POLSKA AKADEMIA NAUK, INSTYTUT MATEMATYCZNY

DISSERTATIONES MATHEMATICAE

(ROZPRAWY MATEMATYCZNE)

CXLII

RYSZARD ZIELIŃSKI

Global stochastic approximation

WARSZAWA 1977

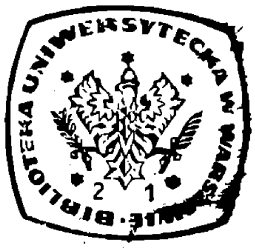
PAŃSTWOWE WYDAWNICTWO NAUKOWE

The paper

R. ZIELIŃSKI

Global stochactic approximation

has been written at the Mathematical Institute of the Polish Academy of Sciences, Problem 06.1.1/02.2.



PRINTED IN POLAND

WROCŁAWSKA DRUKARNIA NAUKOWA

CONTENTS

1. Intuitive background. Statement of the problem

A large number of practical questions lead to the following problem:

(P1) *Given a set* $\mathfrak{X}$ *and a real-valued function* F *on it, find a point in* $\mathfrak{X}$ *at which the function* F *achieves its maximum,*

which should be solved under some more or less awkward restrictions concerning the set $\mathfrak{X}$, the function F and, in consequence, methods which can be used.

The *set* $\mathfrak{X}$ should be considered as an abstract space. Usually there are no practical objections to extending it to a measurable space $(\mathfrak{X}, \mathfrak{B}, \mu)$ but there is no natural way of endowing it with a metric. Consider two simple examples. In the first we deal with a chemical process and the problem consists in finding the temperature and velocity for conducting the process at which it yields the maximal efficiency (it does not matter for us what this means). Although the couple (t, v) = (temperature, velocity) can be considered as a point in R^2, there is no real sense in measuring a distance between two such points. On the other hand, there are no objections to recognizing e.g. the interval $(20 < t < 40,\ 100 < v < 200)$ as twice as large as the interval $(20 < t < 30,\ 100 < v < 200)$, and thus the extension of the set of feasible values of parameters to the Lebesgue measurable space is quite natural. In the second example a number of machine tools and a number of jobs to be done are given and the problem consists in finding an assignment of each job to a machine such that all jobs will be finished as soon as possible (e.g. so-called optimal sequencing problem). Now $\mathfrak{X}$ is a finite set of all feasible assignments and it is quite natural to consider $(\mathfrak{X}, \mathfrak{B}, \mu)$, $\mathfrak{B}$ being the family of all subsets of $\mathfrak{X}$ and μ being the counting measure (i.e., $\mu(S)$ for $S \in \mathfrak{B}$ is equal to the number of elements in the set S).

The difficulty concerning the *function* F is that in many practical situation it is not known except that its values at every point $x \in \mathfrak{X}$ can be "observed". Usually the observation consists in performing some real experiments so that the observed value of the function is subjected to a random error of measurements. In such circumstances a more adequate statement of the problem might be:

(P2) *Given a measurable space* $(\mathfrak{X}, \mathfrak{B}, \mu)$ *and a family* $\{Y_x, x \in \mathfrak{X}\}$ *of real-valued random variables, find a point in* $\mathfrak{X}$ *at which the function* $F(x) = EY_x$ *attains its maximum.*

Later on we shall give a more precise statement of the problem but for the time being we content ourselves with the above one.

The *methods* used for the solution of our problem have of course to avail themselves only of the above information on $\mathfrak{X}$ and F. It seems that only methods of the following structure are acceptable: choose a point $x_1 \in \mathfrak{X}$ as the first approximation to the solution of the problem; suppose that the points $x_1, x_2, \ldots, x_n$ have been constructed and the values of the random variables $Y_{x_1}, Y_{x_2}, \ldots, Y_{x_n}$ recorded; construct the $(n+1)$-th approximation x_{n+1} to the solution, using no other information on F than that contained in $(x_1, Y_{x_1}, x_2, Y_{x_2}, \ldots, x_n, Y_{x_n})$. If x_{n+1} essentially depends on some of $Y_{x_1}, Y_{x_2}, \ldots, Y_{x_n}$ (or on all of them), we obtain a discrete-time $\mathfrak{X}$-valued random process $(x_n, n = 1, 2, \ldots)$, and the problem would be solved if we knew how to construct x_n so as to obtain a process which would converge in some sense to the desired points in $\mathfrak{X}$.

There are two ramifications in the *short history* of the problem. One of them began with the well-known paper by J. Kiefer and J. Wolfowitz [10], in which the stochastic approximation method invented by H. Robbins and S. Monro [15] had been adapted to the problem of seeking the maximum. Originally $\mathfrak{X} = R^1$, Y_x were random variables with uniformly bounded variances and F was a regular function. The process x_n converges in the sense that $E(x_n - \vartheta)^2 \to 0$, ϑ being a unique point of the (local) maximum of F. Later on convergence with probability one was proved and all results were extended to the more general case $\mathfrak{X} = R^m$, m being a finite integer. A number of investigations deal with the problem of limit distribution of x_n (see e.g. the review in Wasan [18]). with the speed of almost sure convergence (e.g. Heyde [7], Major [11]) and with the processes with improved speed of convergence in mean (e.g. Fabian [4], Zieliński [21]). An idea of extending stochastic approximation methods to the problems of seeking global extrema is presented in Waisbord and Youdin [16], [17]. An up-to-date review of the theory is given e.g. in Fabian [4] and Wasan [18].

