rates and delays in multiple-users fading channels," in *Proc. 2000 Conf. Information Science and Systems*, vol. I, Princeton, NJ, Mar. 2000.

[4] S. Lin and D. Costello, *Error Control Coding: Fundamentals and Applications*. Englewood Cliffs, NJ: Prentice-Hall, 1983.

[5] D. Tse and S. Hanly, "Linear multiuser receivers: Effective interference, effective bandwidth and capacity," *IEEE Trans. Inform. Theory*, vol. 45, pp. 641–657, Mar. 1999.

[6] S. Verdú and S. Shamai (Shitz), "Spectral efficiency of CDMA with random spreading," *IEEE Trans. Inform. Theory*, vol. 45, pp. 622–640, Mar. 1999.

[7] E. Biglieri, J. Proakis, and S. Shamai (Shitz), "Fading channels: Information-theoretic and communications aspects," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2619–2692, Oct. 1998.

[8] J. G. Proakis, *Digital Communications*, 3rd ed. New York: McGraw-Hill, 1995.

[9] S. Verdú, *Multiuser Detection*. Cambridge, U.K.: Cambridge Univ. Press, 1998.

[10] D. Bertsekas and R. Gallager, *Data Networks*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1987.

[11] H. David, *Ordered Statistics*, 2nd ed. New York: Wiley, 1981.

[12] W. Kaplan, *Advanced Calculus*. Reading, MA: Addison-Wesley, 1991.

[13] G. R. Grimmet and D. R. Strizaker, *Probability and Random Processes*, 2nd ed. Oxford, U.K.: Oxford Univ. Press, 1992.

Data-Dependent k_n-NN and Kernel Estimators Consistent for Arbitrary Processes

Sanjeev R. Kulkarni, *Senior Member, IEEE*, Steven E. Posner, and Sathyakama Sandilya

***Abstract*—Let $X_1, X_2, \ldots$ be an arbitrary random process taking values in a totally bounded subset of a separable metric space. Associated with X_i we observe Y_i drawn from an unknown conditional distribution $F(y|X_i = x)$ with continuous regression function $m(x) = E[Y|X = x]$. The problem of interest is to estimate Y_n based on X_n and the data $\{(X_i, Y_i)\}_{i=1}^{n-1}$. We construct appropriate data-dependent nearest neighbor and kernel estimators and show, with a very elementary proof, that these are consistent for *every* process $X_1, X_2, \ldots$.**

***Index Terms*—Arbitrary random processes, consistency, data dependent, kernel estimate, nearest neighbor estimate, nonparametric regression.**

I. INTRODUCTION

Let $X_1, X_2, \ldots$ be an arbitrary random process taking values in a subset of a general separable metric space $(\mathcal{X}, \rho)$. Special cases include nonstationary or nonergodic processes and deterministic sequences. Each $X_i = x_i$ has an associated label Y_i which is a random variable drawn from an unknown conditional distribution $F(y|X_i = x_i)$ taking values in a Hilbert space $\mathcal{Y}$. We consider the nonparametric regression estimation problem of estimating $m(X_n) = E[Y_n|X_n]$ given X_n and previous data pairs $\{(X_i, Y_i)\}_{i=1}^{n-1}$.

Most previous work has considered the case in which the data pairs $\{(X_i, Y_i)\}$ are independent and identically distributed (i.i.d.), although some work has also been done for various weakly dependent data, see [9]–[11], [13], [16], [17]. It is well known that, in this case, various universally consistent regression estimators exist under only a finite moment condition on Y. For example, see Györfi, Härdle, Sarda, and Vieu [6], Roussas [14], and the references therein for results with dependent data. There has been significant interest in analyzing the performance of nearest neighbor and kernel estimators and, in particular, establishing consistency results and obtaining rates of convergence, for instance, see [1]–[3], [5], [7], [18]. Recently, Kulkarni and Posner [8] considered the case in which the process $X_1, X_2, \ldots$ takes values in a compact set, but is otherwise completely arbitrary. For continuous regression functions they have shown that results analogous to the i.i.d. case can be obtained for several standard estimators (such as k_n-nearest neighbor and kernel estimators) using a cumulative loss criterion. These results are also related to other work on individual sequences such as [12] on density estimation and references therein.

