Convergence of a Stochastic Rootfinding Procedure

Burton Simon
Department of Mathematics
University of Colorado at Denver
January 15, 1998

Keywords: natural rootfinding procedure, cluster point property, Robbins-Monro algorithm, stochastic approximation, Kriging

ABSTRACT

Let $F \equiv (f_1, f_2, \ldots, f_m)'$ be a system of unknown functions, $f_i : B \to \Re$, where we interpret $B \subset \Re^d$ to be the parameter space of some stochastic model. Our goal is to find a "root" of F, defined to be a simultaneous solution of the equations $f_i(x) = 0,\ i = 1, 2, \ldots, m$. A "simulation" is available that can approximate $F(x)$ for any $x \in B$. A "natural rootfinding procedure" is an iterative procedure that sets the parameters of each simulation to where a root is thought to be, based on previous simulations. More exactly, the nth simulation is at $X_n \in B$, where X_n is a root of an approximating function, H_{n-1}, constructed from the data accumulated during the first $n-1$ simulations. If H_{n-1} has no roots in B then X_n is assigned randomly. The particular sequence of estimators chosen for the natural rootfinding procedure, $\{H_n\},\ n \geq 1$, therefore specifies the distribution of the random sequence of parameter settings, $\{X_n\} \in B$. We prove that under fairly mild conditions on F and the "simulation", if $\{H_n\}$ satisfies the "cluster point property" then $\| F(X_n) \| \to 0$ *a.s.* and therefore $X_n \to x^*$ *a.s.* if x^* is the unique root. We construct a simple estimator that has the cluster point property when certain likelihood ratio estimates are available, and conjecture that under more general conditions parametric estimators (regressions) and "Kriging" estimators have the cluster point property. We numerically compare the natural rootfinding procedure using Kriging estimators with the classical Robbins-Monro algorithm, which is known to have an optimal asymptotic convergence rate.

1 Introduction

Let $F : B \to \Re^m$ be a system of unknown functions, $F \equiv (f_1, f_2, \ldots, f_m)'$, where $B \subset \Re^d$ is interpreted as the "parameter space" of some stochastic model. There is a "simulation" of the model that can approximate $F(x)$ for any $x \in B$ by setting the d system parameters appropriately. After simulating at $\{X_1, X_2, \ldots, X_n\} \in B$ we have data $\{\mathbf{Y}_1, \mathbf{Y}_2, \ldots, \mathbf{Y}_n\}$ and $\{\Theta_1, \Theta_2, \ldots, \Theta_n\}$ where $\mathbf{Y}_i \equiv (\mathbf{Y}_i(1), \mathbf{Y}_i(2), \ldots, \mathbf{Y}_i(m))'$ is the estimate of $F(X_i)$ from the ith simulation and Θ_i is a random element containing other useful information available from the ith simulation such as variance/covariance estimates, derivative estimates, likelihood ratio estimates, and so on. We can construct H_n, an approximation for F, from $\{X_i, \mathbf{Y}_i, \Theta_i\},\ i \leq n$ in any of a number of ways. It is tempting to use H_n to approximate quantities of interest associated with F. Indeed, we will use the roots of H_n to approximate the roots of F, where roots are defined to be the points simultaneously satisfying $f_i(x) = 0,\ i = 1, 2, \ldots, m$. We will denote the set of roots by

$$S \equiv \{x \in B :\ F(x) = 0\}.$$

We will assume that $S \neq \emptyset$, but will not insist that S is a singleton. Problems of finding roots and finding optimal points are closely related. A rootfinding problem can be turned into the optimization problem, minimize $\sum f_i(x)^2$, and if f is differentiable then optimal points of f (in the interior of B) will satisfy $F(x) = 0$ where $F = \nabla f$.

The classical Robbins-Monro method of stochastic approximation also finds roots of systems of equations. Subject to certain technical conditions, the algorithm converges to the root (typically assumed to be unique), and the error after n simulations is $\mathcal{O}(cn^{-1/2})$ (Robbins and Monro (1951), Andradottir (1995)). Given the usual assumptions about simulation noise, only the constant c can be improved, so asymptotically (when it works) Robbins-Monro is nearly optimal. However, it is an empirical fact that Robbins-Monro can be slow in the "non-asymptotic regime", and the conditions that guarantee convergence are rather strict. Improvements have been made in the scope and short term behavior of stochastic approximation algorithms, while retaining the good asymptotic behavior, e.g., Andradottir (1995), Polyak and Juditsky (1992).

The method of finding a root of F proposed here is different from stochastic approximation. The standard stochastic approximation algorithms are Markovian in the sense that the choice of X_{n+1} depends only on X_n and $\mathbf{Y}_n$. Our procedure (which we call "natural") chooses X_{n+1} based on all available information obtained in the first n simulations, $\{(X_i, \mathbf{Y}_i, \Theta_i)\}$, $i \leq n$. Roughly speaking, a natural rootfinding procedure assigns X_{n+1} to be the "best guess" of the location of a root based on the first n simulations. A reasonable method for guessing the location of a root is to approximate F by H_n, where H_n is constructed "intelligently" from all available data, and solve the approximating problem, $H_n(x) = 0$. If the search region B is bounded then we can ensure that $\{X_n\}$ converges *a.s.* to the set of roots by ensuring that there are no cluster points of the sequence $\{X_n\}$ anywhere else. We define a property of estimators $\{H_n\}$, $n \geq 1$ that guarantees (under mild conditions on F and on the simulation) that if X_{n+1} is chosen to be a solution of $H_n(x) = 0$ then $\| F(X_n) \| \rightarrow 0$ *a.s.*, and therefore $X_n \rightarrow x^*$ *a.s.* if x^* is the unique root. (X_{n+1} is chosen randomly from B if $H_n(x) = 0$ has no solution.) We call this property the "cluster point property". The cluster point property is weaker than uniform convergence or relative compactness in a neighborhood of a cluster point, but stronger than pointwise convergence at cluster points.

In order to use a natural rootfinding procedure in practice, a simulation and a deterministic rootfinding algorithm (such as Newton's method) are needed. We assume that the deterministic algorithm finds a root of the system of equations $H_n(x) = 0$ with precision when there is a root, or states that no root exists. Our method is clearly not competitive with Robbins-Monro if constructing H_n and finding its roots is extremely time consuming since the "overhead" due to the Robbins-Monro algorithm is negligible. However, in many cases the simulation time overwhelms the overhead due to constructing and finding roots of H_n, so the efficiency of both methods can be compared in terms of the error after n simulations (or t units of computer time).

We do not derive rigorous convergence rates for our natural rootfinding procedures here. However, there are results in the nonparametric regression literature that appear (at least heuristically) to be relevant. Consider the case $m = 1$, so $F \equiv f$, $\mathbf{Y}_n \equiv Y_n$ and $H_n(x) \equiv h_n(x)$. If the successively observed points $\{X_n\}$ are i.i.d. random variables with a positive density on B, and the observations, Y_n, satisfy $E(Y_n \mid X_n, \ldots, X_1; Y_{n-1}, \ldots, Y_1) = f(X_n)$, then the pairs, $\{(X_n, Y_n)\}$, $n \geq 1$, are i.i.d versions of a generic pair (X, Y), and

$$f(x) = E(Y \mid X = x). \tag{1.1}$$

Approximations of $f(x)$ based on (1.1) and the data $\{(X_n, Y_n)\}$, $n \geq 1$ are called nonparametric regressions. Stone (1977) gives necessary and sufficient conditions (in terms of the way the $c_{ni}(x)$'s are chosen) for L^p convergence of h_n to f when

$$h_n(x) = \sum_{i=1}^{n} c_{ni}(x) Y_i. \tag{1.2}$$

Yakowitz and Szidarovszky (1985) derive convergence rates for a large class of estimates with the form (1.2). They prove that "kernel estimators" converge uniformly on B at rate $\mathcal{O}(n^{-\frac{2}{d+4}})$, which

is the theoretical optimal rate. A rootfinding scheme based on a kernel estimator and an i.i.d. sequence $\{X_n\}$ would (assuming mild regularity conditions on F) also converge at rate $\mathcal{O}(n^{-\frac{2}{d+4}})$. Note that even for $d = 1$ this is slower than Robbins-Monro, and the rate deteriorates as the number of variables increases. Although the convergence theorems in the nonparametric regression literature are not applicable to the sequence $\{X_n\}$ generated by a natural rootfinding procedure, it is difficult to believe that it would not fare as well. In fact it seems reasonable to conjecture that since the region that X_n is chosen from effectively shrinks as n increases, the order of the convergence rate is strictly better. Our numerical experiments imply that natural rootfinding procedures compare favorably with Robbins-Monro in the short term. The asymptotic convergence rate remains an open problem, though.

