The Uniform Convergence of Nearest Neighbor Regression Function Estimators and Their Application in Optimization

LUC P. DEVROYE

***Abstract*—A class of nonparametric regression function estimates generalizing the nearest neighbor estimate of Cover [12] is presented. Under various noise conditions, it is shown that the estimates are strongly uniformly consistent. The uniform convergence of the estimates can be exploited to design a simple random search algorithm for the global minimization of the regression function.**

I. Introduction

The problem of optimizing a function Q with respect to $x \in R^d$ arises frequently in the synthesis of complex systems. Often the optimization problem cannot be solved by analytical methods because the mathematical description of the function is unknown or extremely complicated. However, in many cases, the value of the function can be determined with a certain accuracy for any given value of x. It is known that in such situations, random search can be successfully used (for a review of the literature, see [1]–[3]). Two large classes of random optimization techniques can be distinguished, the nonsequential methods and the sequential methods. The most primitive nonsequential method is the crude search [4], where one lets the estimate of the minimum of Q be the best X_i among a sequence $X_1, \cdots, X_n$ of independent random vectors, uniformly distributed over the set B of R^d in which the minimum is sought. If $Q(X_i)$ can be exactly determined, then the value of Q at the estimate will approximate the (essential) infimum of Q on B as n grows large. However, if for every x, only noisy estimates $Y_1, Y_2, \cdots$, (independent identically distributed random variables with distribution function F_x) can be obtained, then one could estimate the regression function

$$Q(x) = \int y \, dF_x(y),$$

which is assumed to exist for all x, by the average

$$Q_\lambda(x) = \lambda^{-1} \sum_{i=1}^{\lambda} Y_i.$$

Upon computing such a λ-average for every X_i, it is logical to define the best X_i as the one with the lowest value $Q_\lambda(X_i)$. Again, we can expect that the true value of Q at this best X_i is close to the minimal possible value of Q if λ and n are large enough.

It is not unreasonable in most applications to assume that Q is well behaved (smooth, continuous) so that information about $Q(x)$ can be gathered from the values of $Q(y)$ for all y with $\|y - x\|$ small. If we construct an estimate q_n of Q (q_n is a function of x and of the X_i, $Q_\lambda(X_i)$, $1 \leq i \leq n$) and then minimize q_n, chances are, in view of the smoothness assumption, that $Q(X_n^*)$, the value of Q at X_n^*, the minimum of q_n, is close to the extreme value. One such multiple trial estimate (multiple, because $\lambda > 1$) is the one that lets $q_n(x) = Q_\lambda(X_1^x)$ where X_1^x is the nearest neighbor to x among $X_1, \cdots, X_n$.

Often the cost of obtaining the Y_i is very high, so that the said crude search method, or its modification using a multiple trial estimate, is not economical since both use λn measurements. If we let $\lambda = 1$ and $Q_\lambda(X_i) = Y_i$ (so that $(X_1, Y_1), \cdots, (X_n, Y_n)$ are independent and identically distributed), then we can only hope to satisfactorily recover Q if Q is "almost continuous." To illustrate this, let us briefly review some history of single trial estimates.

Estimates that first partition B up into a grid and then let q_n be constant on each rectangle in the partition have been suggested by McMurtry and Fu [5], Hill [6], Jarvis [7], and others in multimodal optimization theory. If $\{B_1, \cdots, B_N\}$ is the partition, then, for all x in B_i, this histogram estimate uses

$$q_n(x) = \sum_{j=1}^{n} Y_j I_{\{X_j \in B_i\}} \Big/ \sum_{j=1}^{n} I_{\{X_j \in B_i\}}$$

where I is the indicator function. Of course, unless $n \to \infty$, $N \to \infty$, and all the B_i shrink in size, there is no hope or guarantee that $q_n(x)$ will be close to $Q(x)$ for a given x. In the literature, two classes of nonparametric regression function estimates have been developed that possess such asymptotically optimal properties (i.e., such that q_n is close to Q for large n in some probabilistic sense). The first one has evolved from the Parzen–Rosenblatt kernel density estimate [8], [9] and is commonly referred to as the Nadaraya estimate [10], [11] or kernel estimate. Nadaraya lets

$$q_n(x) = \sum_{i=1}^{n} Y_i K((X_i - x)/h_n) \Big/ \sum_{i=1}^{n} K((X_i - x)/h_n) \tag{1}$$

where K is a density on R^d and $\{h_n\}$ is a sequence of positive numbers. In [10], he shows that $\sup |q_n(x) - Q(x)| \xrightarrow{n} 0$ with probability one if $d = 1$, if Q, K, and $\{h_n\}$ satisfy some regularity conditions, and if the X_i have a common density. The drawback of his estimate is that the h_n are

Manuscript received October 29, 1976; revised March 18, 1977. This work was supported in part by the United States Air Force under Grant AFOSR-72-2371.

L. P. Devroye was with the Department of Electrical Engineering, University of Texas, Austin. He is now with the School of Computer Science, McGill University, Montreal, Canada H3C 3G1.

0018-9448/78/0300-0142$00.75 © 1978 IEEE

picked without regard to the data. If K is the uniform density on the sphere $S(0,h_n)$ centered at 0 with radius h_n, then (1) computes the average over all the Y_i corresponding to X_i that take values in the sphere $S(x,h_n)$. Assume now that X_1^x is the nearest neighbor to x and that Y_1^x is the corresponding Y_i, and define $h_n = \|X_1^x - x\|$; then (1) reduces to the simple nearest neighbor estimate

$$q_n(x) = Y_1^x$$

which is noted by Cover [12].

In pattern recognition, $(X_1,Y_1), \cdots, (X_n,Y_n)$, (X,Y) are independent identically distributed random vectors, the Y_i are $\{0,1\}$-valued, and Y is unknown. Upon observing $X = x$, Y is estimated by $\tilde{Y}$, a function of X and of the (X_i,Y_i). One rule for which the probability of error $P\{\tilde{Y} \neq Y\}$ is minimal is the following:

$$\begin{aligned}&\tilde{Y} = 1, \quad \text{if } Q(x) = E\{I_{\{Y=1\}} | X = x\} = P\{Y = 1 | X = x\} \geq \tfrac{1}{2}\\ &\tilde{Y} = 0, \quad \text{otherwise.}\end{aligned}$$

Since the conditional expectation $Q(x)$ is unknown, this rule cannot be realized. The obvious solution is to replace Q by a regression function estimate q_n:

$$\begin{aligned}&\tilde{Y} = 1, \qquad \text{if } q_n(x) \geq \tfrac{1}{2}\\ &\tilde{Y} = 0, \qquad \text{otherwise.}\end{aligned}$$

If $q_n(x) = Y_1^x$ as with the nearest neighbor estimate, we obtain the nearest neighbor discrimination rule which lets $\tilde{Y} = Y_1^x$. For more on the nearest neighbor rule, see [12], [13]. Of course, it is unreasonable to expect that $q_n(x)$ approaches $Q(x)$ as n grows large, unless Q is continuous at x and F_x concentrates its mass at $Q(x)$ (no noise situation; this condition corresponds to the nonoverlapping classes condition in discrimination). To correct for this noise sensitivity, Cover and Hart [13] proposed the use of a k-nearest neighbor rule in pattern recognition. The k-nearest neighbor regression function estimate is defined by

$$q_n(x) = k^{-1} \sum_{i=1}^{n} Y_i I_{\{X_i \text{ is among the } k\text{-nearest neighbors to } x\}} \tag{2}$$

where $k/n \overset{n}{\to} 0$ and $k \overset{n}{\to} \infty$. To estimate nonparametrically a density f at x, Loftsgaarden and Quesenberry [14] used a similar idea, viz., they let the estimate be $k/n\ V_n(x)$ where $V_n(x)$ is the volume of the sphere centered at x with the kth nearest neighbor to x on its surface. The k-nearest neighbor regression function estimate (2) was recently generalized independently by Stone [15] and Devroye [16] as follows. First, reorder the (X_i,Y_i) according to increasing distances $\|X_i - x\|$ (if $\|X_i - x\| = \|X_j - x\|$, then we arbitrarily call X_i closer to x if $i < j$), and obtain $(X_1^x,Y_1^x), \cdots, (X_n^x,Y_n^x)$. Then define

$$q_n(x) = \sum_{i=1}^{n} v_{ni} Y_i^x \tag{3}$$

where $v_n = (v_{n1}, \cdots, v_{nn})$ is a probability vector. Picking $v_{ni} = 1/k$ if $i \leq k$ and 0 otherwise gives us back the k-nearest neighbor estimate.