In this correspondence, we also impose no restrictions on the random process $\{X_i\}$ except that almost surely the set $\{X_1, X_2, \ldots\}$ be to-

Manuscript received March 6, 2001; revised February 2, 2002. This work was supported in part by the National Science Foundation under NYI Grant IRI-9457645 and Grant ECS-9873451, and MURI through the Army Research Office under Grant DAAD19-00-1-0466.

S. R. Kulkarni is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA (e-mail: kulkarni@ee.princeton.edu).

S. E. Posner is with Goldman, Sachs and Co., New York, NY 10004 USA (e-mail: steven.posner@gs.com).

S. Sandilya was with Princeton University, Princeton, NJ 08544 USA. He is now with Siemens Corporate Research, Princeton, NJ 08540 USA (e-mail: sandilya@alumni.princeton.edu).

Communicated by G. Lugosi, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Publisher Item Identifier 10.1109/TIT.2002.802611.

0018-9448/02$17.00 © 2002 IEEE

tally bounded. As in [8], we require the regression function to be continuous, and that given X_i, the label Y_i is conditionally independent of $\{(X_j, Y_j)\}_{i \neq j}$. Although this is a rather general setting, we show, perhaps surprisingly, that appropriate data-dependent estimators for $m(X_n)$ can be constructed that are consistent for *every* totally bounded process $X_1, X_2, \ldots$. Thus, for the price of continuity of the regression function, consistent estimators can be constructed with almost complete generality in the process $X_1, X_2, \ldots$, and interestingly, the proofs are very simple. In Section II, we give a precise formulation of the problem, Sections III and IV consider data-dependent nearest neighbor and kernel estimators, respectively, that are consistent for every $X_1, X_2, \ldots$.

II. Formulation

Consider a sequential estimation problem as follows. Let $X_1, X_2, \ldots$ be an arbitrary random process taking values in a complete separable metric space $(\mathcal{X}, \rho)$ and let $Y_1, Y_2, \ldots$ be a corresponding sequence of random variables taking values in a Hilbert space $\mathcal{Y}$ equipped with an inner-product-induced norm $|\cdot|$. A concrete example of spaces satisfying the above conditions would be $\mathcal{X} = R^k$ and $\mathcal{Y} = R$ with the usual Euclidean norm $|\cdot|$. Our goal is to sequentially estimate $m(X_n) = E[Y_n|X_n]$ using only the data

$$\mathcal{D}_n = X_n \cup \{(X_i, Y_i)\}_{i=1}^{n-1}.$$

This problem formulation is quite general and to make useful statements, we require certain assumptions.

We impose the following assumption on the pairs (X_i, Y_i) which implies that given X_i, the label Y_i is conditionally independent of $\{(X_j, Y_j)\}_{j \neq i}$ and drawn according to $F(y|X_i)$.

(A0) For each i and for every measurable set S

$$\Pr(Y_i \in S | X_1, \ldots, X_n, Y_1, \ldots, Y_{i-1}, Y_{i+1}, Y_n) = \Pr(Y_i \in S | X_i) = \int_S F(dy|X_i).$$

Throughout this correspondence, we also impose the following assumptions on the conditional distribution $F(y|x)$.

(A1) $\sup_{x \in \mathcal{X}} E[|Y|^2 | X = x] < \infty$,

(A2) the regression function $m(x) = E[Y|X = x]$ is a continuous function.

Recall that a totally bounded subset of a metric space is one that can be finitely ϵ-covered for each $\epsilon > 0$. We say that a process $X_1, X_2, \ldots$ is *totally bounded* if almost surely the set $\{X_1, X_2, \ldots\}$ is a totally bounded subset of $\mathcal{X}$.