The second branch in the development of our problem is connected with so-called randomized methods (or Monte Carlo methods) of seeking maxima. An early paper by Brooks [1] initiated an experimental approach to the problem and this developed into widespread random search methods of solution of some kind of engineering problems (e.g. Rastrigin [14]). A more sophisticated and strictly mathematical approach is given e.g. in Driml and Hanš [3], Kharlamov [9], Zieliński [20] or in more recent papers by Karmanov [9] and Perekrest [13]. In the above approach $\mathfrak{X}$ is usually an abstract set, the values of F are observed without any error (i.e., $\operatorname{Var} Y_x = 0$ in our terminology) and the points $x_1, x_2, \ldots$ are assumed to be random elements. A distinguishing feature

of these methods is that the sequence of random elements x_n converges in a sense to the element (or to the set of elements) at which F achieves its global maximum in $\mathfrak{X}$, i.e., to such $\bar{x}$ that $F(\bar{x}) \geqslant F(x)$ for all $x \in \mathfrak{X}$.

Our aim is to elaborate *global stochastic approximation processes* in which the above two ramifications would be combined. We shall consider problem (P2), $\mathfrak{X}$ being an abstract space as in random search problems and $\{Y_x, x \in \mathfrak{X}\}$ being a family of non-degenerate random variables as in stochastic approximation. The function F is only assumed to be $\mathfrak{B}$-measurable and essentially bounded. It can hardly be expected that very strong results will be obtained under such weak assumptions; on the other hand we do not want to pose over-restrictive conditions which might make the results practically useless.

The results are: 1° a general theorem on convergence in distribution of global stochastic approximation processes (Theorem 1); 2° if Y_x are unbounded random variables, the limit distribution is in a sense concentrated "near" the higher values of F (Theorems 2 and 3); 3° if Y_x are essentially bounded random variables, the process converges to the solution (to the global maximum) with probability one (Theorem 4). Besides, we shall give (Theorem 5) a pure Monte Carlo construction of the sequence of independent and identically distributed random variables whose distribution is concentrated (in the sense given in Theorem 3) at higher values of the function F.

2. General structure of global stochastic approximation processes

Let $(\mathfrak{X}, \mathfrak{B}, \mu)$ be a measurable space and $(\Omega, \mathfrak{A}, P)$ a probability space. Let $\{Y_x, x \in \mathfrak{X}\}$ be a family of real-valued random variables (r.v.'s) such that for every $x \in \mathfrak{X}$ the expectation EY_x exists and

$$P\{Y_x \leqslant y\} = G(y - EY_x),$$

G being a continuous distribution function. The function EY_x will be denoted by $F(x)$ and we will assume throughout the paper that

1° F is not essentially constant;

2° $F_0 = \operatorname{ess\,sup} F(x)$ is finite;

3° $\mathfrak{X}_0 = \{x \in \mathfrak{X}: F(x) = F_0\}$ is a non-empty set, although $\mu(\mathfrak{X}_0) = 0$ is allowed.

The set $\mathfrak{X}_0$ will be referred to as the optimal set. The problem consists in finding a point from $\mathfrak{X}_0$.

Consider two sequences $(\xi_n, n = 1, 2, \ldots)$ and $(X_n, n = 1, 2, \ldots)$ of $\mathfrak{X}$-valued random elements. To be in contact with applications we shall

refer to ξ_n as a random point at which the n-th experiment consisting in the observation of the value of the function F is to be performed and to Y_x as the outcome of the experiment at the point x. The r.v.'s Y_x are sometimes called *error random variables* and the distribution of ξ_n is called the *experimental design*. The r.v. X_n will be the n-th approximation to the solution.

Let $\mathfrak{A}_n$ be the sub-σ-field of $\mathfrak{A}$ generated by $(X_1, \xi_1, Y_{\xi_1}, X_2, \ldots, X_{n-1}, \xi_{n-1}, Y_{\xi_{n-1}}, X_n)$ and let $\mathfrak{A}_n^{\xi}$ be the sub-σ-field generated by $(X_1, \xi_1, Y_{\xi_1}, X_2, \ldots, X_{n-1}, \xi_{n-1}, Y_{\xi_{n-1}}, X_n, \xi_n)$. We will consider our problem under the following notation and assumptions:

(A1) $P_1(S) = P\{X_1 \in S\}$, $S \in \mathfrak{B}$, is an initial distribution of the process $(X_n, n = 1, 2, \ldots)$. It will be shown that the investigated properties of the process do not depend on P_1, and so it can be chosen optionally.

(A2) The conditional expectation $E[\varphi(\xi_n)|\mathfrak{A}_n]$, φ being a $\mathfrak{B}$-measurable function (we write $P\{\xi_n \in S|\mathfrak{A}_n\}$ instead of $E[\varphi(\xi_n)|\mathfrak{A}_n]$ if φ is the indicator function of S), is evidently an $\mathfrak{A}_n$-measurable function, and so it depends on $\omega \in \Omega$ only through $(X_1, \xi_1, Y_{\xi_1}, X_2, \ldots, X_{n-1}, \xi_{n-1}, Y_{\xi_{n-1}}, X_n)$. We shall assume even more, namely that it depends on ω only through X_n. It is assumed that

$$P\{\xi_n \in S \,|\, X_n = x\} = Q_x(S),$$

$Q_x(S)$ being a function on $\mathfrak{X} \times \mathfrak{B}$ which for every $x \in \mathfrak{X}$ is a probability measure on $\mathfrak{B}$ and for every $S \in \mathfrak{B}$ is a measurable function on $\mathfrak{X}$.

