Rates of Convergence of Nearest Neighbor Estimation Under Arbitrary Sampling

Sanjeev R. Kulkarni, *Member, IEEE*, and Steven E. Posner, *Student Member, IEEE*

***Abstract*—Rates of convergence for nearest neighbor estimation are established in a general framework in terms of metric covering numbers of the underlying space. Our first result is to find explicit finite sample upper bounds for the classical independent and identically distributed (i.i.d.) random sampling problem in a separable metric space setting. The convergence rate is a function of the covering numbers of the support of the distribution. For example, for bounded subsets of $\mathbb{R}^r$, the convergence rate is $O(1/n^{2/r})$. Our main result is to extend the problem to allow samples drawn from a completely *arbitrary* random process in a separable metric space and to examine the performance in terms of the individual sample sequences. We show that for every sequence of samples the asymptotic time-average of nearest neighbor risks equals twice the time-average of the conditional Bayes risks of the sequence. Finite sample upper bounds under arbitrary sampling are again obtained in terms of the covering numbers of the underlying space. In particular, for bounded subsets of $\mathbb{R}^r$ the convergence rate of the time-averaged risk is $O(1/n^{2/r})$. We then establish a consistency result for k_n-nearest neighbor estimation under arbitrary sampling and prove a convergence rate matching established rates for i.i.d. sampling. Finally, we show how our arbitrary sampling results lead to some classical i.i.d. sampling results and in fact extend them to stationary sampling.Our framework and results are quite general while the proof techniques are surprisingly elementary.**

***Index Terms*—Nearest neighbor, nonparametric regression estimation, rates of convergence, metric entropy, covering numbers, worst case, consistency, deterministic sampling, arbitrary sampling.**

I. Introduction

THIS PAPER deals with the problem of estimating a random variable $Y_n \in \mathcal{Y}$ given a sample $X_n \in \mathcal{X}$, where Y_n is drawn according to an unknown conditional distribution $F(y \mid x)$ given $X_n = x$. In addition to X_n, the estimate can be based on $n-1$ previous pairs of data $\{(X_i, Y_i)\}_{i=1}^{n-1}$ where, given $X_i = x_i$, the label Y_i is drawn independently and distributed according to $F(y \mid X_i = x_i)$. One simple and widely studied nonparametric estimation procedure is the nearest neighbor (NN) rule: selecting as an estimate of Y_n the Y_i associated with the nearest neighbor of X_n. Most existing work generally considers the case in which the samples $\{X_i\}_{i=1}^n$ are drawn independent and identically distributed (i.i.d.) from a probability distribution. We add to that body of work by bounding the rate of convergence of the NN rule in a separable metric space. The main concern of this paper is the formulation of a new estimation problem in which the samples $\{X_i\}_{i=1}^n$ need not be drawn i.i.d. but can be an *arbitrary* random process. We investigate convergence of the NN rule in this framework. Our proof techniques are elementary and are based on deterministic sample path analyses, while the results are quite general. In fact, we easily recover the traditional i.i.d. sampling results from the arbitrary sampling results. We emphasize that while the arbitrary sampling results are useful as they recover traditional i.i.d. results, they are also interesting in and of themselves as they apply to completely general sampling schemes including general nonstationary and nonergodic processes, as well as deterministic and mixed sampling strategies. Furthermore, the approach is in keeping with recent trends in information theory to examine performance for arbitrary individual data sequences (e.g., [6], [10]).

Nearest neighbor classification/estimation has received much attention since it was first studied by Fix and Hodges [11]. Cover and Hart [3], [4] proved that under mild conditions on the distribution the nearest neighbor rule risk converges to twice the Bayes risk under squared error loss and is upper-bounded by twice the Bayes risk for metric loss functions. Some results on the convergence rate of the nearest neighbor rule have been established for the classification problem under various assumptions [5], [12], [21]. Convergence of k_n-nearest neighbor regression estimation has also been the subject of much work. Stone [19] investigated the consistency of a general class of nonparametric regression estimators. In particular, he proved that k_n-NN type estimators are "weakly universally consistent" (converge in $\mathcal{L}_p$ for all $p \geq 1$). This work was continued in [8], [9], [14], where "strong consistency" (almost-sure convergence) of k_n-NN estimators was studied. Another generalization of the nearest neighbor estimate was introduced in [7] where it was shown that under various noise conditions the estimates are strongly uniformly consistent. Other results on the convergence rates of k_n-nearest neighbor regression estimation can be found in [1], [2], [13], [16], [17], [20].

In Section II, we precisely formulate the nearest neighbor estimation problem and define various preliminary quantities. Our results allow $X_1, X_2, \cdots, X_n$ to be a sequence of random variables taking values in a general separable metric space $\mathcal{X}$ equipped with metric ρ. The labels $Y_1, Y_2, \cdots, Y_n$ are a sequence of random variables drawn from the conditional distribution $F(y \mid X_i = x_i)$ taking values in a Hilbert space

Manuscript received March 1, 1994; revised January 2, 1995. This work was supported in part by the National Science Foundation under Grants IRI-9209577 and IRI-9457645 and by the U.S. Army Research Office under Grant DAAL03-92-G-0320. The material in this paper was presented in part at the 1994 International Symposium on Information Theory, Trondheim, Norway, June 27–July 1, 1994.

The authors are with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544 USA.

IEEE Log Number 9412067.

$\mathcal{Y}$. We consider several performance criteria such as average risk, cumulative risk, and time-average risk with respect to squared-error loss. Convergence rates for the various risks are obtained in terms of covering numbers of totally bounded sets (sometimes supports of distributions), and so these quantities are discussed in Section II-C.

Our first goal is to find explicit finite sample bounds, i.e., bounds on the convergence rate, of the large sample risk for the standard formulation with i.i.d. sampling. In Section III, we find an upper bound on the expected NN distance for all distributions with support on a totally bounded subset of a separable metric space in terms of the covering numbers of the support. This in turn implies a bound on the convergence rate for the NN estimator.

The main contribution of this paper is to extend this problem to *arbitrary* sequences of samples which is the subject of Section IV. In this problem, the sequence $\{X_i\}_{i=1}^n$ can be an arbitrary random process taking values in a totally bounded metric space, e.g., the sequence can be chosen deterministically or it can be obtained from a general non-i.i.d. process. In Section IV-A, we bound the sum of the NN distances for any sequence of samples. This leads to a result analogous to the i.i.d. problem showing that the asymptotic worst case time-averaged risk is equal to twice the associated Bayes risk. We also show that for any fixed arbitrary sequence the time-averaged risk is equal to twice the time-average of the conditional Bayes risks of the sequence. In fact, precise convergence rates in terms of the covering numbers of the underlying space are also derived. In Section IV-B we show that the error rates for i.i.d. sampling as well as for stationary sampling are recovered in the arbitrary sampling result.

In Section V we prove the consistency of the k_n-NN estimator under arbitrary sampling. In particular, a convergence rate of $O(n^{-2/(r+2)})$ in bounded subsets of $\mathbf{R}^r$ is obtained which is the same as the convergence rate for i.i.d. sampling previously computed [13]. Our proof techniques are not only more elementary, but the results are more powerful as they allow arbitrary sampling. In addition, we extend the result in [13] to stationary sampling.

Our results in Sections III–V are restricted to totally bounded supports. In Section VI we recover i.i.d. sampling results for probability measures that have unbounded support using the results from Section IV-A, which further demonstrates the strength of the arbitrary sampling results.

II. Problem Formulation and Preliminaries

A. Estimation Problem and Bayes Risks

The problem to be considered is the estimation of a random variable Y taking values in $\mathcal{Y}$ given a sample X taking values in $\mathcal{X}$ with the goal of minimizing a loss between Y and the estimate. In this paper we consider squared-error loss and accordingly are performing regression function estimation. We assume $\mathcal{X}$ is a general separable metric space equipped with metric ρ which we denote as the pair $(\mathcal{X}, \rho)$. For simplicity, let $\mathcal{Y} = \mathbf{R}^s$, for some positive integer s, equipped with the usual Euclidean norm induced from the dot product on $\mathbf{R}^s$, i.e., $\|y\|^2 \equiv y \cdot y$ for any $y \in \mathbf{R}^s$. All our results hold for any Hilbert space with essentially no modifications.

Given $X = x$, then Y is drawn according to a conditional distribution $F(y \mid X = x)$. For a given x, an estimator $\tilde{Y}(x)$ yields a conditional squared-error risk $E[\|Y - \tilde{Y}(x)\|^2 \mid X = x]$. If $F(y \mid x)$ is known then the best estimator is known as the Bayes estimator. The Bayes estimator Y^* minimizes this risk resulting in the conditional Bayes risk

$$r^*(x) = E[\|Y - Y^*(x)\|^2 \mid X = x] \leq E[\|Y - \tilde{Y}(x)\|^2 \mid X = x], \quad \forall \tilde{Y}(x).$$

It is understood that there is an underlying probability space $(\mathcal{X}, \mathcal{F}, \mu)$, where $\mathcal{F}$ is the Borel σ-algebra generated by the open sets of $(\mathcal{X}, \rho)$. If X is drawn according to some fixed distribution μ, the *Bayes risk* is given by

$$R_\mu^* = Er^*(X) = \int r^*(x)\mu\,(dx).$$

This is the minimum average loss obtained when X is drawn according to μ. We may omit the subscript μ if the context is clear.