III. A Data-Dependent Nearest Neighbor Estimator

Let X'_{ni} be the ith closest sample to X_n from $X_1, \ldots, X_{n-1}$, and let Y'_{ni} denote the Y_j associated with $X_j = X'_{ni}$ (where ties are broken arbitrarily). Let $d_n(i; X_1, \ldots, X_n) = \rho(X_n, X'_{ni})$ denote the ith nearest neighbor (NN) distance to X_n from $X_1, \ldots, X_{n-1}$. The k_n-NN estimate of $m(X_n)$ is defined as

$$\hat{m}_n(X_n) = \frac{1}{k_n} \sum_{i=1}^{k_n} Y'_{ni}.$$

In what follows, we will be interested in k_n-NN estimators in which $k_n = k_n(X_1, \ldots, X_n)$ is chosen in a data-dependent manner.

Theorem 1: Let $X_1, X_2, \ldots$ be an arbitrary random process and suppose $(X_1, Y_1), (X_2, Y_2), \ldots$ satisfy (A0)–(A2). If

i) $k_n(X_1, \ldots, X_n) \to_{n \to \infty} \infty$ a.s. and

ii) $d_n(k_n; X_1, \ldots, X_n) \to_{n \to \infty} 0$ a.s.

then the corresponding k_n-NN estimator satisfies

$$\lim_{n \to \infty} E\left[|\hat{m}_n(X_n) - m(X_n)|^2 | X_1, \ldots, X_n\right] = 0 \quad \text{a.s.}$$

Proof: Using i), ii), and the fact that $X_1, X_2, \ldots$ is a totally bounded process, we have that almost surely a realization $\omega = (x_1, x_2, \ldots)$ is a totally bounded set with $k_n(x_1, \ldots, x_n) \to \infty$ and $d_n(k_n; x_1, \ldots, x_n) \to 0$. Fix such a realization.

Since $\mathcal{X}$ is complete, the closure of $\{x_1, x_2, \ldots\}$, denoted by A, is a compact subset of $\mathcal{X}$ (e.g., see [4, Theorem 2.3.1]). Then the continuity of $m(x)$ (assumption A2) implies that $m(x)$ is in fact uniformly continuous on A (e.g., see [4, Corollary 2.4.6]).

Fix $\epsilon > 0$. Then since $m(x)$ is uniformly continuous on A, there exists δ such that for any $u, v \in A$ we have $|m(u) - m(v)| < \epsilon/2$ whenever $\rho(u, v) < \delta$. Let N_1 be such that $k_n = k_n(x_1, \ldots, x_n) > 2M/\epsilon$ for all $n \geq N_1$ where $M = \sup_{x \in \mathcal{X}} E[|Y|^2 \,|\, X = x]$ (which is finite by assumption A2). Let N_2 be such that $d_n(k_n; x_1, \ldots, x_n) < \delta$ for all $n \geq N_2$, and let $N = \max\{N_1, N_2\}$. Then for any $n \geq N$ we have

$$\begin{aligned}
&E\left[|\hat{m}(X_n) - m(X_n)|^2 | \Omega_n = \omega_n\right] \\
&\quad = E\left[\left|\frac{1}{k_n}\sum_{i=1}^{k_n} Y'_{ni} - m(x_n)\right|^2 \Big| \Omega_n = \omega_n\right] \\
&\quad \leq E\left[\left|\frac{1}{k_n}\sum_{i=1}^{k_n} Y'_{ni} - m(x'_{ni})\right|^2 \Big| \Omega_n = \omega_n\right] \\
&\qquad + \left|\frac{1}{k_n}\sum_{i=1}^{k_n} (m(x'_{ni}) - m(x_n))\right|^2 \\
&\quad \leq \frac{M}{k_n} + \max_{1 \leq i \leq k_n} \left|m(x'_{ni}) - m(x_n)\right|^2 \\
&\quad \leq \epsilon/2 + \epsilon/2 = \epsilon
\end{aligned}$$

where the second inequality follows from assumptions (A0) and (A1), and the third inequality follows from the choice of $N = \max\{N_1, N_2\}$. Since $\epsilon > 0$ was arbitrary the result follows. □