This paper introduces a new kind of stochastic rootfinding procedure that lends itself to rigorous analysis and appears to work well in practice. It has come to the author's attention that rootfinding schemes, similar to the ones analyzed here, are used outside academia as easily implemented and intuitively reasonable heuristic approaches for solving "inverse problems" (which are root finding problems), and optimization problems. With this in mind, perhaps only the analysis here is really new. See Pflug (1996) or Kushner and Yin (1997) for background on the more established varieties of stochastic approximation.

In the next section we describe a general mathematical framework for studying (stochastic) sequential adaptive procedures like the one analyzed here. This framework is particularly appropriate for procedures using simulation since one usually has some control over the bias and variance of the resulting data via the length (CPU time) of the simulations. In section 3 we describe the cluster point property for sequences of estimators and prove the convergence theorems for natural rootfinding procedures. In section 4 we demonstrate that in certain cases, estimators constructed as simple pointwise averages have the cluster point property when likelihood ratios allow simultaneous estimation of $F(x)$ in a neighborhood of X_n, and conjecture that "Kriging" estimators and (parametric) regressions have the cluster point property in a much more general setting. We also explicitly construct a nonparametric estimator that ensures $\limsup_n \parallel F(X_n) \parallel < \epsilon$. In section 5 we discuss pros and cons of Robbins-Monro and our new procedure, and compare the methods numerically with some simple examples.

2 A General Mathematical Framework for Simulation Based Adaptive Procedures

In our context, the goal of an adaptive procedure is to estimate some quantity of interest, q, associated with an unknown system of functions $F : B \to \Re^m$ where $B \subset \Re^d$. The procedure may be a search, for example if q is the position of a root or optimal point, or the procedure may estimate some integral of $F(x)$ over B, and so on. We assume it is possible to estimate $F(x) = (f_1(x), f_2(x), \ldots, f_m(x))'$ by simulation for any $x \in B$. The estimate of each $f_i(x)$ from the simulation is typically random and biased, but the variance and bias might be controllable. For example, as the length of a computer simulation (measured in "CPU time") increases, the bias and variance usually decrease; often to zero in the limit.

Let $\{(X_n, t_n, \mathbf{Y}_n, \Theta_n)\}$, $n \geq 1$ be a sequence of random elements with $X_n \in B$, $t_n > 0$, and $\mathbf{Y}_n \in \Re^m$. Interpret X_n to be the position of the nth simulation (i.e., the setting of the d model parameters) and t_n to be the length of the nth simulation (or some other measure of the "quality"

of the nth simulation). $\mathbf{Y}_n$ is the estimate of $F(X_n)$, and has the form

$$\mathbf{Y}_n = F(X_n) + b(X_n, t_n) + C(X_n, t_n)\chi_n \tag{2.1}$$

where χ_n is a vector of independent random variables with $E(\chi_{in}) = 0$ and $E(\chi_{in}^2) = 1$. We interpret $b_i(X_n, t_n)$ to be the bias of $\mathbf{Y}_n(i)$ and $C^2(X_n, t_n)$ to be the covariance matrix for $\mathbf{Y}_n$. These are typically unknown and need to be estimated if they are involved in the adaptive procedure. Θ_n is a collection of random elements such as estimates of $b(X_n, t_n)$ and $C^2(X_n, t_n)$, derivative estimates, likelihood ratio estimates, and so on.

Remark 1 *A reasonable meta-model for a large class of simulations has*

$$b_i(X_n, t_n) \equiv t_n^{-1} b_i(X_n),$$

$$C^2(X_n, t_n) \equiv t_n^{-1} C^2(X_n),$$

and χ_n is a vector of i.i.d. standard Normal random variables (see Iglehart (1978), Koehler, Puhalskii and Simon (1997)).

Let $(\Omega, \mathcal{F}, \mathcal{P})$ be a probability space for the simulations and let $\mathcal{F}_n \subset \mathcal{F}$ contain (among other things) all the information associated with the first n simulations. We allow X_n and t_n to be $\mathcal{F}_{n-1}$ measurable, but the distributions of $\mathbf{Y}_n$ and Θ_n depend only on X_n and t_n. $\mathcal{F}_n$ can contain auxiliary random variables not associated with the simulations. For example, we will need an auxiliary random variable in our root finding scheme since we assign X_{n+1} randomly in B if H_n (described below) has no root.

Let q_n be a $\mathcal{F}_n$ measurable random variable which we interpret as our estimate of the quantity of interest q based on the first n simulations. Let $H_n : B \to \Re^m$, $n = 1, 2, \ldots$ be a sequence of (Lebesgue) measurable functions where for each n and $x \in B$, $H_n(x)$ is $\mathcal{F}_n$ measurable. We interpret H_n as our estimate of F based on the first n simulations. In many cases of interest q_n is evaluated from H_n. For example, we use the roots of H_n as approximations for the roots of F. Our algorithm therefore assigns q_n to be a solution of the nth problem, $H_n(x) = 0$, and then assigns X_{n+1} equal to q_n. The sequences $\{X_n\}$ and $\{q_n\}$ are redundant in our case so there is no reason to work with both of them. To avoid unnecessary notation we will work with the sequence $\{X_n\}$.

Remark 2 *In order for the sequence $\{X_n\}$ to be well defined we need to specify how X_{n+1} is chosen if H_n has more than one root. If B is compact then the set of roots of H_n in B is also compact and therefore has a least element under (for example) dictionary order in the d coordinates. We can therefore use the least element as a choice function. In practice, most deterministic algorithms return the first root they find, which can be thought of as a built in choice function.*

3 Main Results

Our goal is to show that the sequence $\{X_n\}$ of parameter settings converges to the solution set, S. For a natural rootfinding procedure to converge, certain mild restrictions must be imposed on F and the simulations that estimate it; but the crucial restriction is on the sequence of estimators, $\{H_n\}$.

3.1 The Cluster Point Property

We begin by defining what we mean by a cluster point. Let $\| \cdot \|$ be Euclidian distance when applied to a vector, and the corresponding matrix norm when applied to a matrix and define

$$B_\alpha(x) = \{y \in B \ : \ \| x - y \| < \alpha\}.$$

Definition 1 *A point $x \in \overline{B}$ is a cluster point of $\{(X_n, t_n)\}$ if*

$$\forall \epsilon > 0, \quad \sum_{\{i: \ \|X_i - x\| < \epsilon\}} t_i = \infty.$$

Let $\mathcal{L}$ denote the set of cluster points of $\{(X_n, t_n)\}$.

In other words, x is a cluster point if an infinite amount of time is spent simulating in every neighborhood of x. In this paper we will assume that there is some $t_{min} > 0$ such that each $t_i > t_{min}$, so x is a cluster point if it is a cluster point of the sequence $\{X_n\}$. Note that $\mathcal{L}$ is a closed set, so if B is bounded then $\mathcal{L}$ is compact.

Definition 2 *A simulation is "uniformly consistent" on B if the bias and variance terms in (2.1) converge to zero uniformly on B as the length of the simulation increases to infinity, i.e., for every $\epsilon > 0$ there is a $T_\epsilon < \infty$ such that when $t \geq T_\epsilon$, $|b_i(x,t)| < \epsilon$ and $C^2_{ii}(x,t) < \epsilon$, $i = 1, 2, \ldots, m$, for every $x \in B$.*

We assume that our simulations are uniformly consistent on B. In a queueing context the condition is typically satisfied if the system is stable for every $x \in \overline{B}$.