If X can be chosen from an arbitrary distribution, we define the *worst case sampling risk* as

$$R^{w*} = \sup_\mu R_\mu^*.$$

This is the worst Bayes risk taken over all choices of distributions. Equivalently, it is the worst possible conditional Bayes risk over all choices of $x \in \mathcal{X}$. Certainly $R^* \leq R^{w*}$. Of course, if the conditional Bayes risk is constant, $r^*(x) = r^*$, i.e., independent of x, then $R^* = R^{w*}$. This can happen if $Y = E[Y \mid X] + \eta$ where the "noise" η is independent of X and has zero mean.

Define the conditional mean of Y given $X = x$ as

$$m(x) = E[Y \mid X = x]$$

and the conditional variance as

$$\sigma^2(x) = E[\|Y\|^2 \mid X = x] - \|m(x)\|^2.$$

It is easy to show that the Bayes estimator is given by $Y^*(x) = m(x)$, that the conditional Bayes risk is $r^*(x) = \sigma^2(x)$, and that the Bayes risk is $R^* = E\sigma^2(X)$. Throughout this paper we impose the following assumption on $F(y \mid x)$.

Assumption 1 (Lipschitz Type): There exists $K_1, K_2 > 0$ and $0 < \alpha \leq 1$ such that for any $x_1, x_2 \in \mathcal{X}$

$$\|m(x_1) - m(x_2)\| \leq \sqrt{K_1}\rho(x_1, x_2)^\alpha$$

and

$$|\sigma^2(x_1) - \sigma^2(x_2)| \leq K_2\rho(x_1, x_2)^{2\alpha}.$$

B. Nearest Neighbor Estimation

The quantities R^* and R^{w*} are the minimum risks that can be achieved having complete knowledge of the underlying conditional distribution $F(y \mid x)$. If $F(y \mid x)$ is unknown it is often assumed that in addition to X one has available pairs of data $\{(X_i, Y_i)\}_{i=1}^{n-1}$. We impose the following assumption on $\{(X_i, Y_i)\}$ which implies that given X_i, the label Y_i is conditionally independent of $\{(X_j, Y_j)\}_{i \neq j}$ and drawn from $F(y \mid X_i)$.

Assumption 2: For any measureable set S

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

for each i.

In this paper, we consider both the case in which the sequence $X_1, X_2, \cdots$ is drawn i.i.d. according to some probability measure μ and the case in which the sequence is drawn from an arbitrary random process.

The nearest neighbor estimate of Y_n using knowledge of $\{(X_i, Y_i)\}_{i=1}^{n-1} \cup X_n$ is defined as follows. Letting

$$j = \arg\min_{i<n} \rho(X_n, X_i)$$

define $X_n' = X_j$. Then X_n' is the nearest neighbor of X_n from the set $\{X_1, \cdots, X_{n-1}\}$. The NN estimate of Y_n is defined as $Y_n' = Y_j$. This is the Y_i associated with the NN of X_n. The NN conditional risk is defined as the expected loss in estimating Y_n by Y_n' given X_n and its NN X_n', that is

$$r_n(X_n, X_n') = r_n(X_1, \cdots, X_n) = E[\|Y_n - Y_n'\|^2 \mid X_1, \cdots, X_n].$$

Define the nearest distance at time n as $d_n = \rho(X_n, X_n')$. The following lemma which is a modified restatement of a result from [3] provides an upper bound on $r_n(X_n, X_n')$ in terms of d_n.

Lemma 1: Let $X_1, \cdots, X_n$ be an arbitrary random process in $(\mathcal{X}, \rho)$ with $(X_1, Y_1), \cdots, (X_n, Y_n)$ satisfying Assumption 2. For any $F(y \mid x)$ satisfying Assumption 1 we have

$$r_n(X_n, X_n') \leq 2\sigma^2(X_n) + (K_1 + K_2) d_n^{2\alpha}.$$

Proof: We have that

$$\begin{aligned} r_n(X_n, X_n') &= E[\|Y_n - Y_n'\|^2 \mid X_1, \cdots, X_n] \\ &= E[\|(Y_n - m(X_n)) + (m(X_n) - m(X_n')) \\ &\quad + (m(X_n') - Y_n')\|^2 \mid X_1, \cdots, X_n]. \end{aligned}$$

Assumption 2 implies that given $X_1, \cdots, X_n$, each X_i is conditionally independent of Y_j for $i \neq j$ and that $\{Y_i\}$ are mutually conditionally independent. This implies that the expected value of the cross terms is zero. It also implies that the index of the nearest neighbor is independent of $\{Y_i\}$. Hence we get that

$$\begin{aligned} r_n(X_n, X_n') &= \sigma^2(X_n) + \|m(X_n) - m(X_n')\|^2 + \sigma^2(X_n') \\ &= 2\sigma^2(X_n) + \|m(X_n) - m(X_n')\|^2 \\ &\quad + [\sigma^2(X_n') - \sigma^2(X_n)] \\ &\leq 2\sigma^2(X_n) + (K_1 + K_2)\rho(X_n, X_n')^{2\alpha} \end{aligned}$$

from Assumption 1. ∎

For i.i.d. sampling define the expected loss given X_n as $r_n(X_n) = E[r_n(X_n, X_n') \mid X_n]$, and the expected loss at time n

$$R_n = E r_n(X_n).$$

The result in [3] states that for squared-error loss and for Lipschitz smooth regression functions, we have $\lim_{n\to\infty} R_n = 2R^*$.

For arbitrary sampling we can naturally extend the definition of loss at time n given x_n as

$$R_n^w = \sup_{x_n, x_n' \in \mathcal{X}} r_n(x_n, x_n').$$

However, we cannot expect to get bounds on R_n^w under arbitrary sampling. One can always construct sequences $x_1, x_2, \cdots$ (and hence a corresponding process $X_1, X_2, \cdots$) which will incur small loss initially and then inflict large loss to beat any bound. This is characteristic of arbitrary sampling schemes. However, we will show that bounds can be derived for the cumulative loss for arbitrary sampling. The cumulative NN risk for an arbitrary sequence $\{X_1, X_2, \cdots, X_n\} \subset \mathcal{X}$ is defined as

$$C_n(X_1, \cdots, X_n) = \sum_{i=2}^{n} r_i(X_i, X_i').$$

(Note that we leave out $r_1(X_1, X_1')$ from the sum since X_1' is not defined and we cannot expect a good estimate.) In the i.i.d. case this gives

$$E[C_n(X_1, \cdots, X_n)] = \sum_{i=2}^{n} R_i.$$

Thus the cumulative risk C_n can be used as a common criterion with which to compare the performances of each sampling scheme. The above relationship motivates the definition of the *time-average NN risk* of $X_1, \cdots, X_n$ as

$$\bar{R}_n(X_1, \cdots, X_n) = \frac{1}{n} C_n(X_1, \cdots, X_n).$$

For i.i.d. sampling it can be seen that

$$\lim_{n\to\infty} E[\bar{R}_n(X_1, \ldots, X_n)] = \lim_{n\to\infty} R_n.$$

It will be shown that $\bar{R}_n$ plays the same role in arbitrary sampling as does R_n in i.i.d. sampling.

In [3], two types of loss functions—metric loss and squared-error loss—are considered under i.i.d. sampling. It was shown that under a metric loss

$$\lim_{n\to\infty} R_n \leq 2R^*$$

and under squared error loss

$$\lim_{n\to\infty} R_n = 2R^*.$$

As mentioned, we consider only the squared error loss case. Under i.i.d. sampling, R_n is upper-bounded by $2R^*$ plus a term that goes to zero. We bound this latter term thereby giving a

bound on the convergence rate. We will also prove that for any sequence of samples

$$\lim_{n\to\infty} \sup_{x_1,\cdots,x_n} \bar{R}_n(x_1,\ldots,x_n) = 2R^{w*}.$$

Furthermore, almost surely

$$\lim_{n\to\infty} R_n(X_1,\cdots,X_n) = \lim_{n\to\infty} \frac{2}{n}\sum_{i=2}^{n} r^*(X_i)$$

which is twice the average of the conditional Bayes risks for the sequence. In fact, we will provide finite sample upper bounds for $\bar{R}_n$ in terms of the sum of this average of Bayes risks plus a second term that goes to zero. In the limit $\bar{R}_n$ equals twice the average of the Bayes risks, so in that sense that second term again gives a bound on the convergence rate. It turns out that a similar analysis in the metric loss case leads to similar finite sample upper bounds for $\bar{R}_n$, although they do not lead to convergence rates since we only have an upper bound on

$$\lim_{n\to\infty} \bar{R}_n(X_1,\cdots,X_n).$$

C. Covering Numbers and Supports

We now introduce various topological definitions and properties that will be used in this paper. Define the open ball of radius ϵ about a point $x \in \mathcal{X}$ as

$$\mathcal{B}(x,\epsilon) = \{y \in \mathcal{X} \mid \rho(x,y) < \epsilon\}.$$

These balls are also known as ϵ-balls. We next define the important and well-known notions of covering numbers and metric entropy which characterize the massiveness of a set. Following Kolmogorov [15], these quantities have been extensively studied and used in various applications.