A global measure of the accuracy of a curve estimate q_n is its distance in L_r $(l \leq r \leq \infty)$ from Q, provided that both q_n and Q belong to L_r:

$$\|q_n - Q\|_r = \begin{cases}\left(\int |q_n(x) - Q(x)|^r \, dG(x)\right)^{1/r}, & r < \infty\\ \operatorname*{ess\,sup}_G |q_n(x) - Q(x)|, & r = \infty,\end{cases}$$

where G is the common distribution function of the X_i. In the context of this paper, we will define $\|q_n - Q\|_\infty$ by $\sup_{x \in B} |q_n(x) - Q(x)|$ where B is the support of G. Obviously, if q_n and Q are continuous, then both definitions are equivalent.

If the F_x are such that

$$E\{|Y_1|^r\} = \int\int |y|^r \, dF_x(y) \, dG(x), \qquad \text{for some } r < \infty,$$

and if

i) $v_{n1} \geq v_{n2} \geq \cdots \geq v_{nn}$,
ii) $\max_i v_{ni} \overset{n}{\to} 0$, and
iii) $\Sigma_{i=k_n+1}^{n} v_{ni} \overset{n}{\to} 0$, and $k_n/n \overset{n}{\to} 0$ for some integer sequence $\{k_n\}$,

then Stone [15] shows that $E\{\|q_n - Q\|_r\} \overset{n}{\to} 0$. This result is quite surprising because Q is not required to be "almost continuous" or smooth, the assumption that was at the basis of our use of k-nearest neighbor estimates. In addition, the F_x and G need not have densities as with the Nadaraya estimate. Condition i) insures that more weight is attached to nearer neighbors; the tails of the probability vector v_n must become negligible as n grows large (first part of iii)) so that only an increasingly small proportion (k_n/n) of the samples plays a role in the estimation of $Q(x)$. However, the noise on the observations can only be averaged out if k_n diverges and if, among the k_n nearest neighbors, there is none whose weight dominates the other weights. But this follows if we make the vote v_{ni} of every (X_i^x,Y_i^x) asymptotically negligible (condition ii)).

Implicit in [16] is the following result. If Q is G-almost everywhere continuous, if ess sup $|Y_1| < \infty$ and if ii) and iii) hold, then $|q_n(X_1) - Q(X_1)| \overset{n}{\to} 0$ in probability. If in addition

$$\sum_n \exp\left(-\alpha \Big/ \max_i v_{ni}\right) < \infty, \qquad \text{for all } \alpha > 0,$$

then the convergence is with probability one as well. The main result of this paper is that $\|q_n - Q\|_\infty \overset{n}{\to} 0$ with probability one if Q is uniformly continuous, if G has compact support B, and if $\{v_n\}$ and the F_x satisfy some regularity conditions (for instance, it suffices to pick the v_{ni} as with a k_n-nearest neighbor estimate and ask that $k_n/n \overset{n}{\to} 0$, that $k_n/\log n \overset{n}{\to} \infty$, and that Y_1^x is a bounded random variable).

The nearest neighbor estimates are useful in applications because, as will be shown below, all their powerful properties remain valid for a large class of dependent sampling procedures. If the estimate q_n is used as a guide in optimization, then rather than use a nonsequential random search technique, it may be more economical to estimate and optimize Q in a sequential manner by repeating the following steps:

i) make one observation Y_n from the distribution function F_{X_n}, where the distribution function of X_n itself is picked as a function of q_{n-1},
ii) find q_n and update the current best estimate of the minimum of Q.

The details of this sequential optimization procedure will be discussed in Section VII. We require that the distribution function of X_n be of the form $\alpha_n G + (1 - \alpha_n) G_n$ where G is a distribution function as for crude search in B (e.g., uniform in hypercube), $\Sigma_n \alpha_n = \infty$ (this will insure enough crude search), and the G_n are arbitrary distribution functions, possibly depending upon $(X_1,Y_1),\cdots,(X_{n-1},Y_{n-1})$. If G_n is Gaussian, centered at the old best-estimate X^*_{n-1} of the minimum and with a gradually decreasing variance σ_n^2, then the frequency of samples X_i in the area of interest for the optimization of Q will increase as n grows large.

We will study the asymptotic properties of $\|q_n - Q\|_\infty$, first for the noiseless case (that is, when $Y_1 = Q(X_1)$ with probability one), next for the noisy case, and finally for the noisy case with dependent sampling. A brief section is devoted to the study of the rate of convergence. In the final section, we show that all uniformly good regression function estimates (estimates for which $\|q_n - Q\|_\infty \xrightarrow{n} 0$ in some sense) can be used to design asymptotically optimal random search procedures. For clarity, all proofs are deferred to the Appendix.

II. Estimation in the Absence of Noise

Assume that $X_1,\cdots,X_n$ are independent random vectors with a common distribution function G whose support B is a subset of R^d. Assume further that Q is a Borel measurable function and that, for all x, F_x puts mass 1 at $Q(x)$. This implies that $Y_i = Q(X_i)$ with probability one for all i. The following condition on the sequence of weight vectors $\{v_n\}$ will be needed throughout.

Condition C1: The sequence $\{v_n\}$ of probability vectors $(v_{n1},\cdots,v_{nn})$ is such that for some sequence $\{k_n\}$ of positive integers, i)

$$k_n/n \xrightarrow{n} 0$$

and ii)

$$\sum_{i=k_n+1}^{n} v_{ni} \xrightarrow{n} 0.$$

Thus the tail of the vector v_n must be asymptotically negligible. This condition is satisfied if $v_n = (a_1,\cdots,a_k,0,\cdots,0)$ for some fixed probability vector $(a_1,\cdots,a_k)$. In particular, the nearest neighbor estimate of Cover has $k = 1$ (and thus $a_1 = 1$). The main result for estimates satisfying C1 is the following.

Theorem 1: If Q is continuous, G has compact support, condition C1 holds, and F_x puts mass 1 at $Q(x)$, then $\|q_n - Q\|_\infty \xrightarrow{n} 0$ with probability one.

In [15], [16] it is indicated that for noisy situations, the influence of a single (X_i,Y_i) on the estimate must become negligible as n grows large ($\max_i v_{ni} \xrightarrow{n} 0$), and this of course forces k_n to grow unbounded in condition C1. Examples of sequences $\{v_n\}$ satisfying C1 while $\max_i v_{ni} \xrightarrow{n} 0$ are plentiful:

i) *rectangular weight vector:* $v_{ni} = 1/k_n$, $1 \le i \le k_n$, and $v_{ni} = 0$ otherwise; the sequence k_n satisfies $k_n \xrightarrow{n} \infty$ and $k_n/n \xrightarrow{n} 0$;
ii) *triangular weight vector:* $v_{ni} = 2(k_n - i + 1)/(k_n + k_n^2)$, $1 \le i \le k_n$, and $v_{ni} = 0$ otherwise; $k_n \xrightarrow{n} \infty$ and $k_n/n \xrightarrow{n} 0$;
iii) *exponential weight vector:* $v_{ni} = a_n(1 + a_n)^{-i}(1 - (1 + a_n)^{-n})^{-1}$; $a_n \xrightarrow{n} 0$ and $na_n \xrightarrow{n} \infty$. To see that this sequence satisfies the said conditions, let $k_n \sim \sqrt{n/a_n}$.

III. Estimation in Noisy Conditions

Assume that $(X_1,Y_1),\cdots,(X_n,Y_n)$ are independent R^{d+1}-valued random vectors with a common distribution function. The distribution function G of X_1 has support B, a subset of R^d, and, given that $X_1 = x$, Y_1 has distribution function F_x. We distinguish between several types of noise. We say that the noise is

absent if

$$\sup_{x \in B} \int |y - Q(x)|^2 \, dF_x(y) = 0,$$

exponential if for all $\epsilon > 0$ there exists a $c(\epsilon) > 0$ such that

$$\sup_{x \in B} \int e^{s(y-Q(x))} \, dF_x(y) \le e^{|s|\epsilon}, \qquad \text{all } |s| \le c(\epsilon),$$

bounded if there exist finite nonnegative numbers K_1 and K_2 such that

$$\sup_{x \in B} \int_{Q(x)+K_1}^{\infty} dF_x(y) + \int_{-\infty}^{Q(x)-K_2} dF_x(y) = 0,$$

in L_t (where $t > 0$) if, for some finite K,

$$\sup_{x \in B} \int |y - Q(x)|^t \, dF_x(y) \le K,$$

uniformly integrable if

$$\lim_{s \to \infty} \sup_{x \in B} \int_{|y-Q(x)| \ge s} |y - Q(x)| \, dF_x(y) = 0, \text{ and}$$

additive if there exists a distribution function F such that $F(y) = F_x(y - Q(x))$ for all real y and all $x \in B$.