(A3) The conditional expectations $E[\psi_1(Y_{X_n})|\mathfrak{A}_n]$ and $E[\psi_2(Y_{X_n}, Y_{\xi_n})|\mathfrak{A}_n^{\xi}]$, ψ_1 and ψ_2 being Borel functions, are assumed to depend on $\omega \in \Omega$ only through X_n and (X_n, ξ_n), respectively; and

(a) for every $x \in \mathfrak{X}$ the conditional expectation $E[Y_{X_n}|\mathfrak{A}_n]$ considered as a function on $\mathfrak{X}$ coincides with EY_x:

$$E[Y_{X_n} \,|\, X_n = x] = F(x).$$

Consequently, the left-hand term will be denoted shortly by EY_x;

(b) for every $x \in \mathfrak{X}$ the conditional probability $P\{Y_{X_n} \leqslant y \,|\, \mathfrak{A}_n\}$ considered as a function on $\mathfrak{X}$ coincides with $P\{Y_x \leqslant y\}$:

$$P\{Y_{X_n} \leqslant y \,|\, X_n = x\} = G\big(y - F(x)\big).$$

Consequently, the left-hand term will be denoted shortly by $P\{Y_x \leqslant y\}$;

(c) the r.v.'s Y_{X_n} and Y_{ξ_n} are conditionally (given X_n, ξ_n) independent, i.e., for every $x, t \in \mathfrak{X}$ and every $y, y' \in R^1$ we have

$$P\{Y_{\xi_n} \leqslant y,\, Y_{X_n} \leqslant y' \,|\, \xi_n = t,\, X_n = x\} = P\{Y_{\xi_n} \leqslant y \,|\, \xi_n = t\} P\{Y_{X_n} \leqslant y' \,|\, X_n = x\}.$$

We shall use the short notation:

$$P\{Y_t \leqslant y,\ Y_x \leqslant y'\} = P\{Y_t \leqslant y\} \cdot P\{Y_x \leqslant y'\};$$

(d) the function

$$r(x, t) = P\{Y_{\xi_n} > Y_{X_n} |\ \xi_n = t,\ X_n = x\},$$

or shortly $r(x, t) = P\{Y_t > Y_x\}$, is a function which does not depend on n and is measurable with respect to the σ-field generated by $\mathfrak{B} \times \mathfrak{B}$. As the function $r(x, t)$ we choose a variant of the conditional probability $P\{Y_t > Y_x\}$ which is a non-increasing function with respect to $F(x)$ for given t and a non-decreasing function with respect to $F(t)$ for given x, i.e., for every t

$$r(x_1, t) \leqslant r(x_2, t) \quad \text{whenever } F(x_1) > F(x_2)$$

and for every x

$$r(x, t_1) \geqslant r(x, t_2) \quad \text{whenever } F(t_1) > F(t_2).$$

(A4) The sequence $(X_n, n = 1, 2, \ldots)$ is defined recursively

$$X_{n+1} = \begin{cases} \xi_n & \text{if } Y_{\xi_n} > Y_{X_n}, \\ X_n & \text{otherwise.} \end{cases} \tag{1}$$

Later we shall consider sequences X_n defined in a more sophisticated way and the above definition will become a special case.

The process $(X_n, n = 1, 2, \ldots)$ will be called the *global stochastic approximation* (g.s.a.) *process*. The practical realization of this process may be described in the following way:

1° choose a point $X_1 \in \mathfrak{X}$ according to the distribution P_1 and observe the r.v. Y_{X_1} at that point ("perform an appropriate experiment at the point X_1 and record its outcome as Y_{X_1}");

2° suppose that the points $X_1, X_2, \ldots, X_n$ are already selected. Sample the r.v. ξ_n according to the distribution Q_{X_n} and observe the r.v. Y_{ξ_n}. Put $X_{n+1} = \xi_n$ if $Y_{\xi_n} > Y_{X_n}$ and $X_{n+1} = X_n$ otherwise.

We shall show that for some experimental designs $(X_n, n = 1, 2, \ldots)$ is a reasonable process of approximation of the global maximum of F, viz. that X_n approaches in a sense the set $\mathfrak{X}_0$.

Note that under the above assumptions $(X_n, n = 1, 2, \ldots)$ is a discrete-time $\mathfrak{X}$-valued Markov chain with the initial probability distribution P_1 and the stationary transition probability function

$$P(x, S) = R(x)\, 1_S(x) + \int_S r(x, t) Q_x(dt), \tag{2}$$

where $1_S(x)$ is the indicator function of the set S, $R(x)$ is the $\mathfrak{B}$-measurable function defined as

$$R(x) = \int [1 - r(x, t)] Q_x(dt) \tag{3}$$

and the integral sign without limits means, as usual, integration over the whole space. This assertion results from the following argumentation.

Suppose $X_n = x$ and $x \in S$, $S \in \mathfrak{B}$. The event $\{X_{n+1} \in S\}$ occurs iff either $Y_{\xi_n} \leqslant Y_{X_n}$ (ξ_n arbitrary) or $\xi_n \in S$ and $Y_{\xi_n} > Y_{X_n}$. Thus

$$\begin{aligned} P\{X_{n+1} \in S \mid X_n = x\} &= P\{Y_{\xi_n} \leqslant Y_{X_n}, \xi_n \in \mathfrak{X} \mid X_n = x\} + \\ &\quad + P\{Y_{\xi_n} > Y_{X_n}, \xi_n \in S \mid X_n = x\} \\ &= \int P\{Y_{\xi_n} \leqslant Y_{X_n} \mid \xi_n = t, X_n = x\} P\{\xi_n \in dt \mid X_n = x\} + \\ &\quad + \int_S P\{Y_{\xi_n} > Y_{X_n} \mid \xi_n = t, X_n = x\} P\{\xi_n \in dt \mid X_n = x\}, \end{aligned}$$

and now by (A2), (A3, d) and (3)

$$P\{X_{n+1} \in S \mid X_n = x\} = R(x) + \int_S r(x, t) Q_x(dt).$$

If $X_n = x$ and $x \notin S$, the first term on the right-hand side of the above equality disappears. The probabilities $P\{X_{n+1} \in S \mid X_n = x\}$ do not depend on n and we can denote them by $P(x, S)$. So we obtain formula (2).