Definition 1: Let A be a subset of metric space $(\mathcal{X},\rho)$. The *metric covering number* $\mathcal{N}(\epsilon) \equiv \mathcal{N}(\epsilon, A, \rho)$ is defined as the smallest number of open balls of radius ϵ that cover the set A. That is

$$\mathcal{N}(\epsilon) = \inf\{k : \exists\, x_1,\cdots,x_k \in \mathcal{X} \text{ s.t. } A \subset \bigcup_{i=1}^{k} \mathcal{B}(x_i,\epsilon)\}.$$

The logarithm of the metric covering number is often referred to as the *metric entropy* or *ϵ-entropy*. A set A is said to be *totally bounded* if $\mathcal{N}(\epsilon, A, \rho) < \infty$ for all $\epsilon > 0$. In particular, every compact set is totally bounded. All totally bounded sets are bounded. Bounded sets in $\mathbf{R}^r$ are totally bounded. Note that as a function of ϵ, $\mathcal{N}(\epsilon, A, \rho)$ is a nonincreasing, piecewise-constant, and right-continuous function. Accordingly, there is no well-defined inverse function. However, we define the following discrete function called the metric covering radius which can be interpreted as a "pseudoinverse" of the metric covering number.

Definition 2: The *metric covering radius* $\mathcal{N}^{-1}(k) \equiv \mathcal{N}^{-1}(k, A, \rho)$ is defined as the smallest radius such that there exists k balls of this radius which cover the set A. That is

$$\mathcal{N}^{-1}(k) = \inf\{\epsilon : \exists\, x_1,\cdots,x_k \in \mathcal{X} \text{ s.t. } A \subset \bigcup_{i=1}^{k} \mathcal{B}(x_i,\epsilon)\}.$$

Note that $\mathcal{N}^{-1}(\cdot)$ is a nonincreasing discrete function of k. In particular, $\mathcal{N}^{-1}(1)$ is the radius of the smallest ball to cover A and is referred to as the radius of A.

Example 1: [15] For any bounded set A in Euclidean r-space, the covering number of A satisfies $\mathcal{N}(\epsilon, A) \le (\beta/\epsilon)^r$ for all $\epsilon \le \beta = \mathcal{N}^{-1}(1, A)$ and the covering radius satisfies $\mathcal{N}^{-1}(n, A) \le \beta/n^{1/r}$. In addition, if A contains an interior point in $\mathbf{R}^r$ then $\mathcal{N}(\epsilon, A) \ge (\beta_1/\epsilon)^r$ for some $\beta_1 > 0$, and $\mathcal{N}^{-1}(n, A) \ge \beta_1/n^{1/r}$.

Note that, in general, the best constants for upper and lower bounds on $\mathcal{N}(\epsilon, A)$ and $\mathcal{N}^{-1}(n, A)$ may be hard to find. In this paper, we are not concerned with finding tight constants and so a conservative value of $\beta = \mathcal{N}^{-1}(1, A)$ (the radius of the set) is certainly valid. Throughout this paper, for simplicity we denote $\mathbf{R}^r$ to be r-space equipped with the Euclidean metric, although results hold as well for any equivalent metric (e.g., ℓ_p induced metrics). All that is required is to have the same order of growth for the covering numbers, although of course the precise constants may depend on the particular metric.

The next lemma simply states that the metric covering radius of a totally bounded set (and hence a compact set) goes to zero as $n \to \infty$. This result will be used in the proofs of subsequent theorems.

Lemma 2: Let A be a totally bounded subset of $(\mathcal{X},\rho)$, then

$$\lim_{n\to\infty} \mathcal{N}^{-1}(n, A, \rho) = 0.$$

Proof: Assume the statement is false. Then since $\mathcal{N}^{-1}(n)$ is nonincreasing, there exists $\epsilon > 0$ such that $\mathcal{N}^{-1}(n) \ge \epsilon$ for all n. But this implies that $\mathcal{N}(\epsilon) \ge n$ for all n, i.e., $\mathcal{N}(\epsilon) = \infty$, which contradicts the fact that A is totally bounded. ∎

We next define the standard notion of the support of a measure (e.g., [7]).

Definition 3: The support of the probability measure μ defined on $(\mathcal{X},\rho)$ is defined as

$$\mathcal{K}(\mu) = \{x \in \mathcal{X} : \forall \epsilon > 0, \mu(\mathcal{B}(x,\epsilon)) > 0\}.$$

III. I.I.D. Sampling

In this section we consider the problem in which the samples $\{X_i\}_{i=1}^n$ are drawn i.i.d. according to some unknown probability measure μ on $(\mathcal{X},\rho)$. This is the classical NN problem for which Cover [3] proved that

$$\lim_{n\to\infty} R_n = 2R^*.$$

The purpose of this section is to find an upper bound on the pointwise-finite sample performance in terms of only the support of μ.

The following lemma which is actually a corollary of Lemma 1 bounds the convergence rate for the expected squared-error loss in terms of the expected nearest neighbor distance.

Lemma 3: Let $X_1, X_2, \cdots, X_n$ be i.i.d. according to a probability measure μ with $\mathcal{K}(\mu)$ a subset of $(\mathcal{X},\rho)$ with $(X_1,Y_1),\cdots,(X_n,Y_n)$ satisfying Assumption 2. For any $F(y \mid x)$ satisfying Assumption 1, we have

$$R_n \le R_\infty + (K_1 + K_2)(Ed_n^2)^\alpha.$$

Proof: Taking expected values on the conclusion of Lemma 1 and using Jensen's inequality since $h(t) = t^\alpha$ is concave for $0 < \alpha \leq 1$, the statement follows by also noting that

$$R_\infty = \lim_{n\to\infty} R_n = 2E[\sigma^2(X)] = 2R_\mu^*. \quad \blacksquare$$

We consider the case in which the support of μ, $\mathcal{K}(\mu)$, is totally bounded. The following theorem provides a bound on Ed_n and Ed_n^2.

Theorem 1: Let $X_1, X_2, \cdots, X_n$ be i.i.d. according to a probability measure μ with $\mathcal{K}(\mu)$ a totally bounded subset of $(\mathcal{X}, \rho)$. Then

$$Ed_n \leq \frac{3}{n}\sum_{i=1}^{n} \mathcal{N}^{-1}(i, \mathcal{K}(\mu))$$

and

$$Ed_n^2 \leq \frac{8}{n}\sum_{i=1}^{n-1} [N^{-1}(i, \mathcal{K}(\mu))]^2.$$

Proof: Observe that for any $X_n \in \mathcal{K}(\mu)$

$$\Pr[d_n > \epsilon \mid X_n] = (1 - \mu(\mathcal{B}(X_n, \epsilon)))^{n-1}.$$

But $\Pr[X_n \in \mathcal{K}(\mu)] = 1$. The proof for this is argued in [4]: the separability of $\mathcal{K}(\mu)$ implies that $[\mathcal{K}(\mu)]^c$ is contained in a countable union of sets of measure zero.