In random optimization and probabilistic automata theory, the collection $\{F_x \mid x \in B\}$ of distribution functions is called a random environment, but in order not to confuse the reader with more technical jargon, we will use the term noise to denote both this collection of distribution functions and the sequence of random variables $Y_1, \cdots, Y_n$.

To situate the types of noise relative to each other, we recall that bounded noise is always exponential, that exponential noise is L_t noise for all positive t, that L_s noise is L_t noise if $t \leq s$, and that if the noise is in L_t for some $t > 1$, then it must be uniformly integrable. Further, additive noise is uniformly integrable since $\int |y| dF(y) < \infty$ by hypothesis. It will also be L_t noise for $t > 1$ if $\int |y|^t dF(y) < \infty$. An interesting case for engineers is when all the F_x are Gaussian with variance σ_x^2 and mean $Q(x)$. It is easy to see that this type of noise is exponential if $\sup_{x \in B} \sigma_x^2 < \infty$. In most practical situations, all the F_x put their weight on an interval of length d_x. This type of noise is exponential if $\sup_{x \in B} d_x < \infty$.

Let us for the moment consider estimates (3) for which condition C2 holds.

Condition C2: The sequence $\{v_n\}$ of probability vectors satisfies $v_{ni} = 1/k_n$ for $1, \leq i \leq k_n$, and $v_{ni} = 0$ for $i > k_n$, where $\{k_n\}$ is a sequence of positive integers with $k_n/n \xrightarrow{n} 0$ and $k_n \xrightarrow{n} \infty$.

If the norm $\|\cdot\|$ that is used to reorder the data is the maximum component norm, then the following is true.

Theorem 2: If Q is continuous, G has compact support, condition C2 holds, the noise is exponential, and

$$k_n/\log n \xrightarrow{n} \infty,$$

then $\|q_n - Q\|_\infty \xrightarrow{n} 0$ with probability one for estimate (3).

Theorem 3: If Q is continuous, G has compact support, condition C2 holds, the noise is in L_t for some $t > 2d + 1$, and

$$k_n^{t-1}/n^{2d} \xrightarrow{n} \infty,$$

then $\|q_n - Q\|_\infty \xrightarrow{n} 0$ in probability for estimate (3). If in addition

$$\sum_{n=1}^{\infty} n^{2d}/k_n^{t-1} < \infty,$$

then $\|q_n - Q\|_\infty \xrightarrow{n} 0$ with probability one.

For Theorem 3 to apply, the noise must be at least in L_{2d+1}. The question remains whether the conclusion of the theorem remains valid for the class of L_2 noises which is so important in control engineering applications. If the L_2 norm is used instead of the maximum component norm on R^d, then the factor n^{2d} in Theorem 3 can be replaced by n^{d+1} and the condition $t > 2d + 1$ must be replaced by the condition $t > d + 2$.

The nearest neighbor multiple-trial estimate with λ_n trials satisfies similar properties. Obviously, in the absence of noise, it is senseless to let $\lambda_n > 1$, while for $\lambda_n = 1$ the classical nearest neighbor estimate is obtained to which Theorem 1 applies. In noisy situations, the k_n-nearest neighbor estimate eliminates the effect of the noise due to the averaging of $Y_1^x, \cdots, Y_{k_n}^x$. With the nearest neighbor multiple trial estimate, the noise reduction is achieved via averaging of $Y_1, \cdots, Y_{\lambda_n}$. Thus we can expect that λ_n will replace k_n in the conditions of convergence, as is seen from the following theorems.

Theorem 2': If Q is continuous, G has compact support, the noise is exponential, and

$$\lambda_n/\log n \xrightarrow{n} \infty,$$

then $\|q_n - Q\|_\infty \xrightarrow{n} 0$ with probability one for the nearest neighbor multiple-trial estimate.

Theorem 3': If Q is continuous, G has compact support, the noise is in L_t for some $t > 1$, and

$$\lambda_n^{t-1}/n \xrightarrow{n} \infty,$$

then $\|q_n - Q\|_\infty \xrightarrow{n} 0$ in probability for the nearest neighbor multiple-trial estimate. If in addition

$$\sum_{n=1}^{\infty} n/\lambda_n^{t-1} < \infty,$$

then $\|q_n - Q\|_\infty \xrightarrow{n} 0$ with probability one.

The nearest neighbor multiple-trial estimate with λ trials and n samples $X_1, \cdots, X_n$ uses λn measurements and can, in data collection cost, be compared with the λ-nearest neighbor single trial estimate with λn samples. It has the advantage however that for a given x, to find $q_n(x)$, it suffices to find the nearest neighbor to x among $X_1, \cdots, X_n$ and to look up the value $Q_\lambda(X_1^x)$ that is already stored in a memory. Also, the convergence of $\|q_n - Q\|_\infty$ can be assured for all L_t noises ($t > 1$) if λ_n grows fast enough. *Thus the nearest neighbor multiple trial estimate seems better suited for situations with heavy noise, easy access to data, and relatively more expensive computing time.* Notice that in some problems, the engineer has no access to more than one Y_i for every X_i so that he is forced to use a single-trial estimate.

IV. A Simplified Regression Function Estimate

Estimate (3) requires for every x the reordering of $X_1, \cdots, X_n$ and the computation of a sum of n terms. Consider the following simplified estimate derived from (3)

$$\tilde{q}_n(x) = q_n(X_1^x) \tag{4}$$

which has none of these drawbacks because

i) the $q_n(X_i)$, $1 \leq i \leq n$, can be computed in advance and stored in a memory, and
ii) to find $\tilde{q}_n(x)$, it suffices to find the nearest neighbor X_1^x and look up the value of $q_n(X_1^x)$.

Theorem 4: Let Q be continuous and let G have compact support B. Assume that condition C1 holds, that the noise is exponential, that

$$\left(\max_i v_{ni}\right) \log n \overset{n}{\to} 0$$

or that the noise is in L_t for some $t \geq 2$, and that

$$\sum_{n=1}^{\infty} n \left(\max_i v_{ni}\right)^{t-1} < \infty; \tag{5}$$

then $\|\tilde{q}_n - Q\|_\infty \overset{n}{\to} 0$ with probability one. If, instead of (5) we have

$$n \left(\max_i v_{ni}\right)^{t-1} \overset{n}{\to} 0,$$

then $\|\tilde{q}_n - Q\|_\infty \overset{n}{\to} 0$ in probability.

The conditions of convergence of Theorem 4 do not depend upon d or the norm $\|\cdot\|$ that is used to reorder $X_1, \cdots, X_n$ (actually, any norm on R^d can be used). The reason that the conditions of convergence are weaker than the ones given in Theorems 2 and 3 is because $\tilde{q}_n$ in (4) is better behaved than q_n in (3) for large n (for one thing, $\tilde{q}_n$ can take only n values while q_n can take almost n^{2d} values).

V. Rate of Convergence

Knowing that $\|q_n - Q\|_\infty \overset{n}{\to} 0$ with probability one assures the engineer that taking n large enough will force q_n to be uniformly close to Q. Two questions immediately arise.

i) How large should n be such that, for given ϵ, $\delta > 0$, $P\{\|q_n - Q\|_\infty > \epsilon\} < \delta$? That is, how fast does $P\{\|q_n - Q\|_\infty > \epsilon\}$ tend to 0 as n grows large?

ii) How fast can we make ϵ_n decrease to 0 in order that $P\{\|q_n - Q\|_\infty > \epsilon_n\}$ still tends to 0 as $n \to \infty$?