3. The fundamental theorem on convergence in distribution

Let $P_1(x, S) = P(x, S)$ and

$$P_{n+1}(x, S) = \int P_n(x, dy) P(y, S).$$

We shall see that under rather general assumptions there exists a limit distribution of $P_n(x, S)$ as $n \to \infty$ which in a sense is concentrated at the optimal set $\mathfrak{X}_0$. We begin with a definition.

DEFINITION 1. A family of distributions $\{Q_x, x \in \mathfrak{X}\}$ on $(\mathfrak{X}, \mathfrak{B})$ *uniformly dominates a measure* μ on $(\mathfrak{X}, \mathfrak{B})$ if for every $A \in \mathfrak{B}$ there exists a positive constant α_A such that $Q_x(A) \geqslant \alpha_A \mu(A)$ for all $x \in \mathfrak{X}$.

Note that if $Q_x \equiv Q$, then the uniform dominance of the family $\{Q_x, x \in \mathfrak{X}\}$ is equivalent to the usual dominance of Q with respect to μ: $Q \gg \mu$, i.e., the absolute continuity of μ with respect to Q: $\mu \ll Q$.

THEOREM 1. *If the family of distributions* $\{Q_x, x \in \mathfrak{X}\}$ *uniformly dominates the measure* μ, *then the sequence of distributions* $(P_n(x, \cdot), n = 1, 2, \ldots)$

converges uniformly (with respect to $\mathfrak{B}$) to a limit distribution $\overline{P}(\cdot)$. The limit distribution $\overline{P}(\cdot)$ does not depend on the initial value of x.

Proof. The proof is based on the method given in Doob [2].

We begin with the statement of two facts (a) and (b) below; then the assertion of the theorem will be a simple consequence.

1. Let $m_S^{(n)} = \inf_{x \in \mathfrak{X}} P_n(x, S)$ and $M_S^{(n)} = \sup_{x \in \mathfrak{X}} P_n(x, S)$. Because of

$$m_S^{(n)} = \inf_{x \in \mathfrak{X}} \int P(x, dy) P_{n-1}(y, S) \geqslant \int P(x, dy) m_S^{(n-1)} = m_S^{(n-1)}$$

and similarly $M_S^{(n)} \leqslant M_S^{(n-1)}$, we have

(a) $$m_S^{(1)} \leqslant m_S^{(2)} \leqslant \ldots \leqslant M_S^{(2)} \leqslant M_S^{(1)}.$$

2. For given $x, y \in \mathfrak{X}$ consider the following function on $\mathfrak{B}$

$$\psi_{x,y}(S) = P(x, S) - P(y, S).$$

The function $\psi_{x,y}$ is a countably additive set function; thus there exists a set A^+ such that for every $S \in \mathfrak{B}$, $S \subset A^+$ we have $\psi_{x,y}(S) \geqslant 0$ and for every measurable subset S of $A^- = \mathfrak{X} - A^+$ we have $\psi_{x,y}(S) \leqslant 0$. Obviously $\psi_{x,y}(A^+) + \psi_{x,y}(A^-) = 0$. We shall show that

(b) there exists a number $\varepsilon \in (0, 1)$ such that $\psi_{x,y}(A^+) \leqslant 1 - \varepsilon$.

According to (A3, b) we have $P\{Y_x \leqslant y\} = G\big(y - F(x)\big)$, $F(x)$ being the expectation of Y_x and G a continuous function. It is obvious that there exist positive constants δ and δ_1 such that for every $x \in \mathfrak{X}$

$$P\{Y_x > F(x) + \delta\} \geqslant \delta_1 \quad \text{and} \quad P\{Y_x < F(x) - \delta\} \geqslant \delta_1.$$

Consider the set $C_\delta = \{x \in \mathfrak{X}\colon F(x) \geqslant F_0 - \delta\}$. From the definition of F_0 as $\operatorname{ess\,sup} F$ it follows that $\mu(C_\delta) > 0$. The distribution Q_x dominates μ; hence $Q_x(C_\delta) > 0$.

For $x \in \mathfrak{X} - C_\delta$ and $t \in C_\delta$ we have $F(x) < F(t)$. The intersection of the events $\{Y_t > F(t)\}$ and $\{Y_x < F(x)\}$ implies the event $\{Y_t > Y_x\}$ and by (A3, d)

$$\begin{aligned} r(x, t) &= P\{Y_t > Y_x\} \geqslant P\{Y_t > F(t),\ Y_x < F(x)\} \\ &= P\{Y_t > F(t)\} \cdot P\{Y_x < F(x)\}, \end{aligned}$$

so that $r(x, t) = \delta_1^2$. Similarly, if $x \in C_\delta$ and $t \in C_\delta$, then $F(x) - \delta \leqslant F(t)$. The intersection of the events $\{Y_t > F(t) + \delta\}$ and $\{Y_x < F(x) - \delta\}$ implies the event $\{Y_t > Y_x\}$; thus, as above, $r(x, t) \geqslant \delta_1^2$. So the last inequality holds for all $x \in \mathfrak{X}$ and $t \in C_\delta$.