Fix $\epsilon > 0$. Now take an $\epsilon/2$-covering of $\mathcal{K}(\mu)$, $B_1, B_2, \cdots, B_{\mathcal{N}(\epsilon/2)}$. Then for $X_n \in \mathcal{K}(\mu)$, there exists an i such that $B_i \subset \mathcal{B}(X_n, \epsilon)$. Let $N \equiv \mathcal{N}(\epsilon/2)$. Now define an $\epsilon/2$-partition as follows. For each $i = 1, \cdots, N$ let

$$P_i = B_i - \bigcup_{k=1}^{i-1} B_k.$$

Then $P_i \subset B_i$

$$\bigcup_{i=1}^{N} B_i = \bigcup_{i=1}^{N} P_i$$

and $P_i \bigcap P_j = 0$. Also

$$\sum_{i=1}^{N} \mu(P_i) = 1.$$

Then for $X_n \in \mathcal{K}(\mu)$ there exists an i such that $P_i \subset B_i \subset \mathcal{B}(X_n, \epsilon)$ and in turn $p_i \equiv \mu(P_i) \leq \mu(\mathcal{B}(X_n, \epsilon))$. Hence

$$\Pr[d_n > \epsilon \mid X_n \in P_i] \leq (1 - p_i)^{n-1}$$

and

$$\Pr[d_n > \epsilon] \leq \sum_{i=1}^{N} p_i(1 - p_i)^{n-1}.$$

As $d_n \geq 0$, then

$$Ed_n = \int_0^\infty d\epsilon \Pr[d_n > \epsilon].$$

We now bound $\Pr[d_n > \epsilon]$ by bounding

$$\sum_{i=1}^{N} p_i(1 - p_i)^{n-1}$$

for all $\{p_i\}$ such that

$$\sum_{i=1}^{N} p_i = 1.$$

We now prove that

$$\sum_{i=1}^{N} p_i(1 - p_i)^{n-1} \leq \begin{cases} 1, & n \leq N \\ \frac{N}{2n}, & n > N. \end{cases}$$

The case when $n \leq N$ is obvious. If $n > N$, then

$$\begin{aligned} \sum_{i=1}^{N} p_i(1 - p_i)^{n-1} &\leq \sum_{i=1}^{N} \max_{p_i} p_i(1 - p_i)^{n-1} \\ &= \sum_{i=1}^{N} \frac{1}{n}\left(1 - \frac{1}{n}\right)^{n-1} \\ &\leq \frac{N}{2n}. \end{aligned}$$

Hence we have that

$$\begin{aligned} &\Pr[d_n > \epsilon] \\ &\quad \leq \sum_{i=1}^{\mathcal{N}(\epsilon/2)} p_i(1 - p_i)^{n-1} \leq \begin{cases} 1, & n \leq \mathcal{N}(\epsilon/2) \\ \frac{\mathcal{N}(\epsilon/2)}{2n}, & n > \mathcal{N}(\epsilon/2). \end{cases} \end{aligned}$$

Since $\Pr[d_n > \epsilon] = 0$ for $\epsilon > 2\mathcal{N}^{-1}(1)$, we have

$$\begin{aligned} Ed_n &= \int_0^\infty d\epsilon \Pr[d_n > \epsilon] \leq \int_0^{2\mathcal{N}^{-1}(n)} d\epsilon \\ &\quad + \int_{2\mathcal{N}^{-1}(n)}^{2\mathcal{N}^{-1}(1)} d\epsilon \frac{\mathcal{N}(\epsilon/2)}{2n} \\ &= 2\mathcal{N}^{-1}(n) + \frac{1}{n}\int_{\mathcal{N}^{-1}(n)}^{\mathcal{N}^{-1}(1)} d\epsilon \mathcal{N}(\epsilon). \end{aligned}$$

Since $\mathcal{N}(\epsilon) = i$ for $\mathcal{N}^{-1}(i) \leq \epsilon < \mathcal{N}^{-1}(i-1)$ we get

$$\begin{aligned} \int_{\mathcal{N}^{-1}(n)}^{\mathcal{N}^{-1}(1)} d\epsilon \mathcal{N}(\epsilon) &= \sum_{i=2}^{n} \int_{\mathcal{N}^{-1}(i)}^{\mathcal{N}^{-1}(i-1)} d\epsilon \mathcal{N}(\epsilon) \\ &= \sum_{i=2}^{n} i[\mathcal{N}^{-1}(i-1) - \mathcal{N}^{-1}(i)] \\ &= \mathcal{N}^{-1}(1) - n\mathcal{N}^{-1}(n) + \sum_{i=1}^{n-1} \mathcal{N}^{-1}(i). \end{aligned}$$

Hence

$$Ed_n \leq \mathcal{N}^{-1}(n) + \frac{\mathcal{N}^{-1}(1)}{n} + \frac{1}{n}\sum_{i=1}^{n-1} \mathcal{N}^{-1}(i) \leq \frac{3}{n}\sum_{i=1}^{n} \mathcal{N}^{-1}(i). \tag{1}$$

Similarly

$$\begin{aligned} Ed_n^2 &= \int_0^\infty d\epsilon \Pr[d_n^2 > \epsilon] = \int_0^\infty d\epsilon \Pr[d_n > \sqrt{\epsilon}] \\ &= 4(\mathcal{N}^{-1}(n))^2 + \frac{4}{n}\int_{(\mathcal{N}^{-1}(n))^2}^{(\mathcal{N}^{-1}(1))^2} d\epsilon \mathcal{N}(\sqrt{\epsilon}) \end{aligned}$$

and together with

$$\int_{(\mathcal{N}^{-1}(n))^2}^{(\mathcal{N}^{-1}(1))^2} d\epsilon \mathcal{N}(\sqrt{\epsilon}) = [\mathcal{N}^{-1}(1)]^2 - n[\mathcal{N}^{-1}(n)]^2 + \sum_{i=1}^{n-1}[N^{-1}(i)]^2$$

gives

$$Ed_n^2 \le \frac{4[\mathcal{N}^{-1}(1)]^2}{n} + \frac{4}{n}\sum_{i=1}^{n-1}[N^{-1}(i)]^2. \quad (2)$$

■

For example, take $\mathcal{K}(\mu)$, a bounded subset of $\mathbf{R}^r$ for some integer $r > 1$, then

$$E[d_n] \le \frac{3\beta r}{r-1} n^{-1/r}$$

where β is the radius of $\mathcal{K}(\mu)$. It is well known that the support is closed [7], hence supports which are bounded subsets of $\mathbf{R}^r$ are in fact compact. Note that the theorem statement can be tightened by using the bounds in (1) and (2) of the proof.

The following corollary obtains a bound on the finite sample convergence rate for the expected squared-error loss using the results from Theorem 1 applied to Lemma 3.

Corollary 1: Let $X_1, X_2, \cdots, X_n$ be i.i.d. according to a probability measure μ with $\mathcal{K}(\mu)$ a totally bounded subset of $(\mathcal{X}, \rho)$ with $(X_1, Y_1), \cdots, (X_n, Y_n)$ satisfying Assumption 2. For any $F(y \mid x)$ that satisfies Assumption 1 we have

$$R_n \le R_\infty + (K_1 + K_2)\frac{8^\alpha}{n}\sum_{i=1}^{n-1}[\mathcal{N}^{-1}(i, \mathcal{K}(\mu))]^{2\alpha}.$$

For example, if $\mathcal{K}(\mu) \subset R^r (r > 2\alpha)$ then

$$R_n \le R_\infty + (K_1 + K_2)\frac{r(8\beta)^\alpha}{r - 2\alpha} n^{-2\alpha/r}$$

where $\beta = \mathcal{N}^{-1}(1, \mathcal{K}(\mu))$.

IV. Arbitrary Sampling

In the previous section, the sequence $\{X_i\}$ was assumed to be i.i.d. We now consider a formulation in which there are no restrictions on the sequence of random variables $\{X_i\}$. That is, the samples $\{X_i\}$ can be chosen from a completely *arbitrary* random process taking values in a general separable metric space. This includes cases in which $\{X_i\}$ is any nonstationary or nonergodic process. It also allows completely arbitrary deterministic sampling strategies. However, as before, Assumption 2 must hold which implies that given X_i, the label Y_i must be conditionally independent of (X_j, Y_j) for $i \ne j$.

Our proof technique is a deterministic sample path analysis. In fact, our theorems make almost sure (a.s.) statements for any random process which is equivalent to making arbitrary deterministic sample path statements.

Recall that

$$X_n' = \arg\min_{i<n} \rho(X_i, X_n)$$

and that $d_n = \rho(X_n, X_n')$. In effect, each sequence of points $\{X_1, \cdots, X_n\}$ induces a sequence of minimum distances $\{d_2, d_3, \cdots, d_n\}$. Note that d_1 is obviously not defined. The next lemma states that under squared-error loss the cumulative loss bound is in terms of the conditional Bayes risks of the sequence and the nearest neighbor distances.