Definition 3 *A sequence of approximations $\{H_n\}$ for a continuous function F has the "cluster point property" if for each n, H_n is continuous, $H_n(x)$ is $\mathcal{F}_n$ measurable for eaxh $x \in B$, and*

$$\forall x \in \mathcal{L}, \quad \lim_{\alpha \to 0} \limsup_n \sup_{y \in B_\alpha(x)} \| H_n(y) - F(y) \| = 0 \quad a.s. \tag{3.1}$$

We will use the notation

$$\{H_n\} \in \mathcal{C}$$

to mean $\{H_n\}$ has the cluster point property. If $\{H_n\}$ satisfies the weaker condition,

$$\forall x \in \mathcal{L}, \quad \lim_{\alpha \to 0} \limsup_n \sup_{y \in B_\alpha(x)} \| H_n(y) - F(y) \| < \epsilon \quad a.s. \tag{3.2}$$

we say that $\{H_n\} \in \mathcal{C}_\epsilon$.

It follows immediately from the definition that $\{H_n\} \in \mathcal{C}$ if and only if $\{H_n\} \in \mathcal{C}_\epsilon$ for every $\epsilon > 0$. Also, if $\{H_n\} \in \mathcal{C}$ then $H_n(x) \to F(x)$ *a.s.* whenever x is a cluster point, although the converse is clearly not true in general. A useful sufficient condition for $\{H_n\} \in \mathcal{C}$ is given by the following lemma.

Lemma 1 *If F is continuous and for each $x \in \mathcal{L}$,*
(i) $H_n(x) \to F(x)$, a.s., and
(ii) $\{H_n\}$ is a.s. relatively compact in a neighborhood of x.
then $\{H_n\} \in \mathcal{C}$.

Proof: Let $x \in \mathcal{L}$ and suppose $\{H_n\}$ is *a.s.* relatively compact in $B_\delta(x)$. From the Arzella-Ascolli Theorem we have

$$\lim_{\alpha \to 0} \limsup_n \sup_{\substack{|s-t| < \alpha \\ s,t \in B_\delta(x)}} \| H_n(s) - H_n(t) \| = 0 \quad a.s. \tag{3.3}$$

To verify (3.1) we write

$$\lim_{\alpha \to 0} \limsup_n \sup_{y \in B_\alpha(x)} \| H_n(y) - F(y) \|$$

$$\leq \lim_{\alpha \to 0} \limsup_n \sup_{y \in B_\alpha(x)} (\| H_n(y) - H_n(x) \| + \| H_n(x) - F(x) \| + \| F(x) - F(y) \|) . \tag{3.4}$$

Equation (3.3) implies that the first quantity in (3.4) goes to zero *a.s.*. Condition (i) and the continuity of F take care of the other terms.
□

Remark 3 *$\{H_n\} \in \mathcal{C}$ is strictly weaker than relative compactness. For example (in one dimension) let $\mathcal{L} = \{0\}$, $f(x) = x$ and $h_n(x) = x \sin \frac{1}{nx}$. Then the limit in (3.3) is $2\delta > 0$ so $\{h_n\}$ is not relatively compact in any neighborhood of x. However, $\{h_n\}$ does satisfy (3.1), so $\{h_n\} \in \mathcal{C}$.*

Lemma 2 *Let F be continuous and let $\{U_n\}$, $n \geq 1$ be a sequence of i.i.d. random variables with positive densities satisfying $P(U_n \in B) = 1$. If $\{H_n\} \in \mathcal{C}$, B is compact, and there is a sequence $n_i \to \infty$ where $X_{n_i} = U_{n_i}$ then $H_n \to F$ uniformly on B, a.s..*

Proof: Since $\Re^d$ is separable we can find a countable base $\{\mathcal{O}_1, \mathcal{O}_2, \ldots\}$ for the relative topology on B. Let $X = \{X_1, X_2, \ldots\}$. Then $\mathcal{L} = B$ if for every i, $X \cap \mathcal{O}_i \neq \emptyset$. Since U_n has a positive density on B, for each i, $P(U_n \in \mathcal{O}_i) > 0$ and so $P(X \cap \mathcal{O}_i = \emptyset) = 0$. Thus

$$P(\mathcal{L} \neq B) \leq \sum_{i=1}^{\infty} P(X \cap \mathcal{O}_i = \emptyset) = 0.$$

Choose $\epsilon > 0$. Since $\{H_n\} \in \mathcal{C}$, for each $x \in B$ there is an $\alpha_x > 0$ such that

$$\limsup_n \sup_{y \in B_{\alpha_x}(x)} \| H_n(y) - F(y) \| < \epsilon \quad a.s.$$

Since B is compact we can find $\{x_1, x_2, \ldots, x_k\} \in B$ such that

$$\bigcup_{i=1}^{k} B_{\alpha_{x_i}}(x_i) = B.$$

Let $N_i < \infty$ be large enough so that $n > N_i$ implies

$$\sup_{y \in B_{\alpha_{x_i}}(x_i)} \| H_n(y) - F(y) \| < \epsilon$$

Thus, if $n > \max\{N_1, N_2, \ldots, N_k\}$ then

$$\sup_{y \in B} \| H_n(y) - F(y) \| < \epsilon$$

Since ϵ is arbitrary the result follows.
□

3.2 Convergence Theorems for Natural Rootfinding Procedures

We now show that an intuitively appealing algorithm for finding roots of $F = (f_1, f_2, \ldots, f_m)'$ based on $\{H_n\}$ converges *a.s.* if the H_n's satisfy the cluster point property.

Definition 4 *Let $\{U_n\}$, $n \geq 0$ be a sequence of i.i.d. random variables with positive densities satisfying $P(U_n \in B) = 1$. The "natural root finding procedure based on $\{H_n\}$" assigns $\{X_n\}$ as follows. $X_1 = U_0$, and then for $n > 1$,*

- *If H_{n-1} has a root then X_n is chosen to be a solution of $H_{n-1}(x) = 0$,*
- *Otherwise, $X_n = U_{n-1}$.*

Remark 4 *U_{n-1} is an auxiliary random variable that is not associated with $\mathbf{Y}_{n-1}$ or Θ_{n-1}. We use U_{n-1} instead of U_n as a convention so that X_n is $\mathcal{F}_{n-1}$ measurable.*

Each X_n is therefore "assigned" or "guessed randomly" depending on whether or not $H_{n-1}(x) = 0$ has a solution. Define $\mathcal{A}$ and $\mathcal{G}$ by

$$\mathcal{A} = \{n : \ H_{n-1}(X_n) = 0\}$$

and

$$\mathcal{G} = \{n : \ X_n = U_{n-1}\}.$$

Theorem 1 *Let $\mathcal{L}_{\mathcal{A}}$ be the set of cluster points of $\{X_n\}_{n \in \mathcal{A}}$. If*
(a) F is continuous,
(b) B is compact,
(c) $\{H_n\} \in \mathcal{C}$, and
(d) $\{X_n\}$ is chosen by the natural rootfinding procedure using $\{H_n\}$,
then $\mathcal{L}_{\mathcal{A}} \subset S$.

Proof: Let $S_\epsilon = \{x \in B : \ \| F(x) \| < \epsilon\}$ and define

$$\mathcal{L}_\epsilon = \mathcal{L} - S_\epsilon.$$

Since B is compact we can choose $\eta > 0$ small enough so that for $x, y \in B$, if $\| x - y \| < \eta$ then $\| F(x) - F(y) \| < \epsilon/3$. Let

$$\hat{\mathcal{L}}_\epsilon = \{y \in B : \ (\exists x \in \mathcal{L}_\epsilon)[\| x - y \| < \eta]\},$$

so that

$$\inf_{y \in \hat{\mathcal{L}}_\epsilon} \| F(y) \| > 2\epsilon/3. \tag{3.5}$$

The cluster point property (3.1) gives us

$$(\forall x \in \mathcal{L}_\epsilon)(\exists \alpha_x > 0, N_x < \infty)[n > N_x \Rightarrow \sup_{y \in B_{\alpha_x}} \| H_n(y) - F(y) \| < \epsilon/3] \quad a.s.$$

Without loss of generality we can take $\alpha_x < \eta$, so

$$\mathcal{L}_\epsilon \subset \bigcup_{x \in \mathcal{L}_\epsilon} B_{\alpha_x}(x) \subset \hat{\mathcal{L}}_\epsilon.$$