Theorem 2: For every totally bounded process $X_1, X_2, \ldots$, if

$$k_n(X_1, \ldots, X_n) = \arg\min_k \frac{1}{k} + d_n(k; X_1, \ldots, X_n)$$

then we have

i) $k_n(X_1, \ldots, X_n) \to_{n \to \infty} \infty$ a.s. and

ii) $d_n(k_n; X_1, \ldots, X_n) \to_{n \to \infty} 0$ a.s.

Proof: Let

$$J_n = \min_k \frac{1}{k} + d_n(k; X_1, \ldots, X_n).$$

We need only show that $\lim_{n \to \infty} J_n = 0$ almost surely. We will do this by showing that for any $\epsilon > 0$, almost surely we have $J_n < \epsilon$ for sufficiently large n.

Fix $\epsilon > 0$ and a realization $\omega = (x_1, x_2, \ldots)$. Since $X_1, X_2, \ldots$ is a totally bounded process, almost surely there exists a totally bounded set A (which may depend on the particular realization ω) such that

$x_1, x_2, \ldots \in A$. Let $B_1, \ldots, B_{N(\epsilon/4)}$ denote balls of radius $\epsilon/4$ forming a finite cover of A. Then $x_n \in \bigcup_{i=1}^{N(\epsilon/4)} B_i$ for all n.

For any fixed k, the number of times an x_i falls in some ball B_i with fewer than k previous elements from $x_1, \ldots, x_{i-1}$ is bounded by $kN(\epsilon/4)$. Hence, there is a finite n_0 such that for all $n \geq n_0$, the number of $x_1, \ldots, x_{n-1}$ within $\epsilon/2$ of x_n is greater than k. Thus, for $n \geq n_0$ we have

$$J_n \leq \frac{1}{k} + \frac{\epsilon}{2}.$$

The result follows by taking $k > 2/\epsilon$. □

The next result follows immediately from Theorems 1 and 2.

Corollary 1: Suppose (A0)–(A2) are satisfied. For

$$k_n(X_1, \ldots, X_n) = \arg\min_k \frac{1}{k} + d_n(k; X_1, \ldots, X_n)$$

the corresponding k_n-NN estimator is consistent for every totally bounded process $X_1, \ldots, X_n$.

Note here that a consistent estimator can also be obtained by choosing

$$k_n(X_1, \ldots, X_n) = \arg\min_k f\left(\frac{1}{k}\right) + g(d_n(k; X_1, \ldots, X_n))$$

where f and g are any strictly positive monotonic functions that decrease to 0 as their arguments approach 0.

IV. A Data-Dependent Kernel Estimator

For an arbitrary process $X_1, X_2, \ldots$, define the kernel weights

$$W_{ni}(\Omega_n) = \frac{\phi(\frac{\rho_{ni}}{\epsilon_n})}{\sum_{j=1}^{n-1} \phi(\frac{\rho_{nj}}{\epsilon_n})}$$

where $\{\epsilon_n\}$ is a sequence of positive numbers, ϕ: $R_+ \to R_+$ is a nonnegative kernel function, and $\rho_{ni} = \rho(X_i, X_n)$. Note that in the expression for the weights and in the sequel we treat $0/0$ as 0. The kernel regression estimate is defined as

$$\hat{m}_n(X_n) = \sum_{i=1}^{n-1} W_{ni}(\Omega_n) Y_i = \sum_{i=1}^{n-1} \frac{\phi(\frac{\rho_{ni}}{\epsilon_n}) Y_i}{\sum_{j=1}^{n-1} \phi(\frac{\rho_{nj}}{\epsilon_n})}.$$

For simplicity, we will assume that the kernel ϕ satisfies the following conditions:

- ϕ has compact support and $\sup_{t \geq 0} \phi(t) < \infty$;
- ϕ is bounded away from zero on $[0, 1]$.