Lemma 4: Let $X_1, \cdots, X_n$ be an arbitrary random process in $(\mathcal{X}, \rho)$ with $(X_1, Y_1), \cdots, (X_n, Y_n)$ satisfying Assumption 2. For any $F(y \mid x)$ satisfying Assumption 1, we have a.s. that

$$C_n(X_1, \cdots, X_n) \le 2\sum_{i=2}^{n} r^*(X_i) + (K_1 + K_2)\sum_{i=2}^{n} d_i^{2\alpha}.$$

Furthermore, if

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=2}^{n} d_i^{2\alpha} = 0$$

then a.s. we have

$$\lim_{n\to\infty} \bar{R}_n(X_1, \cdots, X_n) = \lim_{n\to\infty} \frac{2}{n}\sum_{i=2}^{n} r^*(X_i)$$

if the limit exists, and

$$\lim_{n\to\infty} \sup_{x_1,\cdots,x_n} \bar{R}_n(x_1, \cdots, x_n) = 2R^{w*}.$$

Proof: Fixing $X_1 = x_1, \cdots, X_n = x_n$, then Lemma 1 and summing terms gives

$$C_n(x_1, \cdots, x_n) \le 2\sum_{i=2}^{n} \sigma^2(x_i) + (K_1 + K_2)\sum_{i=2}^{n} d_i^{2\alpha}.$$

For the second part of the lemma, we have from the proof of Lemma 1 the following equality:

$$\bar{R}_n(x_1, \cdots, x_n) = \frac{1}{n}\sum_{i=2}^{n}\{2\sigma^2(x_i) + \|m(x_i) - m(x_i')\|^2 + [\sigma^2(x_i') - \sigma^2(x_i)]\}.$$

But by hypothesis

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=2}^{n} \|m(x_i) - m(x_i')\|^2 \le \lim_{n\to\infty} \frac{K_1}{n}\sum_{i=2}^{n} d_i^{2\alpha} = 0$$

and

$$\lim_{n\to\infty} \left|\frac{1}{n}\sum_{i=2}^{n}[\sigma^2(x_i) - \sigma^2(x_i')]\right| \le \lim_{n\to\infty} \frac{K_2}{n}\sum_{i=2}^{n} d_i^{2\alpha} = 0$$

which imply that

$$\lim_{n\to\infty} \bar{R}_n(x_1, \cdots, x_n) = \lim_{n\to\infty} \frac{2}{n}\sum_{i=2}^{n} \sigma^2(x_i)$$

if the limit exists. Repeating the above computations but taking sups before the limit gives

$$\lim_{n\to\infty} \sup_{x_1,\cdots,x_n} \bar{R}_n(x_1, \cdots, x_n) = \lim_{n\to\infty} \sup_{x_1,\cdots,x_n} \frac{2}{n} \times \sum_{i=2}^{n} \sigma^2(x_i) = 2R^{w*}.$$ ■

A. Bound on Cumulative Distances

Observe that with arbitrary sampling, the samples must be from a totally bounded subset A of metric space $(\mathcal{X}, \rho)$ to be able to bound the cumulative nearest neighbor distances. We require the total boundedness of A to prevent the points from being too "spread out." In fact, the next theorem quantifies the clustering of nearest neighbor distances of arbitrarily chosen points in totally bounded metric spaces. In some sense this is analogous to the statement that if $\{X_i\}$ is i.i.d., then $d_n \to 0$ a.s. [4].

Theorem 2: Let A be a totally bounded subset of $(\mathcal{X}, \rho)$. Then for an arbitrary sequence $x_1, x_2, \cdots, x_n \in A$ and for any $\gamma > 0$

$$\sum_{i=2}^{n} d_i^\gamma \le \sum_{i=1}^{n-1} [2\mathcal{N}^{-1}(i, A)]^\gamma$$

where $d_i = \rho(x_i, x_i')$. In particular, for bounded subsets of $\mathbf{R}^r$ with radius $\beta > 0$

$$\sum_{i=2}^{n} d_i^\gamma \le \begin{cases} \frac{(2\beta)^\gamma}{r-\gamma}[rn^{1-\gamma/r} - \gamma], & \gamma < r \\ (2\beta)^\gamma[1 + \log n], & \gamma = r. \end{cases}$$

Proof: We prove the result by induction. The inequality is true for $n = 2$. Certainly $d_2 \le 2\mathcal{N}^{-1}(1, A)$ which is twice the radius of A. Now assume it holds for some $n > 2$, and fix a sequence $\{x_i\}_{i=1}^{n+1}$. The sequence induces a sequence of nearest neighbor distances $\{d_i\}_{i=2}^{n+1}$. Let

$$d_m = \min_{2 \le i \le n+1} d_i$$

where $m = \arg\min d_i$. We now show that $d_m \le 2\mathcal{N}^{-1}(n, A)$. First note that

$$d_m = \min_{1 \le i,j \le n+1} \rho(x_i, x_j).$$

Cover A with n balls of radius $\mathcal{N}^{-1}(n, A)$. Then at least one ball contains two points. Hence $d_m \le 2\mathcal{N}^{-1}(n, A)$.

Now define the new sequence

$$a_i = \begin{cases} x_i, & i = 1, \cdots, m-1 \\ x_{i+1}, & i = m, \cdots, n. \end{cases}$$

Then $\{a_i\}_{i=1}^{n}$ induces nearest neighbor distances $\{b_i\}_{i=2}^{n}$ where

$$\begin{aligned} b_i &= d_i, \quad i = 2, \cdots, m-1 \\ b_i &\ge d_{i+1}, \quad i = m, \cdots, n. \end{aligned}$$

That is, by removing the mth point we can only increase all the subsequent nearest neighbor distances. But by the induction hypothesis

$$\sum_{i=2}^{n} b_i^\gamma \le \sum_{i=1}^{n-1} [2\mathcal{N}^{-1}(i, A)]^\gamma.$$

Hence

$$\begin{aligned} \sum_{i=2}^{n+1} d_i^\gamma &\le \sum_{i=2}^{n} b_i^\gamma + d_m^\gamma \\ &\le \sum_{i=1}^{n-1} [2\mathcal{N}^{-1}(i, A)]^\gamma + [2\mathcal{N}^{-1}(n, A)]^\gamma \\ &= \sum_{i=1}^{n} [2\mathcal{N}^{-1}(i, A)]^\gamma. \end{aligned}$$

In particular for bounded $A \subset \mathbf{R}^r$ we have from Example 1 that $\mathcal{N}^{-1}(k, A) = \beta/k^{1/r}$ thus

$$\sum_{k=2}^{n} d_k^\gamma \le \sum_{k=1}^{n-1} \frac{\beta^\gamma}{k^{\gamma/r}} = \begin{cases} O(\log n), & \gamma = r \\ O(n^{1-\gamma/r}), & \gamma < r. \end{cases} \quad \blacksquare$$

In fact, Theorem 2 can be generalized so that for any increasing function g such that $g(0) = 0$ we have

$$\sum_{i=1}^{n} g(d_i) \le \sum_{i=1}^{n-1} g(2\mathcal{N}^{-1}(i, A)).$$

The following lower bound demonstrates that the established upper bound is tight up to a factor of 2.

Proposition 1: Let A be a totally bounded subset of $(\mathcal{X}, \rho)$. For every $\delta > 0$ there exists a sequence $x_1, x_2, \cdots, x_n \in A$ such that

$$\sum_{i=2}^{n} d_i \ge \sum_{i=1}^{n-1} \mathcal{N}^{-1}(i, A) - \delta.$$

Proof: Fix a positive sequence $\{\delta_i\}$ such that

$$\sum_{i=1}^{\infty} \delta_i \le \delta$$

for some $\delta > 0$. Let $\epsilon_i = \mathcal{N}^{-1}(i, A)$ for each $i = 1, \cdots, n$. Fix $x_1 \in A$ arbitrarily. Choose x_2 from

$$A \cap [\mathcal{B}(x_1, \epsilon_1 - \delta_1)]^c.$$

Existence is ensured from the definition of $\mathcal{N}^{-1}(\cdot)$. Then $d_2 \ge \epsilon_1 - \delta_1$. Similarly, choosing x_n from

$$A \cap \left[\bigcup_{i=1}^{n-1} \mathcal{B}(x_i, \epsilon_{n-1} - \delta_{n-1}) \right]^c$$

gives $d_n \ge \epsilon_{n-1} - \delta_{n-1}$. Hence

$$\sum_{i=2}^{n} d_i \ge \sum_{i=1}^{n-1} \mathcal{N}^{-1}(i, A) - \delta. \quad \blacksquare$$

The next corollary states that arbitrary sampling in a totally bounded set under squared-error loss has cumulative loss in terms of the conditional Bayes risks of the sequence and the metric covering radii. In particular, the asymptotic time-average of the risk equals twice the time-average of the conditional Bayes risks of the fixed sequence.

Corollary 2: Let $X_1, \cdots, X_n$ be an arbitrary random process in a totally bounded subset A of $(\mathcal{X}, \rho)$ with $(X_1, Y_1), \cdots, (X_n, Y_n)$ satisfying Assumption 2. For any $F(y \mid x)$ satisfying Assumption 1, we have a.s.

$$C_n(X_1, \cdots, X_n) \leq 2\sum_{i=2}^{n} r^*(X_i) + (K_1 + K_2)\sum_{i=1}^{n-1} [2\mathcal{N}^{-1}(i, A)]^{2\alpha}.$$

Furthermore

$$\lim_{n\to\infty} \bar{R}_n(X_1, \cdots, X_n) = \lim_{n\to\infty} \frac{2}{n}\sum_{i=2}^{n} r^*(X_i) \quad \text{a.s.}$$

if the limit exists, and

$$\lim_{n\to\infty} \sup_{x_1,\ldots,x_n} \bar{R}_n(x_1, \cdots, x_n) = 2R^{w*}.$$

Proof: Fix $X_1 = x_1, \cdots, X_n = x_n \in A$. The first statement follows from Lemma 4 and Theorem 2. The second statement follows from Lemma 4 as the hypothesis is satisfied using Theorem 2 and Lemma 2

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=2}^{n} d_i^{2\alpha} \leq \lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n-1} [2\mathcal{N}^{-1}(i, A)]^{2\alpha} = 0. \quad \blacksquare$$

Interestingly, this suggests that to get large nearest neighbor distances (and hence large cumulative error) it suffices to use some i.i.d. process rather than a non-i.i.d. sampling technique or some deterministic method.