Since $\mathcal{L}_\epsilon$ is compact, for some $k < \infty$ we can find $\{x_1, x_2, \ldots, x_k\} \in \mathcal{L}_\epsilon$ so that

$$\mathcal{L}_\epsilon \subset \mathcal{L}_\epsilon^* \equiv \bigcup_{i=1}^{k} B_{\alpha_{x_i}}(x_i) \subset \bigcup_{x \in \mathcal{L}_\epsilon} B_{\alpha_x}(x) \subset \hat{\mathcal{L}}_\epsilon. \tag{3.6}$$

Let $N = \max(N_{x_1}, N_{x_2}, \ldots, N_{x_k})$. It follows that for $n > N$,

$$\sup_{y \in \mathcal{L}_\epsilon^*} \| H_n(y) - F(y) \| < \epsilon/3. \tag{3.7}$$

From (3.5), (3.6) and (3.7) it follows that for $n > N$

$$\sup_{y \in \mathcal{L}_\epsilon^*} \| H_n(y) \| \geq \inf_{y \in \mathcal{L}_\epsilon^*} \| F(y) \| - \sup_{y \in \mathcal{L}_\epsilon^*} \| H_n(y) - F(y) \| > \epsilon/3$$

so that if $n > N$ and $n \in \mathcal{A}$ then (d) implies that $X_n \notin \mathcal{L}_\epsilon^*$ which shows that $\mathcal{L}_\epsilon = \emptyset$. Since ϵ is arbitrary the result follows.
□

Corollary 1.1 *If condition (c) in theorem 1 is replaced by $\{H_n\} \in \mathcal{C}_\epsilon$ then $\mathcal{L}_\mathcal{A} \subset S_\epsilon$.*

Proof: Let $\delta > 0$ and choose $\eta > 0$ so that $\| x - y \| < \eta \Rightarrow \| F(x) - F(y) \| < \delta/2$. Since $\{H_n\} \in \mathcal{C}_\epsilon$ we have

$$(\forall x \in \mathcal{L}_{\epsilon+\delta})(\exists \alpha_x > 0, N_x < \infty)[n > N_x \Rightarrow \sup_{y \in B_{\alpha_x}(x)} \| H_n(y) - F(y) \| < \epsilon] \quad a.s.$$

Following the proof of Theorem 1 we get

$$\sup_{y \in \mathcal{L}_{\epsilon+\delta}^*} \| H_n(y) - F(y) \| < \epsilon \quad \text{and} \quad \inf_{y \in \mathcal{L}_{\epsilon+\delta}^*} \| F(y) \| > \epsilon + \delta/2,$$

so

$$\sup_{y \in \mathcal{L}_{\epsilon+\delta}^*} \| H_n(y) \| > \delta/2.$$

Since δ is arbitrary, $\{X_n\}_{n \in \mathcal{A}}$ is finite outside every neighborhood of S_ϵ and the corollary follows.
□

Corollary 1.2 *Let $m = 1$ so that $F \equiv f$ and $H_n \equiv h_n$ (d is arbitrary). If in addition to conditions (a), (b), (c) and (d) of theorem 1 we have*

(e1) there exists $x^+, x^- \in B$ such that $f(x^+) > 0$ and $f(x^-) < 0$, and

(f1) B is connected, then

$$|f(X_n)| \to 0 \quad a.s.$$

and so if x^* is the unique root then

$$X_n \to x^* \quad a.s.$$

Proof: In view of theorem 1 it is sufficient to show that $\mathcal{G}$ is finite *a.s.*. Suppose not, so that $h_n(x) = 0$ has no solution infinitely often. Then there is a sequence n_i such that $X_{n_i} = U_{n_i - 1}$. From lemma 2 we conclude that there is *a.s.* an $M < \infty$ such that if $n > M$ then $h_n(x^+) > 0$ and $h_n(x^-) < 0$. Since B is connected and each h_n is continuous the intermediate value theorem implies that h_n has a root for every $n > M$, contradicting the hypothesis.
□

When $m > 1$ one cannot use the intermediate value theorem directly to ensure the existence of a root. The technical conditions on F apparently need to be more restricting to ensure $\mathcal{G}$ is finite.

Theorem 2 *Suppose $F : B \to \Re^d$ has a root at $x^* \in B$, and in addition to conditions (a), (b), (c) and (d) of theorem 1 we can find $\eta > 0$ so that*

(e2) the Jacobian matrix for $F(x)$,

$$J_{ij}(x) = \frac{\partial f_i}{\partial x_j}(x),$$

is continuous in $B_\eta(x^*)$ and is nonsingular at x^*, and

(f2) there exists $c < \infty$ such that for $x \in B_\eta(x^*)$,

$$\| x^* - \phi(x) \| < c \| x^* - x \|^2,$$

where

$$\phi(x) = x - AF(x),$$

and

$$A = J(x^*)^{-1},$$

then

$$\| F(X_n) \| \to 0 \quad a.s.,$$

and if x^* is the unique root then

$$X_n \to x^* \quad a.s.$$

Proof: As with corollary 1.1 it suffices to show that $\mathcal{G}$ is finite *a.s.*. Let

$$\beta < \min(\eta, \frac{1}{2c}).$$

Conditions (e2) and (f2) imply that $\phi(x)$ is a contraction on $B_\beta(x^*)$ with

$$\phi(B_\beta(x^*)) \subseteq B_{\beta/2}(x^*). \tag{3.8}$$

Suppose that $\mathcal{G}$ is infinite. Lemma 2 implies that *a.s.* for some $M < \infty$, if $n > M$ then

$$\sup_{x \in B_\beta(x^*)} \| H_n(x) - F(x) \| \quad < \quad \| A \|^{-1} \beta/2. \tag{3.9}$$

Let

$$\begin{aligned} \psi_n(x) &= x - AH_n(x) \\ &= \phi(x) - A(H_n(x) - F(x)). \end{aligned} \tag{3.10}$$

Since A is nonsingular, a fixed point of ψ_n is a solution of $H_n(x) = 0$. From (3.8), (3.9) and (3.10) we have for $n > M$ and each $x \in B_\beta(x^*)$,

$$\| x^* - \psi_n(x) \| \quad \leq \quad \| x^* - \phi(x) \| + \| A \| \cdot \| H_n(x) - F(x) \| \quad < \quad \beta,$$

so

$$\psi_n(B_\beta(x^*)) \subset B_\beta(x^*).$$

Brouwer's fixed point theorem implies that ψ_n has a fixed point in $B_\beta(x^*)$, so $H_n(x) = 0$ has a solution for $n > M$, and $\mathcal{G}$ is therefore finite *a.s.*.

□

Remark 5 *If (e2) is satisfied then a sufficient condition for (f2) is a $K < \infty$ such that*

$$\max_{i,j,k} \sup_{x \in B_\eta(x^*)} \left| \frac{\partial^2 f_i}{\partial x_j \partial x_k}(x) \right| < K. \tag{3.11}$$

The constant c in (f2) can be bounded in terms of K and $\| A \|$, (Asaithambi (1995)).

Corollary 2.1 *Suppose $F : B \to \Re^m$, $m < d$ has a root at $x^* \in B$, and in addition to conditions (a), (b), (c) and (d) of theorem 1 we can find $\eta > 0$ so that*

(e3) the $m \times d$ Jacobian matrix J given by $J_{ij} = \frac{\partial f_i}{\partial x_j}$ is continuous in $B_\eta(x^*)$,

(f3) rank$[J(x^*)] = m$, and

(g3) the second order derivatives of F are bounded as in (3.11),

then

$$\| F(X_n) \| \to 0 \quad a.s.$$

Proof: From condition (f3) we can find a $(d-m) \times d$ matrix W so that

$$\hat{J}(x) = \begin{bmatrix} J(x) \\ W \end{bmatrix}$$

is nonsingular at x^*. Define "auxiliary" functions

$$f_{m+\ell}(x) = \sum_{j=1}^{d} (x_j - x_j^*) W_{\ell j}, \quad \ell = 1, 2, \ldots, d-m.$$

The auxiliary functions have roots at x^* and $\hat{J}$ is the Jacobian matrix of the augmented set of functions

$$\hat{F} = (F'; f_{m+1}, \ldots, f_d)'.$$

Since the second order derivatives of $f_{m+\ell}$ are zero, we conclude from Remark 5 that $\hat{F}$ satisfies conditions (e2) and (f2) of Theorem 2. As before, it suffices to show that $\mathcal{G}$ is finite *a.s.*.