Given $X_1, \ldots, X_n$ and ϵ_n, let $L_n(\epsilon_n; X_1, \ldots, X_n)$ denote the number of X_i for $i = 0, \ldots, n-1$ such that $\rho(X_i, X_n) < \epsilon_n$.

Theorem 3: Let $X_1, X_2, \ldots$ be an arbitrary random process and suppose $(X_1, Y_1), (X_2, Y_2), \ldots$ satisfy (A0)–(A2). Let $\phi(\cdot)$ be a kernel function satisfying the preceding conditions. If

i) $\epsilon_n(X_1, \ldots, X_n) > 0$ and $\epsilon_n(X_1, \ldots, X_n) \to_{n\to\infty} 0$ a.s. and

ii) $L_n(\epsilon_n; X_1, \ldots, X_n) \to_{n\to\infty} \infty$ a.s.

then the corresponding kernel estimator satisfies

$$\lim_{n\to\infty} E\left[|\hat{m}_n(X_n) - m(X_n)|^2 | X_1, \ldots, X_n\right] = 0 \quad \text{a.s.}$$

Proof: First, as before, we have

$$\begin{aligned}
&E\left[|\hat{m}(X_n) - m(X_n)|^2 | \Omega_n = \omega_n\right] \\
&\quad = E\left[\left|\sum_{i=1}^{n-1} W_{ni}(\Omega_n) Y_i - m(x_n)\right|^2 \Big| \Omega_n = \omega_n\right] \\
&\quad = E\left[\left|\sum_{i=1}^{n-1} W_{ni}(\Omega_n)(Y_i - m(x_i))\right|^2 \Big| \Omega_n = \omega_n\right] \\
&\qquad + \left|\sum_{i=1}^{n-1} W_{ni}(\Omega_n) m(x_i) - m(x_n)\right|^2 \\
&\quad = \sum_{i=1}^{n-1} W_{ni}^2(\Omega_n) E[|Y_i - m(x_i)|^2 | \Omega_n = \omega_n] \\
&\qquad + \left|\sum_{i=1}^{n-1} W_{ni}(\omega_n) m(x_i) - m(x_n)\right|^2 \\
&\quad \leq M \sum_{i=1}^{n-1} W_{ni}^2(\Omega_n) + \max_{1 \leq i < n, W_{ni} > 0} |m(x_i) - m(x_n)|^2.
\end{aligned}$$

Also,

$$\begin{aligned}
\sum_{i=1}^{n-1} W_{ni}^2 &= \sum_{i=1}^{n-1} \left(\frac{\phi(\frac{\rho_{ni}}{\epsilon_n})}{\sum_{j=1}^{n-1} \phi(\frac{\rho_{nj}}{\epsilon_n})}\right)^2 \\
&= \frac{1}{\left(\sum_{j=1}^{n-1} \phi(\frac{\rho_{nj}}{\epsilon_n})\right)^2} \sum_{i=1}^{n-1} \phi^2\left(\frac{\rho_{ni}}{\epsilon_n}\right) \\
&\leq \frac{\sup \phi(t)}{\sum_{i=1}^{n-1} \phi(\frac{\rho_{ni}}{\epsilon_n})}.
\end{aligned}$$

Now in order to show that for large enough N, the error will be small we need to establish that $\sum_{i=1}^{n-1} \phi(\frac{\rho_{ni}}{\epsilon_n})$ diverges. Since ϕ is bounded away from zero on $[0, 1]$, there exist $B > 0$ such that $\phi(t) > B$ for all $t \in [0, 1]$. Hence,

$$\sum_{i=1}^{n-1} \phi\left(\frac{\rho_{ni}}{\epsilon_n}\right) > B L_n(\epsilon_n; X_1, \ldots, X_n) \to \infty$$

since $L_n(\epsilon_n; X_1, \ldots, X_n) \to \infty$.