B. Recovering and Extending i.i.d. Sampling Results

We have considered NN estimation under i.i.d. sampling and arbitrary sampling using different techniques. The results in Section III provide pointwise convergence rates for i.i.d. sampling in totally bounded sets. The results in Section IV-A give bounds on the time-averaged risk for an arbitrary sequence of samples in a totally bounded set. An immediate consequence of Section IV-A is time-averaged bounds for i.i.d. sampling. In this section we prove that the results in Section IV-A imply the traditional pointwise results for i.i.d. sampling and in fact we extend the traditional results to stationary sampling.

Lemma 5: Let $X_1, X_2, \cdots, X_n$ be a stationary process taking values in $(\mathcal{X}, \rho)$. Then $Ed_n \leq Ed_i$ for every $2 \leq i \leq n$.

Proof: Fix $0 \leq i < n - 2$. Then

$$\begin{aligned} Ed_n &= E \min_{2\leq j<n} \rho(X_j, X_n) \leq E \min_{2+i\leq j<n} \rho(X_j, X_n) \\ &= E \min_{2\leq j<n-i} \rho(X_j, X_{n-i}) \\ &= Ed_{n-i}. \end{aligned}$$

The second to last equality follows from the stationary hypothesis. ■

Theorem 3: Let $X_1, X_2, \cdots, X_n$ be a stationary process taking values in a totally bounded subset A of $(\mathcal{X}, \rho)$. Then

$$Ed_n \leq \frac{2}{n-1}\sum_{i=1}^{n-1} \mathcal{N}^{-1}(i, A).$$

For a bounded subset A of $\mathbb{R}^r$ and $\beta = \mathcal{N}^{-1}(1, A)$ we have

$$Ed_n \leq \begin{cases} 4\beta\dfrac{1+\log n}{n}, & r = 1 \\ 8\beta n^{-1/r}, & r > 1. \end{cases}$$

Proof: First note that

$$(n-1)E\left[d_n - \frac{1}{n-1}\sum_{i=2}^{n} d_i\right] = E\left[\sum_{i=2}^{n}(d_n - d_i)\right] \leq 0$$

by Lemma 5. Hence

$$Ed_n \leq \frac{1}{n-1}E\sum_{i=2}^{n} d_i.$$

But by Theorem 2 we have that for every sequence $X_1, \cdots, X_n \in A$, a.s. the induced nearest neighbor distances satisfy

$$\sum_{i=2}^{n} d_i \leq 2\sum_{i=1}^{n-1} \mathcal{N}^{-1}(i, A).$$

This implies the theorem statement. ■

Note that in the case of totally bounded support, this theorem implies the results in [13] for $\mathbb{R}^r$ (namely, that $a_n n^{1/r} d_n \to 0$ in probability for any sequence $a_n \to 0$) and extends the results to separable metric spaces and to stationary processes.

V. k_n-NN Estimation Under Arbitrary Sampling

It is well known that k_n-NN estimators are universally consistent under i.i.d. sampling [19]. The purpose of this section is to prove the consistency of the k_n-NN estimator under arbitrary sampling, provide bounds on the convergence rates, and recover the i.i.d. sampling results. In fact, we will find an upper bound on the convergence rate that coincides with known rates for i.i.d. sampling.

The k_n-nearest neighbor rule is defined as follows. Let k_n be any nondecreasing sequence of numbers. (Assume $k_2 \geq 1$ for simplicity.) Denote the k_n nearest neighbors of X_n from the set $\{X_1, \cdots, X_{n-1}\}$ as $X_n^{[1]}, \cdots, X_n^{[k_n]}$ where $X_n^{[1]}$ is the nearest. We denote $Y_n^{[i]}$ as the label associated with the ith nearest neighbor. The k_n-NN rule estimate is the average of the k_n NN labels

$$\bar{Y}_n^{(k_n)} = \frac{1}{k_n}\sum_{i=1}^{k_n} Y_n^{[i]}.$$

Let $r_n^{(k_n)}(X_n, X_n^{[1]}, \cdots, X_n^{[k_n]})$ be the conditional expected loss in estimating Y_n given X_n and the k_n nearest neighbors of X_n. We then define $C_n^{(k_n)}(X_1, \cdots, X_n)$ as the cumulative k_n-NN loss and $\bar{R}_n^{(k_n)}(X_1, \cdots, X_n)$ as the time-averaged risk of a given arbitrary sequence. Let $d_n(i)$ be the ith nearest neighbor

distance of n samples to X_n, i.e., $d_n(i) = \rho(X_n, X_n^{[i]})$. The next lemma, an extension of Lemma 4, gives results for the k_n-NN rule.

Lemma 6: Let $X_1, \cdots, X_n$ be an arbitrary random process in $(\mathcal{X}, \rho)$ with $(X_1, Y_1), \cdots, (X_n, Y_n)$ satisfying Assumption 2. For any $F(y \mid x)$ satisfying Assumption 1, we have that any k_n-nearest neighbor rule gives a.s.

$$C_n^{(k_n)}(X_1, \cdots X_n) \le \sum_{i=2}^{n} \left(1 + \frac{1}{k_i}\right) r^*(X_i) + (K_1 + K_2) \sum_{i=2}^{n} d_i^{2\alpha}(k_i).$$

Furthermore, if $\{k_n\}$ satisfies

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=2}^{n} d_i^{2\alpha}(k_i) = 0$$

then we have that a.s.

$$\lim_{n\to\infty} \bar{R}_n^{(k_n)}(X_1, \cdots, X_n) = \lim_{n\to\infty} \left(1 + \frac{1}{k_n}\right) \lim_{n\to\infty} \frac{1}{n} \sum_{i=2}^{n} r^*(X_i) \quad (3)$$

if the limit exists, and

$$\lim_{n\to\infty} \sup_{x_1,\cdots,x_n} \bar{R}_n^{(k_n)}(x_1, \cdots, x_n) = \lim_{n\to\infty} \left(1 + \frac{1}{k_n}\right) R^{w*}. \quad (4)$$

Proof: Fix $X_1 = x_1, \cdots, X_n = x_n$. As in Lemma 4, under Assumptions 1 and 2 it easy to show that

$$C_n^{(k_n)}(x_1, \cdots, x_n) = \sum_{i=2}^{n} \left\{ \left(1 + \frac{1}{k_i}\right) \sigma^2(x_i) + \frac{1}{k_i^2} \left\| \sum_{j=1}^{k_i} \left[m(x_i) - m\left(x_i^{[j]}\right) \right] \right\|^2 + \frac{1}{k_i^2} \sum_{j=1}^{k_i} \left[\sigma^2\left(x_i^{[j]}\right) - \sigma^2(x_i) \right] \right\}.$$

We see that

$$\sum_{i=2}^{n} \frac{1}{k_i^2} \left\| \sum_{j=1}^{k_i} m(x_i) - m\left(x_i^{[j]}\right) \right\|^2 \le \sum_{i=2}^{n} \frac{K_1}{k_i} \sum_{j=1}^{k_i} d_i^{2\alpha}(j) \le K_1 \sum_{i=2}^{n} d_i^{2\alpha}(k_i).$$

A similar computation on the variance term gives the first statement in the theorem.

From the hypotheses in the second half of the theorem we can argue as follows. Noting that

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=2}^{n} \frac{1}{k_i^2} \left\| \sum_{j=1}^{k_i} m(x_i) - m(x_i^{[j]}) \right\|^2 \le \lim_{n\to\infty} \frac{K_1}{n} \sum_{i=2}^{n} d_i^{2\alpha}(k_i) = 0,$$

and a similar computation on the variance terms gives

$$\lim_{n\to\infty} \bar{R}_n^{(k_n)}(x_1, \cdots, x_n) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=2}^{n} \left(1 + \frac{1}{k_i}\right) \sigma^2(x_i) = \lim_{n\to\infty} \left(1 + \frac{1}{k_n}\right) \lim_{n\to\infty} \frac{1}{n} \sum_{i=2}^{n} r^*(x_i)$$

if the limit exists since $r^*(\cdot)$ is bounded and k_n is nondecreasing.