Choose β as in the proof of theorem 2 and note that if $\mathcal{G}$ is infinite then (3.9) remains valid for $n > M$. If we define

$$\hat{H}_n = (H_n'; f_{m+1}, \ldots, f_d)'$$

then

$$\| H_n(x) - F(x) \| \;\; = \;\; \| \hat{H}_n(x) - \hat{F}(x) \|,$$

so

$$\hat{\psi}_n(x) = x - A\hat{H}_n(x)$$

continuously maps $B_\beta(x^*)$ into itself and therefore has a fixed point. This implies that $\hat{H}_n$ has a root for $n > M$ and so H_n also has a root for $n > M$. We conclude that $\mathcal{G}$ must be finite *a.s.* and the corollary is proven.

□

Corollary 2.2 *Suppose the conditions of theorem 2 hold except that $\{H_n\} \in \mathcal{C}_\epsilon$, where $\epsilon < \frac{\min(\eta, \frac{1}{2c})}{2\|A\|}$, then*

$$\limsup_n \| F(X_n) \| < \epsilon \quad a.s.$$

Proof: In view of corollary 1.1 it suffices to show that $\mathcal{G}$ is finite *a.s.*. Our hypotheses assure us that (3.9) is still valid, so the proof that $\mathcal{G}$ is finite *a.s.* in theorem 2 is valid here as well.

□

4 Existence of Estimators Satisfying the Cluster Point Properties

In this section we restrict ourselves to the case $m = 1$ without loss of generality since $\{H_n\} \in \mathcal{C}$ if and only if $\{h_{in}\} \in \mathcal{C},\ i = 1, 2, \ldots, m$. To keep notation to a minimim we will use $F \equiv f$, $\mathbf{Y}_i \equiv Y_i$ and $H_n \equiv h_n$.

Consider fitting a sequence of straight lines $\{h_n\}$ to the data $\{X_i, Y_i\},\ i \leq n$ by weighted least-squares, and suppose the sequence $\{X_n\}$ has cluster points at x_1^*, x_2^* and x_3^*. Unless $(x_i^*, f(x_i^*))\ i = 1, 2, 3$ are colinear it is impossible to have $h_n(x_i^*) \to f(x_i^*)$ at all three points. In order to show that linear estimators (or any other parametric estimator) has the cluster point property one must (among other things) show that the kind of scenario just described is impossible. Of course, if $\{h_n\}$ has the cluster point property then theorem 2 shows that the scenario is impossible since $\mathcal{L} \subset S$; but the argument is circular.

We have thus far been unable to construct a noncircular argument yielding sufficient conditions for a parametric estimator to have the cluster point property. However, we conjecture that under suitable conditions least squares estimators have the cluster point property even if f is not a member of the parametric family. All our numerical work supports the conjecture. We now discuss some examples of estimators that can be used in natural rootfinding procedures. We first describe Kriging estimators, which appear to be ideal in most circumstances, although we do not have a proof that they satisfy the cluster point property. We then explicitly construct an estimator which (provably) has the cluster point property when certain likelihood ratios are available from the simulation, and an estimator in the class $\mathcal{C}_\epsilon$ that is applicable in a very general setting. The two estimators we found that provably satisfy a cluster point property are nonparametric and the Y_i values near one cluster point cannot adversely affect the estimate at another cluster point.

4.1 Kriging Estimators

Readers unfamiliar with the "Kriging" method are referred to Ripley (1989) for a historical overview. (The technique was virtually unknown outside the geostatistics community from the time of its invention in the 1950's until the 1980's.) Stein (1990), Yakowitz and Szidarovszky (1985) and Koehler, Puhalskii and Simon (1997) discuss mathematical issues associated with Kriging estimators that are relevant to our present discussion.

Suppose we consider f to be a realization of a random function, Z. If Z is an "intrinsic random function" (Yakowitz and Szidarovszky (1985)) then one can write down an expression for the "best linear unbiased predictor" of Z based on the (simulation) data. The assumption that Z is an intrinsic random function means that

$$Z(x) = m(x) + \xi(x), \quad x \in B$$

where $\xi(x)$ is a zero mean random process with covariance function

$$\rho(x, y) = \mathrm{Cov}(\xi(x), \xi(y)),$$

and $m(x) = E(Z(x))$ has the form

$$m(x) = \sum_{i=1}^{k} \beta_i \phi_i(x),$$

where

$$\phi(x) = (\phi_1(x), \ldots, \phi_k(x))'$$

is given, and $\beta = (\beta_1, \ldots, \beta_k)$ are unknown constants. Let

$$h_n(x) = \left[\gamma_n(x)\Gamma_n^{-1} + (\phi(x)' - \Phi_n\Gamma_n^{-1}\gamma_n(x))'(\Phi_n\Gamma_n^{-1}\Phi_n)^{-1}\Phi_n\Gamma_n^{-1}\right]\vec{Y}_n,$$

where

$$\vec{Y}_n = (Y_1, Y_2, \ldots, Y_n)',$$
$$\Phi_n = (\phi(X_1), \phi(X_2), \cdots, \phi(X_n)),$$
$$\gamma_n(x) = (\rho(x, X_1), \rho(x, X_2), \ldots, \rho(x, X_n)),$$

and

$$\Gamma_n = R_n + \sigma_n$$

where

$$R_n = (\gamma_n(X_1)', \gamma_n(X_2)', \cdots, \gamma_n(X_n)')$$

and

$$\sigma_n(i, j) = \mathrm{Cov}(Y_i, Y_j).$$

Then $h_n(x)$, $x \in B$ is the best linear unbiased predictor of Z given $\vec{Y}_n$, and is called a "universal Kriging predictor". Note that if $m(x) = 0$ then $h_n(x)$ has the simple form

$$h_n(x) = \gamma_n(x)\Gamma_n^{-1}\vec{Y}_n. \tag{4.1}$$

If Z is a Gaussian random function then $h_n(x)$ is the minimum variance unbiased estimator of any kind, i.e.,

$$h_n(x) = E(Z(x) \mid \mathcal{F}_n).$$

Thus, when Z is Gaussian the Kriging method can be thought of as a kind of Bayesian inference, where Z is the "prior" on f and h_n is the mean of the "posterior" distribution.

Of course f is not really a random function (in most applications) so it is interesting to consider whether there is any validity to the Kriging method. It is the author's opinion that Kriging is no less valid than least squares regression or any other parametric estimation technique. In fact, Kriging can be thought of as a parametric estimation technique where the parameters specify the distribution of Z.

Kriging can also be thought of as a nonparametric estimator (Yakowitz and Szidarovszky (1985)). If $\sigma_n(i, i) = 0$ (i.e., $Y_i = f(X_i)$, $a.s.$) then (4.1) yields $h_n(X_i) = f(X_i)$. Thus, in principle $h_n(x)$ can be made arbitrarily close to f at arbitrarily many points in B. There are results in the literature much stronger than this simple observation. If f really is an intrinsic random function, Yakowitz and Szidarovszky (1985) and Stein (1990) show that $h_n(x) \to f(x)$ in L^2 at cluster points even if the covariance function $\rho(x, y)$ is misspecified. Koehler, Puhalskii and Simon (1997) show L^2 convergence of integrals of $h_n(x)$ to the corresponding integral of $f(x)$ for any choice of $\rho(x, y)$ for "generalized designs" (closely related to $\{X_n\}$ being dense), even if f is not a random function (for f in a dense subset of L^2). These results show that $\{h_n\}$ is an excellent estimator near cluster points, but they are not directly relevant to deciding if $\{h_n\}$ has the cluster point property.