By these results we have

$$P(x, A^-) = R(x)1_{A^-}(x) + \int_{A^-} r(x, t)Q_x(dt)$$

$$\geqslant \int_{A^-} r(x, t)Q_x(dt) \geqslant \int_{A^- \cap C_\delta} r(x, t)Q_x(dt) \geqslant \delta_1^2 Q_x(A^- \cap C_\delta).$$

Similarly we can obtain $P(y, A^+) \geqslant \delta_1^2 Q_y(A^+ \cap C_\delta)$. Now

$$\psi_{x,y}(A^+) = P(x, A^+) - P(y, A^+) = 1 - P(x, A^-) - P(y, A^+)$$
$$\leqslant 1 - \delta_1^2[Q_x(A^- \cap C_\delta) + Q_y(A^+ \cap C_\delta)].$$

By the uniform dominance of $\{Q_x, x \in \mathfrak{X}\}$

$$Q_x(A^- \cap C_\delta) + Q_y(A^+ \cap C_\delta) \geqslant a\mu(C_\delta),$$

where $a = \min(a_{A^- \cap C_\delta}, a_{A^+ \cap C_\delta})$, so that $\psi_{x,y}(A^+) \leqslant 1 - \varepsilon$ with $\varepsilon = \delta_1^2 a\mu(C_\delta) > 0$. On the other hand, $\psi_{x,y}(A^+)$ is positive, which implies $\varepsilon < 1$. The fact (b) is proved.

For the difference $M_S^{(n)} - m_S^{(n)}$ we have

$$M_S^{(n)} - m_S^{(n)} = \sup_{x,y} \int P_{n-1}(t, S)[P(x, dt) - P(y, dt)]$$

$$= \sup_{x,y} \int P_{n-1}(t, S)\psi_{x,y}(dt)$$

$$\leqslant \sup_{x,y} \Big(\int_{A^+} M_S^{(n-1)}\psi_{x,y}(dt) + \int_{A^-} m_S^{(n-1)}\psi_{x,y}(dt) \Big)$$

$$= \sup_{x,y} \psi_{x,y}(A^+)[M_S^{(n-1)} - m_S^{(n-1)}]$$

$$\leqslant (1 - \varepsilon)[M_S^{(n-1)} - m_S^{(n-1)}].$$

Thus we have

$$M_S^{(n)} - m_S^{(n)} \leqslant (1 - \varepsilon)^{n-1}.$$

It follows that the sequences $m_S^{(n)}$ and $M_S^{(n)}$, $n = 1, 2, \ldots$, have a common limit, say $\bar{P}(S)$, and

$$|P_n(x, S) - \bar{P}(S)| \leqslant M_S^{(n)} - m_S^{(n)} \leqslant (1 - \varepsilon)^{n-1}.$$

The function $\bar{P}$ is a non-negative function on $\mathfrak{B}$ such that $\bar{P}(\mathfrak{X}) = 1$. It is countably additive as the limit of a uniformly convergent sequence of countably additive functions. So $\bar{P}$ is a probability distribution on $(\mathfrak{X}, \mathfrak{B})$. ■

Note that, because of the equality $P_{n+1}(x, S) = \int P_n(x, dy)P(y, S)$ and the convergence just proved, we have $\bar{P}(S) = \int \bar{P}(dy)P(y, S)$, so that $\bar{P}$ is an invariant probability measure for the transition probability function $P(x, S)$.

4. Absolute continuity of the limit distribution

4.1. Introductory remarks. If we want the process $(X_n, n = 1, 2, \ldots)$ to be a reasonable process of global stochastic approximation, the limit distribution $\overline{P}$ should be in a sense concentrated "near" the optimal set $\mathfrak{X}_0$. The properties of $\overline{P}$ obviously depend on the distributions Q_x. From the point of view of applications it is rather difficult to impose any strong conditions on these distributions, so that eventually we shall confine ourselves to the case $Q_x \equiv Q$, Q being the uniform distribution on $(\mathfrak{X}, \mathfrak{B}, \mu)$, i.e., $Q(S) = \mu(S)/\mu(\mathfrak{X})$ provided $\mu(\mathfrak{X}) < +\infty$. Under this assumption $Q \ll \mu$, so that we start with considering absolute continuity of the measures P_n. It will be shown that under some conditions the limit distribution $\overline{P}$ is also absolutely continuous and properties of its density justify considering the process $(X_n, n = 1, 2, \ldots)$ as a process of an approximation of the global maximum of F on $\mathfrak{X}$, i.e., as a g.s.a. process. The results are formulated in Theorem 2, preceded by two auxiliary lemmas. Since according to Theorem 1 the limit distribution does not depend on the initial value of the process we shall assume P_1 to be an absolutely continuous distribution (in Lemmas 1 and 2) or even the uniform distribution (in Theorem 2).

4.2. General case. Define $P_n(S) = P\{X_n \in S\}$. Then $P_n(S) = \int P_{n-1}(dx)P(x, S) = \int P_1(dx)P_{n-1}(x, S)$. As a resultof Theorem 1 we have

$$|P_n(S) - \overline{P}(S)| \leqslant (1-\varepsilon)^{n-1}, \quad n = 1, 2, \ldots$$

for every $S \in \mathfrak{B}$. We shall prove some properties of the limit distribution $\overline{P}$, characterizing them by the properties of the density $\overline{p}$ of $\overline{P}$. Let p_n denote the density of P_n. All the densities are taken with respect to the measure μ.