By repeating the above computations but with taking the supremum over all sequences prior to the limit taking, we get

$$\lim_{n\to\infty} \sup_{x_1,\cdots,x_n} \bar{R}_n^{k_n}(x_1, \cdots, x_n) = \lim_{n\to\infty} \sup_{x_1,\cdots,x_n} \frac{1}{n} \sum_{i=2}^{n} \left(1 + \frac{1}{k_i}\right) \sigma^2(x_i) = R^{w*} \left(1 + \lim_{n\to\infty} \frac{1}{n} \sum_{i=2}^{n} \frac{1}{k_i}\right) = R^{w*} \lim_{n\to\infty} \left(1 + \frac{1}{k_n}\right),$$

since k_n is nondecreasing. ■

A. Bound on Cumulative Distances

We first consider the following lemma which is an extension of Theorem 2.

Lemma 7: For any $\gamma > 0$, for any $n \ge k + 1 \ge 2$, and for any sequence $x_1, x_2, \cdots, x_n$ in a totally bounded subset A of $(\mathcal{X}, \rho)$

$$\sum_{i=k+1}^{n} d_i^{\gamma}(k) \le \sum_{i=k}^{n-1} [2\mathcal{N}^{-1}(\lfloor i/k \rfloor, A)]^{\gamma}$$

where $d_n(i) = \rho(x_n, x_n^{[i]})$.

Proof: The result follows by a modification of Theorem 2 by observing that if we cover A with $\lfloor n/k \rfloor$ balls of radius $\mathcal{N}^{-1}(\lfloor n/k \rfloor, A)$, then with $n + 1$ points at least one ball will have $k + 1$ points. This implies that

$$d_m \equiv \min_{i \le n+1} d_i(k) \le 2\mathcal{N}^{-1}(\lfloor n/k \rfloor, A).$$

Define a new sequence by omitting x_m. Induction completes the argument much the same as in the proof of Theorem 2. ■

The next theorem shows that if k_n grows slow enough then the k_n-NN rule is consistent in estimation.

Theorem 4: Let $X_1, \cdots, X_n$ be an arbitrary random process in a totally bounded subset A of $(\mathcal{X}, \rho)$ with $(X_1, Y_1), \cdots, (X_n, Y_n)$ satisfying Assumption 2. For any $F(y \mid x)$ satisfying Assumption 1, we have that any k_n-nearest neighbor rule satisfies a.s.

$$C_n^{(k_n)}(X_1, \cdots, X_n) \le \sum_{i=2}^{n} \left(1 + \frac{1}{k_i}\right) r^*(X_i) + (K_1 + K_2) \times \sum_{i=k_n}^{n-1} [2\mathcal{N}^{-1}(\lfloor i/k_n \rfloor, A)]^{2\alpha}.$$

Furthermore, if $\{k_n\}$ satisfies the following two conditions:

$$\text{(C1)} \quad k_n \to \infty, \quad \text{as } n \to \infty$$
$$\text{(C2)} \quad \frac{k_n}{n} \to 0, \quad \text{as } n \to \infty$$

then we have that a.s.

$$\lim_{n\to\infty} \bar{R}_n^{(k_n)}(X_1, \cdots, X_n) = \lim_{n\to\infty} \frac{1}{n} \sum_{i=2}^{n} r^*(X_i) \tag{5}$$

when the limit exists. Furthermore,

$$\lim_{n\to\infty} \sup_{x_1,\cdots,x_n} \bar{R}_n^{(k_n)}(x_1, \cdots, x_n) = R^{w*}. \tag{6}$$

Proof: The first statement follows using Lemmas 6 and 7. Next, define a nonincreasing function f such that $f(i) = \mathcal{N}^{-1}(i)$ for all positive integers. Then

$$\frac{1}{n} \sum_{i=k_n}^{n-1} \mathcal{N}^{-1}(\lfloor i/k_n \rfloor)^{2\alpha} \le \frac{f(1)^{2\alpha}}{n} + \frac{1}{n} \int_{k_n}^{n} f(x/k_n)^{2\alpha} dx$$
$$\le \frac{f(1)^{2\alpha}}{n} + \frac{k_n}{n} \int_{1}^{n/k_n} f(x)^{2\alpha} dx.$$

The limit as $n \to \infty$ goes to zero using Condition (C2) and Lemma 2. The hypothesis in Lemma 6 is then satisfied and the result then follows. ∎

Corollary 3: Let A be a bounded subset of $\mathbf{R}^r$. There exists a k_n-NN rule satisfying (C1) and (C2) such that for an arbitrary random process $X_1, \cdots, X_n \in A$ with $(X_1, Y_1), \cdots, (X_n, Y_n)$ satisfying Assumption 2, and for $F(y \mid x)$ satisfying Assumption 1, we have the convergence rate $O(n^{-\frac{2\alpha}{r+2\alpha}})$, if $r > 2\alpha$.

Proof: From Example 1, $\mathcal{N}^{-1}(n, A) \le \beta/n^{1/r}$ for $\beta = \mathcal{N}^{-1}(1, A)$. Now compute the second term in the cumulative bound given in Theorem 4. Let $k_n = n^t$ for some $t > 0$, then

$$\sum_{i=k_n}^{n-1} [2\mathcal{N}^{-1}(\lfloor i/k_n \rfloor)]^{2\alpha} \le (2\beta)^{2\alpha} n^{2\alpha t/r} \sum_{i=k_n}^{n-1} i^{-2\alpha/r}$$
$$\le (2\beta)^{2\alpha} \frac{r}{r-2\alpha} n^{1+2\alpha(t-1)/r}.$$

Hence the finite sample cumulative bound is

$$O\left(n^{1+(t-1)2\alpha/r} + n^{1-t}\right).$$

The best choice is $t = \frac{2\alpha}{2\alpha+r}$. This gives a cumulative bound as $O(n^{\frac{r}{r+2\alpha}})$. Hence the convergence rate of the k_n-NN rule is $O(n^{-\frac{2\alpha}{r+2\alpha}})$. ∎

Note that the rate obtained Corollary 3 for arbitrary sampling is the rate obtained in [13]. sampling with Lipschitz assumptions on the conditional distribution. In the following theorem we recover the k_n-NN i.i.d. sampling problem and extend the result to stationary processes and separable metric spaces.

Theorem 5: Let $X_1, X_2, \cdots, X_n$ be a stationary process taking values in a totally bounded subset A of $(\mathcal{X}, \rho)$. Then

$$Ed_n(k_n) \le \frac{1}{n-k_n} \sum_{i=k_n}^{n-1} 2\mathcal{N}^{-1}(\lfloor i/k_n \rfloor, A)$$

for any $k_n \ge 1$ such that $k_n/n \to 0$. In particular, for a bounded subset A of $\mathbf{R}^r$ with $\beta = \mathcal{N}^{-1}(1, A)$, we have

$$Ed_n(k_n) \le 8\beta \left(\frac{k_n}{n}\right)^{1/r}.$$

Proof: Fix $k \ge 1$. As in Lemma 5, it is easy to show that $Ed_n(k) \le Ed_i(k)$ for all $k+1 \le i \le n$. By summing and dividing by $n-k$ we get that

$$Ed_n(k) \le \frac{1}{n-k} \sum_{i=k+1}^{n} Ed_i(k).$$

The theorem statement follows from Lemma 7. ∎

Note that the special case of $\mathbf{R}^r$ in the theorem implies the results in [13] and extends them to stationary processes. In addition, Theorem 5 provides new results for separable metric spaces.

VI. Unbounded Support

All of our i.i.d. sampling results obtained in the previous sections explicitly required that the samples be drawn from a probability measure with totally bounded support. The convergence rates depended on the metric covering numbers of the support. We will now consider the case in which the support is not necessarily totally bounded but is σ-compact. ($\mathcal{X}$ is σ-compact if $\mathcal{X}$ is a countable union of compact sets, i.e., if there exists a sequence of compact subsets, $\{A_1, A_2, \cdots\}$, such that $\mathcal{X} = \cup_i A_i$.) For example, this condition is always true whenever $(\mathcal{X}, \rho)$ is a complete separable metric space (e.g., $\mathbf{R}^r$), as it is well known that the support of a probability measure in such a metric space is σ-compact [18, p. 29].

The following theorem states that almost every i.i.d. sequence will eventually have a pointwise bound on the expected NN distance in terms of the covering numbers of sets that depend on the tail of μ.