4.2 An Estimator Utilizing Likelihood Ratio Estimates

Consider a stochastic model involving (among other things) a Poisson process. Construct an appropriate probability space for the model, and let P_λ be the probability measure associated with the model when the rate of the Poisson process is λ. Let T be a stopping time and let ϕ be a function of the sample paths that depends on events up to time T. Let N be the number of Poisson events up to time T. Suppose we wish to find a root of

$$f(\lambda) \equiv E_\lambda(\phi) \equiv \int \phi dP_\lambda, \quad \lambda \in [a,b],$$

where $0 < a < b < \infty$. It is well known (e.g., Reiman and Weiss (1989)) that for any $\lambda, \lambda' \in [a,b]$,

$$f(\lambda) = E_{\lambda'}[\psi(\lambda', \lambda)], \tag{4.2}$$

where

$$\psi(\lambda', \lambda) = \phi e^{(\lambda' - \lambda)T} (\frac{\lambda}{\lambda'})^N,$$

as long as the expected value exists. Let λ_n be the setting of the Poisson rate for the nth simulation, let ϕ_n be the resulting estimate of $f(\lambda_n)$, and define

$$\psi_n(\lambda) = \phi_n e^{(\lambda_n - \lambda)T_n} (\frac{\lambda}{\lambda_n})^{N_n}, \quad \lambda \in [a,b]$$

to be the likelihood estimate of $f(\lambda)$ based on the nth simulation. We will assume (for now) that there is a $K < \infty$ so that

$$\sup_{\lambda, \lambda' \in [a,b]} E(|\psi_n(\lambda)| \mid \lambda_n = \lambda') < K, \qquad \sup_{\lambda, \lambda' \in [a,b]} \mathrm{Var}(\psi_n(\lambda) \mid \lambda_n = \lambda) < K. \tag{4.3}$$

Finally, let

$$h_n(\lambda) = n^{-1} \sum_{i=1}^{n} \psi_i(\lambda) \tag{4.4}$$

be our estimator for $f(\lambda)$ based on the first n simulations.

Theorem 3 *Let* $\hat{\psi} = \phi e^{(b-a)T} \left(\frac{b}{a}\right)^N$. *Suppose*

$$\lim_{\alpha \to 0} \sup_{\lambda \in [a,b]} E_\lambda(\hat{\psi}|(1+\alpha)^N - 1|) = 0, \tag{4.5}$$

$$\lim_{\alpha \to 0} \sup_{\lambda \in [a,b]} E_\lambda(\hat{\psi}|e^{\alpha T} - 1|) = 0, \tag{4.6}$$

and there is a $K < \infty$ *and* $\alpha^* > 0$ *such that*

$$\sup_{\alpha < \alpha^*} \sup_{\lambda \in [a,b]} Var_\lambda(\hat{\psi}|(1+\alpha)^N - 1|) < K, \tag{4.7}$$

$$\sup_{\alpha < \alpha^*} \sup_{\lambda \in [a,b]} Var_\lambda(\hat{\psi}|e^{\alpha T} - 1|) < K. \tag{4.8}$$

Then $\{h_n\} \in \mathcal{C}$.

Proof: Let λ^* be a cluster point of $\{\lambda_n\}$. In view of lemma 1 it suffices to show that $h_n(\lambda^*) \to f(\lambda^*)$ *a.s.* and $\{h_n(\lambda)\}$ is relatively compact on $[a,b]$. We write

$$h_n(\lambda^*) = f(\lambda^*) + n^{-1}\sum_{i=1}^{n}(\psi_i(\lambda^*) - f(\lambda^*)). \tag{4.9}$$

From (4.3) we have $\mathrm{Var}(\psi_i(\lambda^*))$ is uniformly bounded, so the sum in (4.9) goes to zero *a.s.*, (Durrett (1996), Chapter 1, theorem 8.3). To show relative compactness, let $s < s' \in [a,b]$ and write

$$\begin{aligned} |h_n(s') - h_n(s)| &\leq n^{-1}\sum_{i=1}^{n}|\psi_i(s') - \psi_i(s)| \\ &\leq n^{-1}\sum_{i=1}^{n}\hat{\psi}_i\left|\left(\frac{s'}{s}\right)^{N_i} - 1\right| + n^{-1}\sum_{i=1}^{n}\hat{\psi}_i|e^{(s'-s)T_i} - 1|. \end{aligned} \tag{4.10}$$

Choose $\epsilon > 0$ and $\alpha > 0$ small enough so that the expected values in (4.5) and (4.6) are less than $\epsilon/2$. Choose $0 < \delta < \min(b\alpha, \alpha^*)$. If $s' - s < \delta$ then for every $\lambda \in [a,b]$,

$$g_1(\lambda) \equiv E_\lambda(\hat{\psi}(e^{(s'-s)T} - 1)) < \epsilon/2 \tag{4.11}$$

and

$$g_2(\lambda) \equiv E_\lambda(\hat{\psi}((\frac{s'}{s})^N - 1)) < \epsilon/2. \tag{4.12}$$

Thus we can rewrite (4.10) as

$$\begin{aligned} |h_n(s') - h_n(s)| &\leq n^{-1}\sum_{i=1}^{n}(\hat{\psi}_i((\frac{s'}{s})^{N_i} - 1) - g_1(\lambda_n)) + g_1(\lambda_n) \\ &+ n^{-1}\sum_{i=1}^{n}(\hat{\psi}_i(e^{(s'-s)T_i} - 1) - g_2(\lambda_n)) + g_2(\lambda_n) \end{aligned}$$

Assumptions (4.7), (4.8), along with (4.11) and (4.12) imply that

$$\limsup_n \sup_{|s'-s|<\delta} |h_n(s') - h_n(s)| < \epsilon \quad a.s. \tag{4.13}$$

Since ϵ is arbitrary the Arzella-Ascolli Theorem along with (4.13) implies relative compactness. □

In practice equation (4.2) may fail if $|\lambda' - \lambda|$ is too big, so some of the $\psi_i(\lambda)$'s in (4.4) might "ruin" the estimator. One can alleviate the problem by defining $\psi_n(\lambda) = 0$ outside a small neighborhood of λ_n. The resulting estimator still has the cluster point property.

Corollary 3.1 *Let $\delta > 0$ and define $\hat{h}_n(\lambda)$ to be any continuous function satisfying*

$$\hat{h}_n(\lambda) = \frac{1}{\#(I_n^\delta(\lambda))}\sum_{i \in I_n^\delta(\lambda)}\psi_i(\lambda), \quad \text{if } \#(I_n^\delta(\lambda)) > 0,$$

where $I_n^\delta(\lambda) = \{i \leq n : \ |\lambda_i - \lambda| \leq \delta\}$ and $\#(\cdot)$ is the number of elements in a set. Then $\{\hat{h}_n\} \in \mathcal{C}$.

Proof: Since $\{\lambda : \ I_n^\delta(\lambda) > 0\}$ is a finite union of closed intervals and each ψ_i is continuous, $\hat{h}_n$ extends (trivially) to $[a,b]$ by the Tietze Extension Theorem (linearly interpolate between the endpoints). The fact that $\hat{h}_n$ satisfies the conditions of lemma 1 follows from the proof of theorem 3 since $\#(I_n^\delta(\lambda)) \to \infty$ for $|\lambda - \lambda^*| < \delta$ when λ^* is a cluster point. □

4.3 A Nonparametric Estimator in $\mathcal{C}_\epsilon$

We now to explicitly construct an example of $\{h_n\} \in \mathcal{C}_\epsilon$ that is valid in a very general setting. For technical reasons (in the proof of theorem 4) we will need to truncate the simulation data $\{Y_n\}$. Define

$$\tilde{Y}_n^M = \begin{cases} Y_n & \text{if} \quad |Y_n| < M \\ M & \text{if} \quad Y_n \geq M \\ -M & \text{if} \quad Y_n \leq -M \end{cases}$$

Since $E(|Y_n|) < \infty$ and the simulations are uniformly consistent, for any $\alpha > 0$ we can choose $M < \infty$ and $T < \infty$ large enough so that

$$E(|\tilde{Y}_n^M - f(X_n)| \mid X_n = x, t_n = T) < \alpha \tag{4.14}$$

Choose $\delta > 0$, let $\#(\cdot)$ denote the number of elements in a set, and define

$$g_n(x) = \begin{cases} \frac{1}{\#(I_n^\delta(x))} \sum_{i \in I_n^\delta(x)} \tilde{Y}_i^M & \text{if } \#(I_n^\delta(x)) > 0 \\ \\ \tilde{Y}_{i_n(x)}^M & \text{if } \#(I_n^\delta(x)) = 0 \end{cases} \tag{4.15}$$

where

$$I_n^\delta(x) = \{i \leq n \ : \ X_i \in B_\delta(x)\}$$

and $i_n(x)$ is the index $i \leq n$ that minimizes $|X_i - x|$. Finally, define

$$h_n(x) = \int_{B_\delta(x)} g_n(y) w_x(y) dy \tag{4.16}$$

where $w_x(y)$ is a "weighting function" with support on $B_\delta(x)$ that integrates to one.