Some authors prefer to use ii) in the study of rates of convergence of random sequences but, in the context of this paper, i) seems to be a far more interesting question. Assuming for the moment that the conditions of Theorem 2 are fulfilled, we see from a quick inspection of the proof of Theorem 2 that for every $\epsilon > 0$, we can find an N large enough and positive constants a_i, all depending upon ϵ, Q, G, and the collection of F_x, such that, for the Stone–Devroye estimate (3),

$$P\{\|q_n - Q\|_\infty > \epsilon\} \leq a_1 e^{-a_2 n} + a_3 n^{2d} e^{-a_4 k_n}, \qquad \text{all } n \text{ with } k_n \leq n/N.$$

Truly practical expressions for the a_i and N can be obtained only if additional assumptions are made regarding Q, G, and the noise. Assume for instance that Q is Lipschitz with constant C (that is, $|Q(x) - Q(y)| \leq C\|x - y\|$ for all x and y), that G is the uniform distribution function on $[-M,M]^d$, and that all the F_x put their mass on an interval of length at most D containing $Q(x)$. For

$$k_n = \gamma n(\epsilon/4MC)^d/2, \qquad 0 < \gamma \leq 1,$$

the following bound is valid (see the proofs of Theorems 1 and 2):

$$P\{\|q_n - Q\|_\infty > 2\epsilon\} \leq 2(8MC/\epsilon)^d e^{(-n(\epsilon/4MC)^{2d})/2} + 2(1+n)^{2d} e^{-\gamma n \epsilon^2 (\epsilon/4MC)^d / D^2}.$$

The best choice of γ seems to be 1, but since we want k_n to be just large enough so that the noise averaging effect and the influence of the variation of Q on the k_nth nearest neighbor are about equal, it is logical to try to pick γ such that both terms in the given bound are equal. Matching the exponents would give $\gamma = \min(1;(\epsilon/4MC)^d(D/\epsilon)^2/2)$. This is not a surprise since, with small noise (D small) and highly irregular Q (C large), the engineer will intuitively prefer to use a smaller k_n in the k_n-nearest neighbor estimate. Similar finite sample studies can be made for other types of noise, other estimates, and under other conditions on G and Q.

VI. Estimation with Dependent Sampling

In the introduction, we indicated why it is important in some applications to gradually take more samples from a certain region of B such as the region close to the global minimum of Q. Consider thus the following model for a dependent sampling procedure. Let $\{\alpha_n\}$ be a sequence from $[0,1]$ with cumulative sums β_n, $n \geq 1$, and let $Z_1, Z_2, \cdots$, be a sequence of independent binary-valued random variables with $P\{Z_n = 1\} = \alpha_n$, $n \geq 1$. If $Z_n = 1$, then X_n is independent of $X_1, \cdots, X_{n-1}$ and has distribution function G; while if $Z_n = 0$, then X_n has an arbitrary distribution function G_n, possibly depending upon $(X_1,Y_1), \cdots, (X_{n-1},Y_{n-1})$. Thus the distribution function of X_n is $\alpha_n G + (1 - \alpha_n) G_n$. Given $X_1, \cdots, X_n$, the $Y_1, \cdots, Y_n$ are independent random variables with distribution functions $F_{X_1}, \cdots, F_{X_n}$. In the Appendix, we prove the following generalization of Theorems 1–3 for the Stone–Devroye estimate (3).

Theorem 5: Let G have compact support B, let X_1, $X_2, \cdots$, take values (with probability one) in a closed set B_0 containing B, and let Q be continuous and bounded on B_0.

i) If condition C1 holds, if $k_n/\beta_n \overset{n}{\to} 0$, if $\beta_n \overset{n}{\to} \infty$, and if the noise is absent, then $\|q_n - Q\|_\infty \overset{n}{\to} 0$ in probability for estimate (3). If in addition $\beta_n/\log n \to \infty$, then $\|q_n - Q\|_\infty \overset{n}{\to} 0$ with probability one.

ii) If condition C2 holds, if $k_n/\beta_n \overset{n}{\to} 0$, if $k_n/\log n \overset{n}{\to} \infty$, and if the noise is exponential, then $\|q_n - Q\|_\infty \overset{n}{\to} 0$ with probability one.

iii) If condition C2 holds, if $k_n/\beta_n \overset{n}{\to} 0$, if $k_n^{t-1}/n^{2d} \overset{n}{\to} \infty$, and if the noise is in L_t for some $t > 2d + 1$, then $\|q_n - Q\|_\infty \overset{n}{\to} 0$ in probability. The convergence is with probability one if in addition $\Sigma_n n^{2d}/k_n^{t-1} < \infty$.

VII. An Application in Optimization

In estimation, the distribution function G of the X_i is often unknown, while in crude search of a regression function the distribution function G is picked in such a way that it covers the area in which the minimum of Q is sought, e.g., G is uniform on a hypercube B so that the search area and the support of G coincide.

Consider the following general setup for a sequential random optimization scheme that uses all the past information in an intelligent way. Let $X_1^*, X_2^*, \cdots$, be the sequence of best estimates of the minimum of Q in R^d, and let $(X_1,Y_1),(X_2,Y_2),\cdots$, be a sequence of R^{d+1}-valued random vectors. Given X_{n-1}^* and $(X_1,Y_1),\cdots,(X_{n-1},Y_{n-1})$, take three steps to find X_n^*, the next best estimate of the minimum.

i) Make one observation (X_n,Y_n) where X_n has distribution function $\alpha_n G + (1-\alpha_n)G_n$ with $\alpha_n \in [0,1]$. G_n is an arbitrary distribution function and a Borel measurable function of $(X_1,Y_1),\cdots,(X_{n-1},Y_{n-1})$, X_{n-1}^*. Given that $X_n = x$, Y_n is an independent random variable with distribution function F_x.

ii) Let γ_n be a number from $[0,1]$, and let H_n be an arbitrary distribution function and a Borel measurable function of $(X_1,Y_1),\cdots,(X_n,Y_n)$, X_{n-1}^*. Let further $W_1^*,\cdots,W_n^*$ be a sequence of independent random vectors with common distribution function G. With probability γ_n, define $W_n = W_n^*$. W_n, called the candidate best estimate, has distribution function $\gamma_n G + (1-\gamma_n)H_n$.

iii) Compute $q_n(W_1),\cdots,q_n(W_n)$ and let X_n^* be the W_i for which $q_n(W_i)$ is minimal (in the case of ties, break them arbitrarily).

The search is started with $\alpha_1 = \gamma_1 = 1$. As an example, let $\alpha_n = \gamma_n$, $H_n = G_n$, and $W_n = X_n$ so that, on the basis of q_n, X_n^* is the best choice among $X_1,\cdots,X_n$. Usually the search area B is a hypercube, and G is a global search oriented distribution function such as the uniform distribution function on B. G_n is a local search oriented distribution function such as the one that is Gaussian with mean $X_{n-1}^* + D_n$ (D_n is an R^d-valued random vector called the bias) and variance S_n, where

$$S_n = \begin{cases} 0, & \text{if } X_{n-1}^* \neq X_{n-2}^* \\ \sigma^2, & \text{if } X_{n-1}^* = X_{n-2}^*, \end{cases}$$

and

$$D_n = \delta(X_{n-1}^* - X_{n-2}^*), \qquad \text{for some } \delta > 1, \sigma > 0.$$

We can let α_n tend to 0 (insuring however that $\Sigma\alpha_n = \infty$) so that in the beginning there is a larger portion of global search, and in later stages the emphasis will be on local search. To control the portion of estimation relative to the effort spent on optimization, H_n can be the distribution that puts mass 1 at X_{n-1}^*. The net result of this is that the number of different values taken by the W_i, $1 \le i \le n$, is approximately equal to $\Sigma_{i=1}^n \gamma_i$.

The minimum of Q is defined by

$$q_{\min} = \operatorname*{ess\,inf}_G Q(x).$$

Thus if $q_{\min}$ is finite, it is the unique number with the property that, for all $\epsilon > 0$, $P\{Q(X_1) \le q_{\min} - \epsilon\} = 0$ and $P\{Q(X_1) \le q_{\min} + \epsilon\} > 0$. If Q is continuous, then $q_{\min}$ is the infimum, over the support B of G, of $Q(x)$. If G is atomic, then regardless of whether Q is continuous or not, $q_{\min}$ will be the infimum of $Q(x)$ over all x for which $P\{X_1 = x\} > 0$.

Theorem 6 shows why it is important that $\|q_n - Q\|_\infty \xrightarrow{n} 0$ if the estimate q_n is going to be used in optimization.