DEFINITION 2. R.v.'s $\{Y_x, x \in \mathfrak{X}\}$ are said to be *unbounded* if $P\{Y_x > a\} > 0$ for μ-almost all $x \in \mathfrak{X}$ and for every real number a. In terms of the function G (cf. (A3, b)) this is equivalent to $G(a) < 1$ for each a.

LEMMA 1. *If $P_1 \ll \mu$ and $Q_x \ll \mu$ for all $x \in \mathfrak{X}$, then $P_n \ll \mu$, $n = 1, 2, \ldots$ The densities p_n satisfy the following equations*

$$p_{n+1}(s) = R(s)p_n(s) + \int r(x, s)q_x(s)p_n(x)\mu(dx), \tag{4}$$

$q_x(s)$ being the densities of Q_x, $x \in \mathfrak{X}$.

Proof. P_1 is assumed to be absolutely continuous. Suppose $P_n \ll \mu$. We have

$$\begin{aligned} P_{n+1}(S) &= \int P_n(dx)P(x, S) \\ &= \int_S R(x)p_n(x)\mu(dx) + \int_{t \in S} \int_{x \in \mathfrak{X}} r(x, t)p_n(x)q_x(t)\mu(dx)\mu(dt), \end{aligned}$$

so that $P_{n+1} \ll \mu$ and p_{n+1} is given by (4). ■

4.3. Uniform experimental design.

LEMMA 2. *Assume that* (i) $Q_x \equiv Q \ll \mu$ *and* $P_1 \ll \mu$; (ii) $Q \gg \mu$; (iii) *the densities* p_1 *and* q *of* P_1 *and* Q, *respectively, are* μ*-almost everywhere bounded*; (iv) *the r.v.'s* $\{Y_x, x \in \mathfrak{X}\}$ *are unbounded*; (v) $\inf_{x \in \mathfrak{X}} F(x) > -\infty$. *Then the limit distribution* $\bar{P}$ *is absolutely continuous and its density* $\bar{p}$ *is a solution of the integral equation*

$$\bar{p}(s) = \int \frac{q(s)r(x,s)}{1-R(s)} \bar{p}(x)\mu(dx) \tag{5}$$

satisfying $\int \bar{p}(s)\mu(ds) = 1$. *The limit density* $\bar{p}$ *is positive whenever* q *is positive.*

Proof. The limit distribution $\bar{P}$ exists by hypothesis (ii) and Theorem 1. First of all we will show that

$$\beta < r(x,t) < 1-\beta \quad \text{and} \quad \beta < R(x) < 1-\beta \tag{6}$$

for μ-almost all x, $t \in \mathfrak{X}$, β being a positive number less than $\frac{1}{2}$.

For any real a we have

$$r(x,t) = P\{Y_t > Y_x\} > P\{Y_t > a > Y_x\}.$$

By (A3, c) the right-hand term is equal to $P\{Y_t > a\} \cdot P\{Y_x < a\}$. By (A3, b) the distribution function G is continuous and by assumption (iv) the r.v.'s Y_x are unbounded, so that for μ-almost all x and t we have

$$\begin{aligned} P\{Y_t > a\} \cdot P\{Y_x < a\} &= \big[1 - G\big(a - F(t)\big)\big] G\big(a - F(x)\big) \\ &\geqslant [1 - G(a - \inf F)] G(a - F_0), \end{aligned}$$

and this is greater than a positive β. It is clear that taking a large enough we can get $\beta < \frac{1}{2}$, which will be assumed. Thus $r(x,t) > \beta > 0$ for μ-almost all x, $t \in \mathfrak{X}$ and, by symmetry, $1 - r(x,t) = 1 - P\{Y_t > Y_x\} = P\{Y_x > Y_t\} + P\{Y_x = Y_t\} \geqslant P\{Y_x > Y_t\} = r(t,x) > \beta > 0$. This gives the first part of (6); the second part follows from the definition of $R(x)$.

According to Lemma 1 we have

$$p_{n+1}(s) = R(s)p_n(s) + \int r(x,s)q(s)p_n(x)\mu(dx), \tag{7}$$

and now by (6)

$$\begin{aligned} p_{n+1}(s) &< (1-\beta)[p_n(s) + q(s)] \\ &< (1-\beta)^n p_1(s) + q(s) \sum_{j=1}^{n} (1-\beta)^j \\ &= (1-\beta)^n p_1(s) + q(s) \frac{1-\beta}{\beta} [1 - (1-\beta)^n], \end{aligned} \tag{8}$$

so that, by hypothesis (iii), $p_{n+1}(s)$ is bounded μ-almost everywhere. Letting $n\to\infty$ in (7), we obtain

$$\overline{p}(s) = R(s)\overline{p}(s) + \int r(x, s)q(s)\overline{p}(x)\mu(dx),$$

and this is equation (5).

By (8) we obtain

$$\overline{p}(s) = \lim_{n\to\infty} p_n(s) \leqslant \frac{1-\beta}{\beta} q(s)$$

and

$$\overline{P}(S) \leqslant \frac{1-\beta}{\beta} Q(S).$$

Absolute continuity of Q implies that of $\overline{P}$.

The equality $\int \overline{p}(s)\mu(ds) = 1$ follows from the fact that $\overline{P}$ is absolutely continuous and $\overline{P}(\mathfrak{X}) = 1$ (cf. Theorem 1).