Say ϕ is zero outside the interval $[0, K]$. As in the proof of Theorem 1, we have that the set A, which is the closure of $\{x_i\}$, is compact and hence m is uniformly continuous on A. Fix $\epsilon > 0$. Now, there exists δ such that $|m(y) - m(x)|^2 < \epsilon/2$ if $|y - x| < \delta$. Pick N_1 so that $K\epsilon_n < \delta$ for all $n > N_1$. Further, pick N_2 so that

$$B L_n > \frac{2M}{\epsilon \sup \phi(t)}, \qquad \text{for all } n > N_2$$

where $M = \sup_{x \in \mathcal{X}} E[|Y|^2 | X = x]$ (which is finite by assumption A2). Let $N = \max\{N_1, N_2\}$. Thus for all $n > N$ we have

$$E\left[|\hat{m}(X_n) - m(X_n)|^2 | \Omega_n = \omega_n\right] \leq \epsilon/2 + \epsilon/2 = \epsilon. \qquad \square$$

Theorem 4: For every totally bounded process $X_1, X_2, \ldots$, if

$$\alpha_n(X_1, \ldots, X_n) = \arg\min_{\alpha} \alpha + \frac{1}{L_n(\alpha; X_1, \ldots, X_n)}$$

then we have

i) $\alpha_n(X_1, \ldots, X_n) \to_{n\to\infty} 0$ a.s. and

ii) $L_n(\alpha_n; X_1, \ldots, X_n) \to_{n\to\infty} \infty$ a.s.

Proof: The proof of this result is very similar to that of Theorem 2. Let

$$J_n = \min_{\alpha} \alpha + \frac{1}{L_n(\alpha; X_1, \ldots, X_n)}.$$

As before, we need only show that $\lim_{n\to\infty} J_n = 0$ almost surely, and we do this by showing that for any $\epsilon > 0$, almost surely for sufficiently large n we have $J_n < \epsilon$.

Form a finite $\epsilon/4$-cover as in Theorem 2. Then since the cover is finite, it is easy to see that in an $\epsilon/2$ neighborhood of x_i, we can have fewer than $2/\epsilon$ points only finitely many times. Hence, almost surely for sufficiently large n we have $J_n \leq \epsilon$. $\square$

The subsequent result follows immediately from Theorems 3 and 4.

Corollary 2: Suppose (A0)–(A2) are satisfied. Let

$$\alpha_n(X_1, \ldots, X_n) = \arg\min_{k} \frac{1}{k} + d_n(k; X_1, \ldots, X_n)$$

and let

$$\epsilon_n(X_1, \ldots, X_n) = \max\{\alpha_n(X_1, \ldots, X_n), \beta_n\}$$

where β_n is any positive sequence (possibly depending on $X_1, \ldots, X_n$) converging to 0 almost surely. Let ϕ be an admissible kernel function. Then the corresponding kernel estimator is consistent for every totally bounded process $X_1, X_2, \ldots$.

V. Remarks

The notion of a totally bounded process considered here is slightly more general than the condition used in [8] that the X_i take values in some compact set A almost surely. If the process $X_1, X_2, \ldots$ is totally bounded, then almost surely a realization $x_1, x_2, \ldots$ is a totally bounded set, and so is contained in a compact set (namely, the closure of $\{x_1, x_2, \ldots\}$). However, the compact set can depend on the realization so there may be no single compact A with $X_i \in A$ almost surely. It turns out that the consistency results in [8] actually hold for totally bounded processes, although the results on rates of convergence need the stronger condition.