Theorem 6: If q_n is *any* estimate of Q with the property that $\|q_n - Q\|_\infty \xrightarrow{n} 0$ in probability (with probability one), if G has support B, if all W_n take values in B with probability one, and if

$$\sum_{n=1}^{\infty} \gamma_n = \infty,$$

then $\max(Q(X_n^*), q_{\min}) \xrightarrow{n} q_{\min}$ in probability (with probability one).

Notice that we must use $\max(Q(X_n^*), q_{\min})$ since it is possible that $Q(X_n^*)$ is strictly smaller than $q_{\min}$ (e.g., let $B = [0,1]$, let G be uniform on B, and let $Q(x) = 1$ with the exception that $Q(0.3) = 0$. If $\gamma_n = \frac{1}{2}$ and H_n is atomic at 0.3, then $Q(X_n^*)$ tends with probability one to 0 while $q_{\min} = 1$).

One can object that the given optimization procedure requires a growing memory for the storage of the (X_i,Y_i) (and the W_i in some cases). Clearly this is not a major drawback in the presence of ultrahigh-speed and large-capacity computers. Moreover, the loss (in terms of the number of samples to be collected for the same accuracy of the search) incurred by forgetting or not using some of the (X_i,Y_i) may be higher than the cost of time and memory resulting from the computations of the $q_n(W_i)$ and the storage of the (X_i,Y_i). *Thus the specific field of application of this class of optimization techniques seems to be the one in which data collection is expensive and computation is cheap.* A paper is in preparation in which the (X_i,Y_i) and the W_i need not be memorized but in which, upon observing a new (X_n,Y_n), the estimate X_n^* is updated, and (X_n,Y_n) is forgotten. Such a technique has the limited memory flavor of the classical search techniques, but the engineer cannot use it in applications in which for some reason he desires to have an estimate of the regression function that he is minimizing.

Theorem 6 remains valid if X_n^* is picked in such a way that

$$q_n(X_n^*) = \min_{x \in B} q_n(x)$$

when q_n is an estimate which attains its minimum on B (all the estimates discussed in this paper do). However, this would require a subsequent search at every iteration instead of the proposed relatively simple comparison of n or less vectors.

For the selection of α_n, γ_n, G_n, and v_n, the engineer must be guided by his experience. A choice for v_n is suggested in the section on the rate of convergence of estimate (3).

For choices of local search oriented distribution functions G_n, the reader is referred to the random search literature, in particular to the work by Cockrell and Fu [3] and Matyas [17].

Appendix

We start off by showing that if B is the support of G, the common distribution function of $X_1, \cdots, X_n$, then $P\{X_1 \in B\} = 1$, and B is closed. If Q is continuous, then it follows that Q is bounded and uniformly continuous on B whenever B is bounded.

Proof: If $G_x(\epsilon) = P\{\|X_1 - x\| \leq \epsilon\}$, then the support of G is the set of all x with the property that $G_x(\epsilon) > 0$ for all $\epsilon > 0$. It is easy to see that B is closed. Indeed, if y is a cluster point of B and $\epsilon > 0$ is arbitrary, then there exists an x_ϵ in the intersection of B and $S(y,\epsilon/2)$, the closed sphere with center y and radius $\epsilon/2$. Thus $G_y(\epsilon) = P\{\|X_1 - y\| \leq \epsilon\} \geq P\{\|X_1 - x_\epsilon\| \leq \epsilon/2\} = G_{x_\epsilon}(\epsilon/2) > 0$. In conclusion, if there exists a finite M such that $P\{\|X_1\| \leq M\} = 1$, then B, the support of G, is compact.

To show that $P\{X_1 \in B\} = 1$, note that B^c, the complement of B, is the set of all x in R^d for which for some $\epsilon(x) > 0$, $P\{X_1 \in S(x,\epsilon(x))\} = 0$. We also know that R^d is separable, and thus that there exists a countable dense subset D of R^d. Since D is dense, find for each x in B^c a $d(x)$ in D such that $d(x) \in S(x,\epsilon(x)/3)$. Thus, $S(d(x),\epsilon(x)/2)$ is contained in $S(x,\epsilon(x))$, and therefore $P\{X_1 \in S(d(x),\epsilon(x)/2)\} = 0$. Also, x is in $S(d(x),\epsilon(x)/3)$ so that

$$P\{X_1 \in B^c\} \leq \sum_{d=d(x) \text{ for some } x \text{ in } B^c} P\{X_1 \in S(d,a)\} = 0$$

as a countable union of null sets, where

$$a = \sup_{x \text{ in } B^c \text{ for which } d(x)=d} \epsilon(x)/3.$$

Q.E.D.

Next we show that if B is bounded, then $\inf_{x\in B} G_x(\epsilon) > 0$ for all $\epsilon > 0$.

Proof: Assume that $\inf_{x\in B} G_x(\epsilon) = 0$ for some $\epsilon > 0$. Thus there exists a sequence $x_1, x_2, \cdots$, from B with $G_{x_i}(\epsilon) \rightarrow 0$. Since B is compact, the sequence $\{x_i\}$ must have a cluster point y in B. Therefore, there exists a further subsequence $\{x_i^*\}$ such that $G_{x_i^*}(\epsilon) \rightarrow 0$ and $\|x_i^* - y\| \leq \epsilon/2$ for all i. Thus $S(x_1^*,\epsilon/2)$ is contained in the intersection of all the $S(x_i^*,\epsilon)$. Hence, $G_{x_1^*}(\epsilon/2) \leq \liminf_i G_{x_i^*}(\epsilon) = 0$, which contradicts the fact that x_1^* belongs to B.

Q.E.D.

Lemma 1: If $X_1, \cdots, X_n$ are independent zero-mean random variables with the property that for every $\epsilon > 0$ there exists a $c(\epsilon) > 0$ such that

$$E\{e^{sX_i}\} \leq e^{|s|\epsilon}, \qquad \text{for all } |s| \leq c(\epsilon) \text{ and } 1 \leq i \leq n,$$

and if $(a_1, \cdots, a_n)$ is a probability vector, then

$$P\left\{\left|\sum_{i=1}^n a_i X_i\right| > \epsilon\right\} \leq K_1 e^{-K_2/\max_i a_i}$$

for some $K_1, K_2 > 0$ depending upon ϵ.

Lemma 2: If $X_1, \cdots, X_n$ are independent zero-mean random variables with the property that

$$E\{|X_i|^t\} \leq M < \infty, \qquad 1 \leq i \leq n, \text{ some } t > 1,$$

then

$$P\left\{\left|\sum_{i=1}^n X_i/n\right| > \epsilon\right\} \leq K_3/n^{t-1}$$

for some $K_3 > 0$ depending upon L, t, and ϵ. If $(a_1, \cdots, a_n)$ is a probability vector and $t \geq 2$, then

$$P\left\{\left|\sum_{i=1}^n a_i X_i\right| > \epsilon\right\} \leq K_4 \left(\max_i a_i\right)^{t-1}$$

for some $K_4 > 0$ depending upon L, t, and ϵ.

Proof of Lemmas 1 and 2: For Lemma 1, we have that

$$P\left\{\sum_{i=1}^n a_i X_i > \epsilon\right\} \leq e^{-s\epsilon} \prod_{i=1}^n E\{e^{sa_iX_i}\}, \qquad \text{all } s > 0,$$

$$\leq e^{-|s|\epsilon} e^{\Sigma|s|a_i\epsilon/2}, \qquad \text{all } s \text{ with } a_i|s| \leq c(\epsilon/2),\ 1 \leq i \leq n,$$

$$\leq e^{-|s|\epsilon/2} \leq e^{-c(\epsilon/2)/2\, \max_i a_i}.$$

Thus, by symmetry, Lemma 1 holds with $K_1 = 2$ and $K_2 = \epsilon c(\epsilon/2)/2$.