To prove the last assertion of the lemma we use once again (6) and (7):

$$p_{n+1}(s) > \beta[p_n(s) + q(s)] > \beta^n p_1(s) + q(s)\frac{\beta}{1-\beta}(1-\beta^n);$$

hence $\overline{p}(s) = \lim_{n\to\infty} p_n(s) \geqslant \frac{\beta}{1-\beta} q(s)$. ■

THEOREM 2. *If P_1 and Q are uniform distributions on $(\mathfrak{X}, \mathfrak{B}, \mu)$, then there exist variants of densities p_n, $n = 1, 2, \ldots$, such that every p_n is constant on sets $\{x \in \mathfrak{X}: F(x) = \text{const}\}$ and is a non-decreasing function with respect to F, i.e., $p_n(x_1) \geqslant p_n(x_2)$ whenever $F(x_1) > F(x_2)$. If, furthermore, the r.v.'s $\{Y_x, x \in \mathfrak{X}\}$ are unbounded and $\inf_{x\in\mathfrak{X}} F(x) > -\infty$, then there exists a variant $\overline{p}(s)$ of the limit density such that $\overline{p}(s)$ is constant whenever $F(x) = \text{const}$ and is a non-decreasing function with respect to F.*

Proof. By hypothesis, the distributions Q and P_1 are uniform on $(\mathfrak{X}, \mathfrak{B}, \mu)$, so that their denisties are constant μ-almost everywhere and equal to $1/\mu(\mathfrak{X})$. Let q and p_1 be variants of these densities such that $q(x) = p_1(x) = 1/\mu(\mathfrak{X})$ for all $x \in \mathfrak{X}$. Define p_2 by formula (4)

$$p_2(s) = \frac{1}{\mu(\mathfrak{X})}\left[R(s) + \frac{1}{\mu(\mathfrak{X})}\int r(x, s)\mu(dx)\right].$$

Thus the density p_2 is uniquely defined. The functions $r(x, s)$ and $R(s)$ are constant on sets $\{s: F(s) = \text{const}\}$ and non-decreasing with respect to F, hence p_2 has the same properties.

By induction (cf. formula (4)) we obtain the assertion for all p_n, $n = 1, 2, \ldots$

By Theorem 1, the limit distribution $\bar{P}$ exists and by Lemma 2 its density is equal μ-almost everywhere to $\lim_{n\to\infty} p_n(s)$; defining $\bar{p}(s)$ as equal to this limit for all $s \in \mathfrak{X}$, we obtain the Theorem. ■

Theorem 2 reveals a feature of the limit distribution which justifies considering the process $(X_x, n = 1, 2, \ldots)$ as a process of an approximation to the global maximum of F. An obvious consequence of this theorem may be formulated as follows. Let f be a real number belonging to $F(\mathfrak{X})$ and $C_f = \{x \in \mathfrak{X}: F(x) \geqslant F_0 - f\}$. Let $A, B \in \mathfrak{B}$ be such sets that $A \subset C_f$, $B \subset \mathfrak{X} - C_f$ and $\mu(A) = \mu(B)$. Then $\bar{P}(A) \geqslant \bar{P}(B)$.

To see that the assumption that Q is the uniform distribution on $(\mathfrak{X}, \mathfrak{B}, \mu)$ is essential, consider the following simple numerical example. Let $\mathfrak{X} = \{1, 2, 3\}$, $F(x) = x$ and the r.v.'s $\{Y_x, x \in \mathfrak{X}\}$ be such that for $r(x, t)$ defined in (A3, d) we have $r(1, 2) = 1 - r(2, 1) = 0.7$; $r(1, 3) = 1 - r(3, 1) = 0.9$; $r(2, 3) = 1 - r(3, 2) = 0.7$ and obviously $r(1, 1) = r(2, 2) = r(3, 3) = 0.5$. Suppose that $Q(\{1\}) = 0.7$; $Q(\{2\}) = 0.2$ and $Q(\{3\}) = 0.1$. According to (3) $R(x) = 1 - \sum_t r(x, t)Q(t)$, so that $R(1) = 0.42$; $R(2) = 0.62$ and $R(3) = 0.82$. By formula (2) $P(x, y) = r(x, y)Q(y)$ if $x \neq y$ and $P(x, x) = R(x) + r(x, x)Q(x)$. Hence we obtain the following transition probability matrix $\{P(x, y)\}$:

$$\begin{bmatrix} 0.77 & 0.14 & 0.09 \\ 0.21 & 0.72 & 0.07 \\ 0.07 & 0.06 & 0.87 \end{bmatrix}$$

Now for the limit distribution $\bar{P}$ we have $\bar{P}(\{1\}) = \frac{161}{454}$, $\bar{P}(\{2\}) = \frac{118}{454}$ and $\bar{P}(\{3\}) = \frac{175}{454}$ instead of $\bar{P}(\{1\}) \leqslant \bar{P}(\{2\}) \leqslant \bar{P}(\{3\})$, as stated by Theorem 2. The assumption of Theorem 2, however, can be weakened; the conclusion remains true if the densities p_1 and q are constant whenever $F(x) = \text{const}$ and are non-decreasing functions of F; this property, however, does not seem to have a great practical value, while in many practical cases the construction of the uniform distribution does not present any difficulties.

4.4. Improvement by randomization. It is interesting to note that a limit distribution which is better concentrated at the optimal set $\mathfrak{X}_0$ (in the sense given below) can be obtained by a randomization of the process given by formula (1).