Remark 6 *Combining (4.15) and (4.16) and switching the order of the sum and integral we obtain*

$$h_n(x) = \sum_{i=1}^n \left(\int_{B_\delta(x_i) \cap B_\delta(x)} \frac{w_x(y)}{\#(I_n^\delta(y))} dy \right) Y_i,$$

which has the form (1.2).

Theorem 4 *If f is Lipschitz continuous on B with constant κ, and $\delta < \frac{\epsilon}{4\kappa+2}$ then $\{h_n\} \in \mathcal{C}_\epsilon$.*

Proof: First we show that if x is a cluster point of $\{X_n\}$ and $z \in B_\delta(x)$ then

$$\limsup_n |g_n(z) - f(x)| < \frac{\epsilon}{2} \quad a.s. \tag{4.17}$$

We write

$$|g_n(z) - f(x)| = \frac{1}{\#(I_n^\delta(z))} \left| \sum_{i \in I_n^\delta(z)} (\tilde{Y}_i^M - \tilde{f}^M(X_i) + \tilde{f}^M(X_i) - f(x)) \right|$$

$$\leq \frac{1}{\#(I_n^\delta(z))} \left| \sum_{i \in I_n^\delta(z)} (\tilde{Y}_i^M - \tilde{f}^M(X_i)) \right| + \frac{1}{\#(I_n^\delta(z))} \sum_{i \in I_n^\delta(z)} |\tilde{f}^M(X_i) - f(x)| \tag{4.18}$$

where

$$\tilde{f}^M(X_i) = E(\tilde{Y}_i^M \mid X_i, t_n = T). \tag{4.19}$$

If a null sum is defined to be zero then for every n we can bound the second term in (4.18) by

$$\begin{aligned}\frac{1}{\#(I_n^\delta(z))}\sum_{i\in I_n^\delta(z)}|\tilde{f}^M(X_i)-f(x)| &\leq \frac{1}{\#(I_n^\delta(z))}\sum_{i\in I_n^\delta(z)}|\tilde{f}^M(X_i)-f(X_i)| \\ &\quad +\frac{1}{\#(I_n^\delta(z))}\sum_{i\in I_n^\delta(z)}|f(X_i)-f(x)| \qquad (4.20)\end{aligned}$$

Using (4.14) and (4.19), the first term in (4.20) is less than $\delta/2$, and since f is Lipschitz continuous with constant κ the second term in (4.20) is less than $\kappa\delta$. Thus,

$$\limsup_n \frac{1}{\#(I_n^\delta(z))}\sum_{i\in I_n^\delta(z)}|\tilde{f}^M(X_i)-f(x)|<\frac{\epsilon}{2}. \qquad (4.21)$$

Let $\{m_1, m_2, \ldots\}$ be the indices where $X_i \in B_\delta(z)$, i.e.,

$$m_n = \min\{k \ : \ \#(I_k^\delta(z)) = n\}.$$

Since $x \in B_\delta(z)$ is a cluster point, we are assured that $\#(I_n^\delta(z)) \to \infty$. Let

$$\Psi_n = \frac{1}{n}\sum_{i=1}^{n}(\tilde{Y}_{m_i}^M - \tilde{f}^M(X_{m_i})).$$

Ψ_n is a sum of independent zero mean random variables, and using (4.14) and Durrett (1996), chapter 1, theorem 8.3, we have

$$\Psi_n \to 0 \quad a.s.$$

If $n \in [m_k, m_{k+1})$ then

$$\frac{1}{\#(I_n^\delta(z))}\sum_{i\in I_n^\delta(z)}(\tilde{Y}_i^M - \tilde{f}^M(X_i)) = \Psi_k$$

so

$$\frac{1}{\#(I_n^\delta(z))}\left|\sum_{i\in I_n^\delta(z)}(\tilde{Y}_i^M - \tilde{f}^M(X_i))\right| \to 0 \quad a.s., \qquad (4.22)$$

Equation (4.17) follows from (4.21) and (4.22).

We now show that $\{h_n\}$ satisfies (3.2). We begin by showing that if x is a cluster point of $\{X_n\}$ then

$$\limsup_n |h_n(x) - f(x)| < \frac{\epsilon}{2}. \qquad (4.23)$$

From (4.16) we have

$$\begin{aligned}\limsup_n |h_n(x) - f(x)| &= \limsup_n \left|\int_{B_\delta(x)}(g_n(y) - f(x))w_x(y)dy\right| \\ &\leq \limsup_n \int_{B_\delta(x)}|g_n(y) - f(x)|w_x(y)dy.\end{aligned}$$

Since $g_n(y) < M$ we can use the limsup version of Fatou's Lemma along with (4.17) to write

$$\limsup_n \int_{B_\delta(x)}|g_n(y) - f(x)|w_x(y)dy \ \leq \ \int_{B_\delta(x)}\limsup_n |g_n(y) - f(x)|w_x(y)dy \leq \epsilon/2.$$

Next, we show that there exists $\alpha > 0$ such that for any cluster point x of $\{X_n\}$,

$$\limsup_n \sup_{y \in B_\alpha(x)} |h_n(x) - h_n(y)| < \frac{\epsilon}{2}. \tag{4.24}$$

Without loss of generality, $\alpha < \delta$. Write

$$\sup_{y \in B_\alpha(x)} |h_n(x) - h_n(y)| \leq \sup_{y \in B_\alpha(x)} \left| \int_{D_{xy}} g_n(z) w_y(z) dz \right|$$

where D_{xy} is the symmetric difference of $B_\delta(x)$ and $B_\delta(y)$, i.e.,

$$D_{xy} = (B_\delta(x) - B_\delta(y)) \bigcup (B_\delta(y) - B_\delta(x)).$$

Thus,

$$\begin{aligned} \sup_{y \in B_\alpha(x)} |h_n(x) - h_n(y)| &\leq \sup_{y \in B_\alpha(x)} \int_{D_{xy}} |g_n(z)| w_y(z) dz \\ &\leq \int_{D_x^\alpha} |g_n(z)| w_y(z) dz \end{aligned}$$

where

$$D_x^\alpha = \bigcup_{y \in B_\alpha(x)} D_{xy}$$

Since $|g_n(\cdot)| < M$ we have

$$\limsup_n \int_{D_x^\alpha} |g_n(z)| w_y(z) dz \leq M \int_{D_x^\alpha} dz.$$

Since D_x^α is the region between the balls $B_{\delta - \alpha}(x)$ and $B_{\delta + \alpha}(x)$, its volume can be made arbitrarily small, and the required α does not depend on x. Property (3.2) follows from (4.23) and (4.24), and the theorem is proved.
□

5 Discussion

In the previous sections we introduced the "natural rootfinding procedure", established a sufficient condition for its convergence (the "cluster point property") and showed that the condition is satisfied by certain estimators. There remain two significant open problems. The estimator (4.4) that was proven to have the cluster point property is not applicable in general since it involves a likelihood ratio estimate. The nonparametric estimator (4.16) is widely applicable, but is inconvenient to use and is not in the class $\mathcal{C}$. It is conjectured that parametric estimators such as least squares regressions and Kriging estimators enjoy the cluster point property, but a proof is lacking. The second open problem is determining the rate of convergence of the natural rootfinding procedure.

Competing with the Robbins-Monro algorithm is a serious challenge for any new procedure for stochastic rootfinding. If asymptotic convergence rate is the benchmark, the best one can hope for is to be comparable with Robbins-Monro, so any claim to superiority must be based on other considerations. Robbins-Monro has two weaknesses: occasional divergence, and potentially bad short term behavior (even when it does converge). If one is certain that a root exists in a given region one can virtually ensure convergence by simply sticking X_n back in the region if it attempts

to escape. Andradottir (1995) offers a more sophisticated approach for improving the chances of convergence which appears to enhance the short term behavior in some cases as well. The averaging approach of Polyak and Juditsky (1992) improves both the short term behavior and the asymptotic convergence rate.