The first part of Lemma 2 is a direct corollary of a theorem of Wagner [22]. In [22], no explicit expression for K_3 is derived. Values for K_3 for the case $t \geq 2$ can be found in Fuk and Nagaev [19, p. 653]. In addition, they show that

$$P\left\{\left|\sum_{i=1}^n a_i X_i\right| \geq \epsilon\right\} \leq 2(1 + 2/t)^t \sum_{i=1}^n E\{|na_iX_i|^t\}/(n_\epsilon)^t$$
$$+ 2 \exp\left(-2e^{-t} n^2 \epsilon^2/(t + 2) \sum_{i=1}^n E\{|na_iX_i|^2\}\right).$$

By using the facts $\Sigma a_i = 1$, $E\{|X_i|^2\} \leq (E\{|X_i|^t\})^{2/t} \leq M^{2/t}$, $E\{|X_i|^t\} \leq M$, this expression can be overbounded by $K_4'(\max_i a_i)^{t-1} + 2 \exp(-K_5/\max_i a_i)$ where $K_4' = 2(1 + 2/t)^t M/\epsilon^t$ and $K_5 = 2e^{-t}\epsilon^2/(t + .2)M^{2/t}$. Since, for all $x,y > 0$, $e^{-x} \leq (y/ex)^y$, we can further overbound the last expression by $K_4(\max_i a_i)^{t-1}$ if we let $K_4 = K_4' + 2((t - 1)/eK_5)^{t-1}$. Q.E.D.

A last word of caution is in order before we can start to prove Theorems 1–6. It is in general not true that one can specify a collection of distribution functions F_x for x belonging to a closed subset B of R^d, and then claim that there exists a random vector (X,Y) where X has distribution function G on R^d and, given that $X = x$, Y has distribution function F_x. However, this measurability question is easily solved if there exists a probability space $(\Omega,\mathcal{A},P)$ and an $(\Omega \times B, \mathcal{A} \times \mathcal{B}_B^d) - (R,\mathcal{B})$ measurable function h ($\mathcal{B}_B^d$ is the class of all Borel sets contained in B, and $\mathcal{B}$ is the class of all Borel sets of R) such that $Y = h(\omega,X)$. In that case, we define

$$F_x(y) = P\{\omega \mid \omega \in \Omega, h(\omega,x) \leq y\}, \qquad x \in B, y \in R.$$

Throughout we assume that the distribution functions F_x are obtained in this fashion without explicitly mentioning the mapping h. Thus, we can write

$$Q(x) = \int y \, dF_x(y),$$

instead of $Q(x) = \int h(\omega,x) P(d\omega)$. This approach has the advantage that $(X,h(\omega,X))$ is a random vector whenever X is a random vector.

Proof of Theorem 1: Let $\epsilon > 0$ be arbitrary, and find an $a > 0$ such that $|Q(z) - Q(x)| \leq \epsilon/2$, for all $z,x \in B$, and $\|z - x\| \leq a$ (use the uniform continuity of Q on B). Let A_n be the event that $\|X_{k_n}^x - x\| \leq a$ for all $x \in B$, and let $q_{\max} = \max_{x\in B} Q(x)$, and $q_{\min}$

$= \min_{x \in B} Q(x)$. Clearly,

$$\begin{aligned} P\left\{\sup_{x\in B} |q_n(x) - Q(x)| > \epsilon\right\} &\le P\{A_n^c\} \\ &+ P\left\{A_n, \sup_{x\in B} |q_n(x) - Q(x)| > \epsilon\right\} \\ &\le P\{A_n^c\} + P\left\{\left|\sum_{i=k_n+1}^{n} v_{ni}(q_{\max} - q_{\min})\right| > \epsilon/2\right\} \\ &+ P\left\{A_n, \sup_{x\in B} \left|\sum_{i=1}^{k_n} v_{ni}(Q(X_i^x) - Q(x))\right| > \epsilon/2\right\} \\ &\le P\{A_n^c\} \end{aligned}$$

for all n large enough, since $\Sigma_{i=k_n+1}^{n} v_{ni} \overset{n}{\to} 0$, $q_{\max} - q_{\min} < \infty$, and since, on A_n, $|Q(X_i^x) - Q(x)| < \epsilon/2$ for all x in B and $1 \le i \le k_n$. We tacitly use the fact that, with probability one, all the X_i take values in B and that, also with probability one, $Y_i = Q(X_i)$ for all i.

Next,

$$A_n \subseteq \{\mu_n(S(x,a)) < k_n/n \text{ for some } x \text{ in } B\},$$

where $\mu_n(C) = \Sigma_{i=1}^{n} I_{\{X_i \in C\}}/n$ is the empirical measure of a set C with $X_1, \cdots, X_n$, and where $S(x,a)$ is the closed sphere with center x and radius a. If μ is the measure on the Borel sets of R^d that corresponds to G, then we have seen that $\inf_{x\in B}\mu(S(x,a/2)) = c > 0$. Since B is compact, we can find a finite number N_1 of points $x_1, \cdots, x_{N_1}$ from B with the property that, for every x in B, there exists an x_i, $1 \le i \le N_1$, with $\|x - x_i\| \le a/2$. Thus, if n is so large that $k_n/n < c/2$, then

$$\begin{aligned} P\{\|q_n - Q\|_\infty > \epsilon\} &\le P\{A_n^c\} \le P\left\{\inf_{x\in B} \mu_n(S(x,a)) < k_n/n\right\} \\ &\le P\left\{\bigcup_{i=1}^{N_1} \{\mu_n(S(x_i,a/2)) < k_n/n\}\right\} \\ &\le N_1 \sup_{x\in B} P\{\mu_n(S(x,a/2)) < k_n/n\} \\ &\le N_1 \sup_{x\in B} P\{\mu_n(S(x,a/2)) - \mu(S(x,a/2)) < -c/2\} \\ &\le 2N_1 e^{-2n(c/2)^2} \\ &= 2N_1 e^{-nc^2/2} \end{aligned}$$

where we used Hoeffding's inequality [21]. Theorem 1 now follows by the Borel–Cantelli lemma since the last term in the chain of inequalities is summable with respect to n. Q.E.D.

Proof of Theorems 2 and 3: Let ϵ be arbitrary and note that

$$\begin{aligned} P\left\{\sup_{x\in B} |q_n(x) - Q(x)| > 2\epsilon\right\} &\le P\left\{\sup_{x\in B}|q_n(x) - q_{n0}(x)| > \epsilon\right\} \\ &+ P\left\{\sup_{x\in B} |q_{n0}(x) - Q(x)| > \epsilon\right\} \end{aligned}$$

where $q_{n0}(x) = \Sigma_{i=1}^{n} v_{ni} Q(X_i^x)$. The last term on the right side is overbounded as in the proof of Theorem 1 because the conditions of Theorem 1 are fulfilled. The first term, accounting for the noise, is overbounded as follows:

$$\begin{aligned} P&\left\{\sup_{x\in B} |q_n(x) - q_{n0}(x)| > \epsilon\right\} \\ &= P\left\{\sup_{x\in B} \left|\sum_{i=1}^{k_n} (Y_i^x - Q(X_i^x))/k_n\right| > \epsilon\right\} \\ &= E\left\{P\left\{\sup_{x\in B} \left|\sum_{i=1}^{k_n} (Y_i^x - Q(X_i^x))/k_n\right| > \epsilon \,|\, X_1, \cdots, X_n\right\}\right\} \\ &\le E\left\{P\left\{\sup_{A\in J(X_1,\cdots,X_n)} \right.\right. \\ &\cdot \left.\left.\left|\sum_{i\in A} (Y_i - Q(X_i))/k_n\right| > \epsilon \,|\, X_1, \cdots, X_n\right\}\right\} \\ &\le E\left\{s(\mathcal{A},n) \sup_{\text{all subsets } \{j_1,\cdots,j_{k_n}\} \text{ of } \{1\cdots n\}} \right. \\ &\cdot \left. P\left\{\left|\sum_{i=1}^{k_n} (Y_{j_i} - Q(X_{j_i}))/k_n\right| > \epsilon \,|\, X_1, \cdots, X_n\right\}\right\} \\ &\le s(\mathcal{A},n) \sup_{\bar{x}_n = (x_1,\cdots,x_{k_n}) \in B^{k_n}} \\ &\cdot P\left\{\left|\sum_{i=1}^{k_n} (Y_i - Q(X_i))/k_n\right| > \epsilon \,|\, X_1 = x_1, \cdots, X_{k_n} = x_{k_n}\right\}, \end{aligned}$$

where $J(X_1, \cdots, X_n)$ is the collection of all sets of k_n indices from $\{1, \cdots, n\}$ such that $A = \{j_1, \cdots, j_{k_n}\}$ belongs to $J(X_1, \cdots, X_n)$ if and only if there exists an x in R^d for which $X_{j_i}, \cdots, X_{j_{k_n}}$ are the k_n-nearest neighbors (not necessarily in that order) to x among $X_1, \cdots, X_n$; $\mathcal{A}$ is the class of all closed and open spheres in R^d; and $s(\mathcal{A},n)$ is the maximum, over all $(x_1, \cdots, x_n) \in R^{dn}$, of the number of different sets in $\{\{x_1, \cdots, x_n\} \cap A \,|\, A \in \mathcal{A}\}$.