Theorem 6: Let $X_1, X_2, \cdots$ be i.i.d. according to a probability measure μ with $\mathcal{K}(\mu)$ a σ-compact subset of $(\mathcal{X}, \rho)$. Let $A_1 \subset A_2 \subset \cdots \subset \mathcal{X}$ be such that the A_i are compact

$$\mathcal{K}(\mu) \subset \bigcup_{i=1}^{\infty} A_i$$

and

$$\sum_{i=1}^{\infty} \mu(A_i^c) < \infty$$

and let $k_n \ge 1$ satisfy $k_n/n \to 0$. Then for almost every $\omega = x_1, x_2, \cdots$, there exists $N(\omega) > 0$ such that for all $n > N(\omega)$

$$Ed_n(k_n) \le \frac{2}{n-k_n} \sum_{i=k_n}^{n-1} \mathcal{N}^{-1}(\lfloor i/k_n \rfloor, A_n).$$

Proof: As the $\{A_i\}$ are compact, they are in turn totally bounded. By the Borel–Cantelli lemma, almost surely there exists $N_1(\omega) > 0$ such that for all $n > N_1(\omega)$, $x_n \in A_n$. As the $\{A_i\}$ are sequentially embedded, $x_i \in A_n$ for all $N_1(\omega) < i \leq n$. Furthermore, there exists $N_2(\omega) > 0$ such that $x_i \in A_{N_2}$ for all $i \leq N_1(\omega)$. Letting

$$N(\omega) = \max(N_1(\omega), N_2(\omega))$$

we have that for all $n > N(\omega)$, $x_i \in A_n$ for all $i \leq n$. Hence

$$\begin{aligned} Ed_n(k_n) &\leq \frac{1}{n-k_n} \sum_{i=k_n+1}^{n} Ed_i(k_n) \\ &\leq \frac{2}{n-k_n} \sum_{i=k_n}^{n-1} \mathcal{N}^{-1}(\lfloor i/k_n \rfloor, A_n) \end{aligned}$$

by Theorem 5. ∎

For example, take $\mathcal{K}(\mu) = \mathcal{X} = R^r$. The bound in the theorem states that for large enough n

$$Ed_n(k_n) \leq c\mathcal{N}^{-1}(1, A_n)(k_n/n)^{1/r}$$

for some known constant $c > 0$. However, the sequence $\{A_i\}$ must grow to satisfy the theorem hypothesis. For example, for exponentially decaying tails (e.g., Gaussian distributions) this leads to an additional logarithmic term over the rates for compact support. For geometric tails, our upper bounds on the convergence rate have a power law decay which is strictly slower than that for compact support. In fact, it can be shown that these upper bounds are fairly tight. This example illustrates that the expected NN distance depends critically on the tails of the distribution. The following theorem shows that this is not the case for convergence in probability. In fact, it is a recovery of the result in [13] and an extension to separable metric spaces.

Theorem 7: Let $X_1, X_2, \cdots, X_n$ be i.i.d. according to a probability measure μ with $\mathcal{K}(\mu)$ a σ-compact subset of $(\mathcal{X}, \rho)$. For any $a_n \to 0$, $k_n \geq 1$ with $k_n/n \to 0$, and any compact $\{A_i\}$ such that

$$\mathcal{K}(\mu) \subset \bigcup_{i=1}^{\infty} A_i$$

and $\mu(A_n^c) \to 0$, we have

$$\frac{a_n d_n(k_n)}{\frac{1}{n}\sum_{i=k_n}^{n-1} \mathcal{N}^{-1}(\lfloor i/k_n \rfloor, A_n)} \to 0 \quad \text{in probability.}$$

Proof: Fix $\epsilon > 0$. Let

$$\delta_n = \frac{1}{n} \sum_{i=k_n}^{n-1} \mathcal{N}^{-1}(\lfloor i/k_n \rfloor, A_n).$$

Next, let the number of X_i's that land in A_n until time $n-1$ be denoted as M_n. Then

$$\begin{aligned} &\Pr(a_n d_n(k_n)/\delta_n > \epsilon) \\ &\quad = \Pr(a_n d_n(k_n)/\delta_n > \epsilon \mid X_n \in A_n, M_n \geq n/2) \\ &\qquad \times \Pr(X_n \in A_n, M_n \geq n/2) \\ &\qquad + \Pr(a_n d_n(k_n)/\delta_n > \epsilon \mid X_n \in A_n^c \text{ or } M_n < n/2) \\ &\qquad \times \Pr(X_n \in A_n^c \text{ or } M_n < n/2) \\ &\quad \leq \Pr(a_n d_n(k_n)/\delta_n > \epsilon \mid X_n \in A_n, M_n \geq n/2) \\ &\qquad + \Pr(A_n^c) + \Pr(M_n < n/2). \end{aligned}$$

For n large enough so that $k_n/n \leq 1/2$ and using Markov's inequality and Theorem 5 we have

$$\begin{aligned} &\Pr(a_n d_n(k_n)/\delta_n > \epsilon \mid X_n \in A_n, M_n \geq n/2) \\ &\qquad \leq \frac{a_n E[d_n(k_n) \mid X_n \in A_n, M_n \geq n/2]}{\epsilon \delta_n} \\ &\qquad \leq \frac{4a_n}{\epsilon \delta_n n} \sum_{i=k_n}^{n/2-1} \mathcal{N}^{-1}(\lfloor i/k_n \rfloor, A_n) \\ &\qquad \leq \frac{4a_n}{\epsilon}. \end{aligned}$$

Noting that $\Pr(M_n < n/2) \leq \Pr(A_n^c)$, we have that

$$\Pr(a_n d_n(k_n)/\delta_n > \epsilon) \leq \frac{4a_n}{\epsilon} + 2\Pr(A_n^c) \to 0. \quad \blacksquare$$

VII. Conclusion

We have provided rates of convergence for nearest neighbor estimation in two settings. Under i.i.d. sampling in a totally bounded set of a separable metric space, we showed that the convergence rate is in terms of the covering numbers of the underlying sampling set. We then introduced the notion of arbitrary sampling in a totally bounded set. Although pointwise bounds on the NN estimator under arbitrary sampling are not possible, we showed that cumulative bounds are obtainable and are again in terms of the covering numbers of the sampling set. We also showed the consistency of the k_n-NN estimator under arbitrary sampling and bounded its convergence rate. Finally, we connect the i.i.d. and arbitrary sampling problems by proving some classical and new i.i.d. sampling results from our more general framework.

Acknowledgment

The authors wish to thank A. R. Barron, L. Györfi, and the anonymous referees for helpful suggestions.

References

[1] J. Beck, "The exponential rate of convergence of error for k_n-NN nonparametric regression and decision," *Prob. Contr. Inform. Theory*, vol. 8, pp. 303–311, 1979.
[2] P. E. Cheng, "Strong consistency of nearest neighbor regression function estimators," *J. Multivariate Anal.*, vol. 15, pp. 63–72, 1984.
[3] T. M. Cover, "Estimation by the nearest neighbor rule," *IEEE Trans. Inform. Theory*, vol. IT-14, pp. 50–55, Jan. 1968.
[4] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-13, pp. 21–27, Jan. 1967.
[5] T. M. Cover, "Rates of convergence for nearest neighbor procedures," in *Proc. 1st Ann. Hawaii Conf. on Systems Theory*, Jan. 1968, pp. 413–415.
[6] ——, "Universal portfolios," *Math. Finance*, vol. 1, pp. 1–29, Jan. 1991.
[7] L. P. Devroye, "The uniform convergence of nearest neighbor regression function estimators and their application in optimization," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 142–151, Mar. 1978.
[8] ——, "On the almost everywhere convergence of nonparametric regression function estimates," *Ann. Stat.*, vol. 9, 1981.
[9] ——, "Necessary and sufficient conditions for the pointwise convergence of nearest neighbor regression function estimates," *Z. Wahrscheinlichkeitstheorie verw. Gebiete*, vol. 61, pp. 467–481, 1982.

[10] M. Feder, N. Merhav, and M. Gutman, "Universal prediction of individual sequences," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1258–1270, July 1992.

[11] E. Fix and J. Hodges Jr., "Discriminatory analysis, nonparametric discrimination. I. Consistency properties, " USAF School of Aviation Medicine, Texas Project 21-49-004, Rep. 4, Contract AF41(128)-31, 1951.

[12] J. Fritz, "Distribution-free exponential error bound for nearest neighbor pattern classification," *IEEE Trans. Inform. Theory*, vol. IT-21, pp. 552–557, Sept. 1975.

[13] L. Györfi, "The rate of convergence of k_{nr}NN regression estimates and classification rules," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 362–364, May 1981.

[14] ——, "Recent results on nonparametric regression estimate and multiple classification," *Prob. Contr. Inform. Theory*, vol. 10, pp. 43–52, 1981.

[15] A. N. Kolmogorov and V. M. Tihomirov, "ϵ-entropy and ϵ-capacity of sets in functional spaces," *Amer. Math. Soc. Translations*, ser. 2, vol. 17, pp. 277–364, 1961.

[16] 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.

[17] Y. P. Mack, "Local properties of k-NN regression estimates," *SIAM J. Alg. Disc. Meth.*, vol. 2, pp. 311–323, 1981.

[18] K. R. Parthasarathy, *Probability Measures on Metric Spaces*. San Diego, CA: Academic Press, 1967.

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

[20] W. Stute, "Asymptotic normality of nearest neighbor regression function estimates," *Ann. Stat.*, vol. 12, pp. 917–926, 1984.

[21] T. J. Wagner, "Convergence of the nearest neighbor rule," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 566–571, Sept. 1971.