The natural rootfinding procedure cannot be easily dismissed, though. The conditions for convergence in theorem 2 are somewhat less restrictive than the analogous conditions for Robbins-Monro (e.g., theorem 1 in Andradottir (1995)). For example, if x^* is a saddle point of $F(x) = \nabla f(x)$, then Robbins-Monro may not find x^*. (Saddle points do not satisfy condition 3 in theorem 1, Andradottir (1995).) Also, if $f(x)$ is too "steep" near its root (e.g., $f(x) = x^3$) it is well known that Robbins-Monro can diverge unless it is carefully controlled. Furthermore, for Robbins-Monro to work one must know the "orientation" of $F(x)$, e.g., in one dimension it is necessary to know whether $f(x)$ is increasing or decreasing near the root. The natural rootfinding procedure does not share any of these problems.

Intuitively, one expects the short term behavior of the natural rootfinding procedure based on $\{H_n\}$ to be very competitive when $\{H_n\}$ approximates F well. One is free to use any available knowledge of the qualitative properties of F when choosing $\{H_n\}$. Of course, it is advisable to choose an estimator with the cluster point property, which is why it is important to establish this property for as many estimators as possible.

We now present some experimental results comparing the natural rootfinding procedure with Robbins-Monro. Considering the vast selection of estimators to choose from for the natural rootfinding procedure and the various enhanced and modified versions of Robbins-Monro, it is a difficult comparison to make. Our tactic is to use simple generic (perhaps naive) versions of both algorithms and caution the reader not to read too much into the numbers we obtain. For the natural rootfinding procedure we use a simple zero-mean Kriging estimator (equation (4.1)) with covariance function $\rho(x, y) = e^{-\|x-y\|}$. For Robbins-Monro we use a scaling factor of 1 (i.e., $X_{n+1} = X_n - \frac{1}{n}Y_n$), and place X_{n+1} on the boundary of the search region if it attempts to escape.

For our first example we compare the methods when $f(x) = cx$. We set $Y_n = cX_n + \sigma\chi_n$, where the χ_n's are independent standard normal random variates (i.e., we are "simulating a simulation"). The search region is the interval $[-10, 10]$. Figure 1 shows the standard errors of each method after 50, 100 and 200 simulations for various valued of c and σ, based on 25 replications of each experiment.

Our second example is an optimization problem based on the M/M/1 queue. We want to minimize

$$f(\lambda, \mu) = W(\lambda, \mu) + \mu + (\lambda - 1)^2,$$

where λ is the arrival rate, μ is the service rate, and $W(\lambda, \mu)$ is the stationary mean sojourn time. We are therefore looking for a root of the system of equations

$$F(\lambda, \mu) = \nabla f(\lambda, \mu) = \begin{pmatrix} \frac{\partial}{\partial\lambda} W(\lambda, \mu) + 2(\lambda - 1) \\ \\ \frac{\partial}{\partial\mu} W(\lambda, \mu) + 1 \end{pmatrix}.$$

Of course, it is well known that $W(\lambda, \mu) = (\mu - \lambda)^{-1}$ so the optimal solution $(\lambda^*, \mu^*) = (0.5, 1.5)$ is easily calculated. However, we treat $W(\lambda, \mu)$ and its derivatives as unknown quantities that must be evaluated by simulation. We use a regenerative simulation (busy cycles) and likelihood ratios to estimate $\frac{\partial}{\partial\lambda} W(\lambda, \mu)$ and $\frac{\partial}{\partial\mu} W(\lambda, \mu)$. The search region is the triangle $0 \leq \mu \leq 5$ and $0 \leq \lambda \leq .9\mu$. Figure 2 shows the standard errors of each method after 50, 100 and 200 simulations

for the estimates of λ^* and μ^* for varying numbers of busy cycles per simulation, based on 25 replications of each experiment.

REFERENCES

1. Andradottir, S., (1995), A Stochastic Approximation Algorithm with Varying Bounds, *Operations Research*, Vol. 43, 1037-1048.

2. Asaithambi, N.S., (1995), Numerical Analysis, Saunders College Publishing.

3. Durrett, R., (1996), Probability: Theory and Examples (second edition), Duxbury Press.

4. Iglehart, D.L., (1978), The Regenerative Method for Simulation Analysis, in *Current Trends in Programming Methodology - Software Modeling* (K.M. Chandy and R.T. Yeh, editors), Prentice Hall.

5. Koehler, J.R., Puhalskii, A.A. and Simon, B., (1997), Estimating Functions Evaluated by Simulation: a Bayesian/Analytic Approach, *Annals of Applied Probability*, to appear.

6. Kushner, H.J. and Yin, G.G., (1997), Stochastic Approximation Algorithms and Applications, Springer-Verlag.

7. Pflug, G.C., (1996), Optimization of Stochastic Models, Kluwer Academic Publishers.

8. Polyak, B. and Juditsky, A., (1992), Acceleration of Stochastic Approximation by Averaging, *SIAM Journal on Optimization*, Vol. 30, 838-855.

9. Reiman, M.I. and Weiss, A., (1989), Sensitivity Analysis for Simulations via Likelihood Ratios, *Operations Research*, Vol. 37, 830-844.

10. Ripley, B.D., (1981), Spatial Statistics, John Wiley and Sons.

11. Robbins, H. and Monro, S., (1951), A Stochastic Approximation Method, *Annals of Mathematical Statistics*, Vol. 22, 400-407.

12. Stein, M.L., (1990), Uniform Asymptotic Optimality of Linear Prediction of a Random Field Using an Incorrect Second-Order Structure, *Annals of Statistics*, Vol. 18, 850-872.

13. Stone, C.J., (1977), Consistent Nonparametric Regression, *Annals of Statistics*, Vol. 5, 595-620.

14. Yakowitz, S.J. and Szidarovszky, F., (1985), A Comparison of Kriging with Nonparametric Regression Methods, *Journal of Multivariate Analysis*, Vol. 16, 21-53.

Robbins-Monro									
	$\sigma = .1$			$\sigma = 1$			$\sigma = 10$		
slope	$se(50)$	$se(100)$	$se(200)$	$se(50)$	$se(100)$	$se(200)$	$se(50)$	$se(100)$	$se(200)$
0.5	0.0290	0.0211	0.0150	0.326	0.258	0.192	3.171	2.290	1.653
1.0	0.0109	0.0099	0.0086	0.142	0.0929	0.0764	1.377	0.964	0.622
2.0	0.0080	0.0052	0.0042	0.0775	0.0481	0.0343	0.923	0.484	0.308
5.0	0.0039	0.0029	0.0031	0.0486	0.0335	0.0230	0.514	0.261	0.276
10.0	0.0035	0.0020	0.0013	0.0331	0.0249	0.0163	0.282	0.277	0.192
natural rootfinding procedure									
	$\sigma = .1$			$\sigma = 1$			$\sigma = 10$		
slope	$se(50)$	$se(100)$	$se(200)$	$se(50)$	$se(100)$	$se(200)$	$se(50)$	$se(100)$	$se(200)$
0.5	0.0407	0.0277	0.0195	0.315	0.263	0.242	4.571	3.960	2.722
1.0	0.0245	0.0124	0.0072	0.203	0.119	0.118	2.394	2.328	1.166
2.0	0.0076	0.0030	0.0026	0.105	0.0678	0.0400	1.041	0.602	0.387
5.0	0.0038	0.0015	0.0017	0.0197	0.0318	0.0207	0.305	0.169	0.148
10.0	0.0015	0.0010	0.0007	0.0175	0.0084	0.0076	0.191	0.0990	0.0631

Figure 1: rootfinding for linear functions $f(x) = cx$

Robbins-Monro						
cycles/simulation	$se_\lambda(50)$	$se_\mu(50)$	$se_\lambda(100)$	$se_\mu(100)$	$se_\lambda(200)$	$se_\mu(200)$
500	0.181	0.528	0.141	0.339	0.107	0.204
1000	0.150	0.442	0.121	0.281	0.0844	0.170
10000	0.0265	0.0407	0.0147	0.0246	0.0096	0.0128
natural rootfinding procedure						
cycles/simulation	$se_\lambda(50)$	$se_\mu(50)$	$se_\lambda(100)$	$se_\mu(100)$	$se_\lambda(200)$	$se_\mu(200)$
500	0.148	0.0498	0.143	0.0441	0.141	0.0416
1000	0.0885	0.0460	0.0768	0.0273	0.0748	0.0284
10000	0.0222	0.0221	0.0134	0.0099	0.0113	0.0054

Figure 2: optimizing an M/M/1 queue