From Lemmas 1 and 2, we know that, for every $\epsilon > 0$, there exist positive constants K_1, K_2, K_3 such that

$$\sup_{\bar{x}_n \in B^{k_n}} P\left\{\left|\sum_{i=1}^{k_n} (Y_i - Q(X_i))/k_n\right| > \epsilon \,|\, X_1 = x_1, \cdots, X_{k_n} = x_{k_n}\right\}$$

$$\le \begin{cases} K_1 \exp(-K_2 k_n), & \text{if the noise is exponential,} \\ K_3/k_n^{t-1}, & \text{if the noise is in } L_t (t > 1). \end{cases}$$

Again using the fact that $\|q_n - Q\|_\infty \le \sup_{x\in B}|q_n(x) - Q(x)|$ whenever $B = \text{support}(G)$, we have, using a result from Theorem 1, that for all n large enough,

$$P\{\|q_n - Q\|_\infty > 2\epsilon\} \le P\left\{\sup_{x\in B} |q_n(x) - Q(x)| > 2\epsilon\right\}$$

$$\le \begin{cases} 2N_1 \exp(-nc^2/2) + s(\mathcal{A},n)K_1 \exp(-K_2 k_n), \\ \quad \text{if the noise is exponential,} \\ 2N_1 \exp(-nc^2/2) + s(\mathcal{A},n)K_3/k_n^{t-1}, \\ \quad \text{if the noise is in } L_t. \end{cases}$$

With the maximum component norm $s(\mathcal{A},n) \le (1 + n)^{2d}$ and with the standard L_2 norm, $s(\mathcal{A},n) \le (2n)^{d+1}$. Thus collecting bounds completes the proof of Theorems 2 and 3. For the "with probability one" part of the theorems, the Borel–Cantelli lemma is used together with the fact that, for any $a > 0$,

$$\sum_{n=1} n^a e^{-bk_n} < \infty, \qquad \text{for all } b > 0$$

if and only if $k_n/\log n \overset{n}{\to} \infty$. Q.E.D.

Proof of Theorems 2′ and 3′: Copying the proof of Theorems 2 and 3, we see that we need only overbound, for all $\epsilon > 0$,

$P\{\sup_{x\in B}|q_n(x) - q_{n0}(x)| > \epsilon\}$ where $q_{n0}(x) = Q(X_1^x)$. But

$$\begin{aligned}
&P\left\{\sup_{x\in B}|q_n(x) - q_{n0}(x)| > \epsilon\right\}\\
&\quad\le P\left\{\bigcup_{i=1}^{n}\{|Q(X_i) - Q_{\lambda_n}(X_i)| > \epsilon\}\right\}\\
&\quad\le n\sup_{x\in B} P\{|Q(x) - Q_{\lambda_n}(x)| > \epsilon\}\\
&\quad\le \begin{cases} 0, & \text{in the no noise case,}\\ K_1 n\exp(-K_2\lambda_n), & \text{if the noise is exponential,}\\ K_3 n/\lambda_n^{t-1}, & \text{if the noise is in } L_t,\end{cases}
\end{aligned}$$

for some $K_1,K_2,K_3 > 0$. The "in probability" part of the theorems follows trivially. For the "with probability one" version, the Borel–Cantelli lemma is used. Q.E.D.

Proof of Theorem 4: Let $\epsilon > 0$ be arbitrary and note that, with the same a, $q_{\max}$, $q_{\min}$, and A_n as in the proof of Theorem 1,

$$\begin{aligned}
&P\left\{\sup_{x\in B}|\tilde{q}_n(x) - Q(x)| > 3\epsilon\right\}\\
&\qquad\le P\{A_n^c\} + P\left\{A_n, \sup_{x\in B}|\tilde{q}_n(x) - Q(x)| > 3\epsilon\right\}.
\end{aligned}$$

Now, on A_n, $|Q(x) - Q(X_1^x)| \le \epsilon/2$ for all $x \in B$; while, also for all $x \in B$,

$$\begin{aligned}
&|Q(X_1^x) - q_{n0}(X_1^x)|\\
&\qquad\le (q_{\max} - q_{\min})\sum_{i=k_n+1}^{n} v_{ni} + \left(\sum_{i=1}^{k_n} v_{ni}\right)\epsilon/2 \le \epsilon
\end{aligned}$$

if $\Sigma_{i=k_n+1}^{n} v_{ni} \le \epsilon/2(q_{\max} - q_{\min})$. Thus with N_1 and c as in the proof of Theorem 1, we have, for all n large enough,

$$\begin{aligned}
&P\left\{\sup_{x\in B}|\tilde{q}_n(x) - Q(x)| > 3\epsilon\right\}\\
&\qquad\le 2N_1e^{-nc^2/2} + P\left\{\sup_{x\in B}|q_n(X_1^x) - q_{n0}(X_1^x)| > 3\epsilon/2\right\}.
\end{aligned}$$

The last term is overbounded by

$$\begin{aligned}
&P\left\{\bigcup_{i=1}^{n}\{|q_n(X_i) - q_{n0}(X_i)| > 3\epsilon/2\}\right\}\\
&\quad\le nP\{|q_n(X_1) - q_{n0}(X_1)| > 3\epsilon/2\}\\
&\quad\le n\sup_{x\in B} P\left\{\left|\sum_{i=1}^{n} v_{ni}(Y_i^x - Q(X_i^x))\right| > 3\epsilon/2\right\}\\
&\quad\le n\sup_{x\in B} E\left\{P\left\{\left|\sum_{i=1}^{n} v_{ni}(Y_i^x - Q(X_i^x))\right| > 3\epsilon/2|X_1^x,\cdots,X_n^x\right\}\right\}\\
&\quad\le n\sup_{\bar{x}_n=(x_1,\cdots,x_n)\in B^n}\\
&\quad\cdot P\left\{\left|\sum_{i=1}^{n} v_{ni}(Y_i - Q(X_i))\right| > 3\epsilon/2|X_1 = x_1,\cdots,X_n = x_n\right\}\\
&\quad\le \begin{cases} nK_1\exp\left(-K_2\Big/\max_i v_{ni}\right), & \text{if the noise is exponential,}\\ nK_3\left(\max_i v_{ni}\right)^{t-1}, & \text{if the noise is in } L_t (t \ge 2),\end{cases}
\end{aligned}$$

where K_1,K_2,K_3 are constants not depending upon n (use Lemmas 1, 2, and the definitions of exponential and L_t noise). This concludes the proof of Theorem 4. Q.E.D.

Proof of Theorem 5: Let $N_n = \Sigma_{i=1}^{n} Z_i$, and note that N_n has mean β_n and variance $\Sigma_{i=1}^{n}\alpha_i(1 - \alpha_i) \le \beta_n$. Thus, by Bennett's inequality for sums of independent bounded random variables [19]–[21],

$$\begin{aligned}
P\{N_n \le \beta_n/2\} &\le P\{(N_n - E\{N_n\})/n \le -\beta_n/2n\}\\
&\le \exp(-n(\beta_n/2n)^2/(2\beta_n/n + \beta_n/2n)) = \exp(-\beta_n/10).
\end{aligned}$$

Since β_n is monotone, it is clear that $\Sigma_{n=1}^{\infty} e_n^{-\beta_n/10} < \infty$ if and only if $\beta_n/\log n \xrightarrow{n} \infty$. Let $\epsilon > 0$ be arbitrary, and let a be so small that $|Q(x) - Q(z)| \le \epsilon/2$ for all z and x from B_0 with $\|x - z\| \le a$. Define further $c = \inf_{x\in B}\mu(S(x,a/2))$. Then

$$\begin{aligned}
&P\left\{\sup_{x\in B}|q_n(x) - Q(x)| > 2\epsilon\right\}\\
&\quad\le P\left\{\sup_{x\in B}|q_n(x) - q_{n0}(x)| > \epsilon, N_n > \beta_n/2\right\}\\
&\quad+ P\left\{\sup_{x\in B}|q_{n0}(x) - Q(x)| > \epsilon, N_n > \beta_n/2\right\} + P\{N_n \le \beta_n/2\}.
\end{aligned}$$