To get the consistency results presented here, one needs to choose the parameters k_n or ϵ_n in a data-dependent manner. For any data-independent choices, one can construct examples of totally bounded processes $X_1, X_2, \ldots$ for which consistency fails. In contrast, the results of [8] show that for a time-average criterion any k_n-NN or kernel estimator under the standard conditions works.

One cannot, in general, get rates of convergence, even with a Lipschitz assumption on the conditional distribution $F(Y|X)$. To get rates one would also need to put conditions on the process $X_1, X_2, \ldots$ that allow getting rates on $k_n(X, \ldots, X_n)$ and $d_n(k_n; X_1, \ldots, X_n)$ (or analogous quantities). However, as shown in [8], one can get rates on cumulative risk with conditions only on $F(Y|X)$.

One could actually do a similar analysis for the general case of Stone-type estimators [15], and treat the k_n-NN and kernel estimators as special cases, but the main point is the existence of estimators with the properties shown and for simplicity we focus on NN and kernel estimators.

References

[1] J. Beck, "The exponential rate of convergence of error for $\boldsymbol{k}_n$-NN nonparametric regression and decision," *Prob. Contr. Inform. Theory*, vol. 8, pp. 303–311, 1979.

[2] T. M. Cover, "Rates of convergence for nearest neighbor procedures," in *Proc. 1st Annu. Hawaii Conf. Systems Theory*, Jan. 1968, pp. 413–415.

[3] L. Devroye, "Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 61, pp. 467–481, 1982.

[4] R. M. Dudley, *Real Analysis and Probability*. London, U.K.: Chapman & Hall, 1989.

[5] L. Györfi, "The rate of convergence of $\boldsymbol{k}_n$-NN regression estimates and classification rules," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 362–364, May 1981.

[6] L. Györfi, W. Härdle, P. Sarda, and P. Vieu, *Nonparametric Curve Estimation From Time Series*. Berlin, Germany: Springer-Verlag, 1989.

[7] A. Krzyżak, "The rates of convergence of kernel regression estimates and classification rules," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 668–679, Sept. 1986.

[8] S. R. Kulkarni and S. E. Posner, "Rates of convergence of nearest neighbor estimation under arbitrary sampling," *IEEE Trans. Inform. Theory*, vol. 41, pp. 1028–1039, July 1995.

[9] D. Modha and E. Masry, "Memory-universal prediction of stationary random processes," *IEEE Trans. Inform. Theory*, vol. 44, pp. 117–133, Jan. 1998.

[10] G. Morvai, S. R. Kulkarni, and A. B. Nobel, "Regression estimation from an individual stable sequence," *Statistics*, vol. 33, pp. 99–119, 1999.

[11] A. B. Nobel, "Limits to classification and regression estimation from ergodic processes," *Ann. Statist.*, vol. 27, pp. 262–273, 1999.

[12] A. B. Nobel, G. Morvai, and S. Kulkarni, "Density estimation from an individual numerical sequence," *IEEE Trans. Inform. Theory*, vol. 44, pp. 537–541, Mar. 1998.

[13] G. Roussas, "Nonparametric estimation in Markov processes," *Ann. Inst. Statist. Math.*, vol. 21, pp. 73–87, 1967.

[14] G. Roussas, Ed., *Nonparametric Functional Estimation and Related Topics*. Amsterdam, The Netherlands: Kluwer, 1991.

[15] C. J. Stone, "Consistent nonparametric regression," *Ann. Statist.*, vol. 5, pp. 595–645, 1977.

[16] S. Yakowitz, "Nonparametric density and regression estimation from Markov sequences without mixing assumptions," *J. Multivar. Anal.*, vol. 30, pp. 124–136, 1989.

[17] ——, "Nearest neighbor regression estimation for null-recurrent Markov time series," *Stoch. Processes Appl.*, vol. 48, pp. 311–318, 1993.

[18] L. C. Zhao, "Exponential bounds of mean error for the nearest neighbor estimates of regression functions," *J. Multiv. Anal.*, vol. 21, pp. 168–178, 1987.