Proceeding as in the proof of Theorem 1, we have, with the same definition of N_1,

$$\begin{aligned}
&P\left\{\sup_{x\in B}|q_{n0}(x) - Q(x)| > \epsilon, N_n > \beta_n/2\right\}\\
&\quad\le P\left\{\sup_{x\in B}\|X_{k_n}^x - x\| > a, N_n > \beta_n/2\right\}\\
&\quad+ P\left\{\sup_{x\in B}|q_{n0}(x) - Q(x)| > \epsilon, \sup_{x\in B}\|X_{k_n}^x - x\| \le a\right\}\\
&\quad\le \sup_{k>\beta_n/2} P\{\text{for some } x \in B, S(x,a) \text{ contains less than } k_nX_i\text{'s}\\
&\quad\text{with } Z_i = 1|N_n = k\}\\
&\quad\text{(for all } n \text{ large enough by condition C1 and the boundedness}\\
&\quad\text{of } Q \text{ on } B_0)\\
&\quad\le 2N_1\exp(-\beta_n c^2/4)\\
&\quad\text{(for all } n \text{ large enough by } \beta_n \xrightarrow{n} \infty \text{ and } k_n/\beta_n \xrightarrow{n} 0).
\end{aligned}$$

To obtain an upper bound for $P\{\sup_{x\in B}|q_n(x) - q_{n0}(x)| > \epsilon\}$ in the proof of Theorems 2 and 3, we did not use the independence of the X_i. Thus the derived bounds remain valid for all sequences $X_1,\cdots,X_n$ of random vectors taking values in B. In particular, for some positive constants K_1,K_2,K_3 not depending upon n,

$$\begin{aligned}
&P\left\{\sup_{x\in B}|q_n(x) - q_{n0}(x)| > \epsilon\right\}\\
&\quad\le \begin{cases} 0, & \text{in the no noise case,}\\ (1+n)^{2d}K_1\exp(-K_2k_n), & \text{if the noise is exponential,}\\ (1+n)^{2d}K_3/k_n^{t-1}, & \text{if the noise is in } L_t.\end{cases}
\end{aligned}$$

Theorem 5 follows from all these inequalities and the Borel–Cantelli lemma. Q.E.D.

Proof of Theorem 6: We sketch the proof for the "with probability one" part only. The "in probability" part is proved in a similar fashion. Let $\epsilon > 0$ be arbitrary. Then

$$
\begin{aligned}
P&\left\{\bigcup_{k=n}^{\infty}\{Q(X_k^*) > q_{\min} + \epsilon\}\right\}\\
&\le P\left\{\bigcup_{k=n}^{\infty}\bigcup_{i=1}^{k}\{|q_k(W_i) - Q(W_i)| > \epsilon/4\}\right\}\\
&\quad + P\left\{\bigcup_{k=n}^{\infty}\bigcap_{i=1}^{k}\{Q(W_i) > q_{\min} + \epsilon/2\}\right\}\\
&\le P\left\{\bigcup_{k=n}^{\infty}\left\{\sup_{x\in B}|q_k(x) - Q(x)| > \epsilon/4\right\}\right\}\\
&\quad + P\left\{\bigcap_{i=1}^{n}\{Q(W_i) > q_{\min} + \epsilon/2\}\right\},
\end{aligned}
$$

in view of the fact that all the W_i take values in B with probability one. As n grows large, the first probability tends to 0. If $\theta = P\{Q(X_1) \le q_{\min} + \epsilon/2\}$, then

$$
\begin{aligned}
P&\left\{\bigcap_{i=1}^{n}\{Q(W_i) > q_{\min} + \epsilon/2\}\right\}\\
&\le P\left\{\bigcap_{i=1}^{n}\{\{W_i \text{ has d.f. } H_n\} \cup \{W_i \text{ has d.f. } G\right.\\
&\quad \left.\text{and } Q(W_i) > q_{\min} + \epsilon/2\}\}\right\}\\
&\le \prod_{i=1}^{n}(1 - \gamma_i + \gamma_i P\{Q(X_1) > q_{\min} + \epsilon/2\})\\
&\le \prod_{i=1}^{n}(1 - \gamma_i\theta) \le \prod_{i=1}^{n}\exp(-\gamma_i\theta)\\
&= \exp\left(-\theta\sum_{i=1}^{n}\gamma_i\right) \xrightarrow{n} 0. \qquad \text{Q.E.D.}
\end{aligned}
$$

References

[1] G. J. McMurtry, "Adaptive optimization procedures," in *Adaptive, Learning and Pattern Recognition Systems,* J. M. Mendel and K. S. Fu, Eds. New York: Academic, 1970.

[2] R. A. Jarvis, "Optimization strategies in adaptive control: a selective survey," *IEEE Trans. Syst., Man,Cybern.,* vol. SMC-5, no. 1, pp. 83–94, 1975.

[3] L. D. Cockrell and K. S. Fu, "On search techniques in adaptive systems," Technical Report TR-EE-70-1, Purdue University, Lafayette, IN, 1970.

[4] S. H. Brooks, "Discussion of random methods for locating surface maxima," *Operations Research,* vol. 6, pp. 244–251, 1958.

[5] G. J. McMurtry and K. S. Fu, "A variable structure automaton used as a multimodal searching technique," *IEEE Trans. Automat. Contr.,* vol. AC-11, no. 3, pp. 379–387, 1966.

[6] J. D. Hill, "A search technique for multimodal surfaces," *IEEE Trans. Syst., Sci. Cybern.,* vol. SSC-5, no. 1, pp. 2–8, 1969.

[7] R. A. Jarvis, "Adaptive global search in a time-variant environment using a probabilistic automaton with pattern recognition supervision," *IEEE Trans. Syst., Sci. Cybern.,* vol. SSC-6, no. 3, pp. 209–216, 1970.

[8] E. Parzen, "On the estimation of a probability density function and the mode," *Annals of Mathematical Statistics,* vol. 33, pp. 1065–1076, 1962.

[9] M. Rosenblatt, "Remarks on some nonparametric estimates of density function," *Annals of Mathematical Statistics,* vol. 27, pp, 832–837, 1957.

[10] E. A. Nadaraya, "On estimating regression," *Theory of Probability and Its Applications,* vol. 9, pp. 141–142, 1964.

[11] ——, "Remarks on nonparametric estimates for density functions and regression curves," *Theory of Probability and Its Applications,* vol. 15, pp. 134–137, 1970.

[12] T. M. Cover, "Estimation by the nearest neighbor rule," *IEEE Trans. Inform. Theory,* vol. IT-14, pp. 50–55, Jan. 1968.

[13] T. M. Cover and P. E. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inform. Theory,* vol. IT-13, pp. 21–27, Jan. 1967.

[14] D. O. Loftsgaarden and C. P. Quesenberry, "A nonparametric estimate of a multivariate density function," *Annals of Mathematical Statistics,* vol. 36, pp. 1049–1051, 1965.

[15] C. J. Stone, "Consistent weighted average estimators of a regression function," Manuscript, Dep. of Mathematics, Univ. California, Los Angeles, 1976.

[16] L. P. Devroye and T. J. Wagner, "Nonparametric discrimination and density estimation," Technical Rep. #183, Information Systems Research Lab, Univ. Texas, Austin, 1976.

[17] J. Matyas, "Random optimization," *Automation and Remote Control,* vol. 26, no. 2, pp. 244–251, 1965.

[18] V. N. Vapnik and A. Ya. Chervonenkis, "On the uniform convergence of relative frequencies of events to their probabilities," *Theory of Probability and Its Applications,* vol. 16, pp. 264–280, 1971.

[19] D. K. Fuk and S. V. Nagaev, "Probability inequalities for sums of independent random variables," *Theory of Probability and Its Applications,* vol. 16, pp. 643–660, 1971.

[20] G. Bennett, "Probability inequalities for the sums of independent random variables," *J. American Statistical Association,* vol. 57, pp. 33–45, 1962.

[21] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *J. American Statistical Association,* vol. 58, pp. 13–30, 1963.

[22] T. J. Wagner, "On the rate of convergence for the law of large numbers," *Annals of Mathematical Statistics,* vol. 40, pp. 2195–2197, 1969.