Let $(U_n, n = 1, 2, \ldots)$ be a sequence of independent r.v.'s distributed uniformly (with respect to the Lebesgue measure) on the interval $[0, 1]$. We shall assume that the r.v.'s U_n, ξ_n, and Y_{ξ_n} are conditionally (given X_{n-1}, $Y_{X_{n-1}}$) independent. Let w: $R^1 \to [0, 1-a]$ for some $a \in (0, 1]$ be a Lebesgue-measurable non-decreasing function. Modify the process

$(X_n, n = 1, 2, \ldots)$ as follows:

$$X_{n+1} = \begin{cases} X_n & \text{if } \{U_n \leqslant w(Y_{X_n})\} \text{ or} \\ & \{U_n > w(Y_{X_n}) \text{ and } Y_{\xi_n} \leqslant Y_{X_n}\}, \\ \xi_n & \text{otherwise.} \end{cases} \tag{9}$$

By arguments similar to those in Chapter 2 we conclude that the sequence $(X_n, n = 1, 2, \ldots)$ forms a Markov chain with transition probabilities

$$P_w(x, S) = \int P(x, S|\ Y_x = y)\, dP(Y_x \leqslant y), \tag{10}$$

where

(11)

$$P(x, S|\ Y_x = y) = \begin{cases} w(y) + [1 - w(y)]\left(R(x) + \int\limits_S r(x, t)Q(dt)\right) & \text{if } x \in S, \\ [1 - w(y)] \int\limits_S r(x, t)Q(dt) & \text{otherwise.} \end{cases}$$

Denote $E[w(Y_\xi)|\, \xi = x]$ shortly by $E_w(x)$. The transition probability function (10) can be written in the form

$$P_w(x, S) = \left(R(x) + [1 - R(x)]E_w(x)\right)1_S(x) + [1 - E_w(x)] \int\limits_S r(x, t)Q(dt).$$

Put

$$R_w(x) = R(x) + [1 - R(x)]E_w(x),$$

$$r_w(x, t) = [1 - E_w(x)]r(x, t).$$

This gives

$$P_w(x, S) = R_w(x)1_S(x) + \int\limits_S r_w(x, t)Q(dt),$$

which is similar to (2) considered previously, with R_w and r_w instead of R and r. Thus, under the assumptions of Theorems 1 and 2, the randomized process (9) converges in distribution, the limit distribution $\bar{P}_w$ does not depend on the initial value of the process, and we can choose such a variant $\bar{p}_w$ of its density which is constant whenever $F(x) = \text{const}$ and is a non-decreasing function of F. This results from the following arguments.

The crucial points in the proofs were those concerning the function $r(x, t)$. In the proof of Theorem 1 we established that $\psi_{x,y}(A^+) \leqslant 1 - \varepsilon$ for all $x, y \in \mathfrak{X}$, using the estimation $r(x, t) \geqslant \delta_1^2$ in a suitable set. Now $r_w(x, t) = [1 - E_w(x)]r(x, t)$; by the hypothesis w: $R^1 \to [0, 1 - a]$ we have $E_w(x) \leqslant 1 - a$, and an analogous estimation holds. In Lemma 2 we used the facts that $\beta < r(x, t) < 1 - \beta$ and $\beta < R(x) < 1 - \beta$ for some $\beta \in (0, \frac{1}{2})$; now by $0 \leqslant E_w(x) \leqslant 1 - a$ we have $\beta' < r_w(x, t) < 1 - \beta'$ and

$\beta' < R_w(x) < 1-\beta'$ with a suitable β'. To prove the monotonicity of p_n, and consequently that of the limit density $\bar{p}$, we used the monotonicity of $r(x, s)$ and $R(s)$ with respect to $F(s)$. Now $r_w(x, s) = [1-E_w(x)]r(x, s)$ as a function of $F(s)$ behaves like $r(x, s)$. To see that $R(s)$ is a non-decreasing function of $F(s)$ consider s_1 and s_2 such that $F(s_1) > F(s_2)$; then

$$R_w(s_1) - R_w(s_2)$$
$$= [R(s_1) - R(s_2)] \cdot [1 - E_w(s_1)] + [E_w(s_1) - E_w(s_2)] \cdot [1 - R(s_2)] \geqslant 0.$$

The density $\bar{p}_w$ of the limit distribution satisfies the equation

$$[1-E_w(s)]\bar{p}_w(s) = \int \frac{q(s)r(x, s)}{1-R(s)} [1-E_w(x)]\bar{p}_w(x)\mu(dx),$$

while in the former case (non-randomized, i.e., for $w \equiv 0$) we had

$$\bar{p}(s) = \int \frac{q(s)r(x, s)}{1-R(s)} \bar{p}(x)\mu(dx).$$

To establish a correspondence between $\bar{p}$ and $\bar{p}_w$ consider the recurrence formula (7):

$$p_{n+1}(s) = R(s)p_n(s) + \int r(x, s)q(s)p_n(x)\mu(dx).$$

This now takes on the following form:

$$p'_{n+1}(s) = \big[R(s) + \big(1-R(s)\big)E_w(s)\big]p'_n(s) + q(s)\int [1-E_w(x)]r(x, s)p'_n(x)\mu(dx).$$

An augmentation of the right-hand side by adding the term $\big(p'_{n+1}(s) - -p'_n(s)\big)E_w(s)$, which tends to zero as $n \to \infty$, does not affect the limit distribution $\bar{p}_w(s) = \lim\limits_{n\to\infty} p'_n(s)$, but then we have

$$[1-E_w(s)]p'_{n+1}(s)$$
$$= R(s)[1-E_w(s)]p'_n(s) + q(s)\int [1-E_w(x)]r(x, s)\,p'_n(x)\mu(dx),$$

which is the same as (7) with $[1-E_w(s)]p'_n(s)$ instead of $p_n(s)$. Let C be a constant such that $C\int [1-E_w(s)]p'_1(s)\mu(ds) = 1$; then $C\int [1-E_w(s)] \cdot p'_1(s)\mu(ds) = 1$ for all $n = 1, 2, \ldots$ and in consequence $C\int [1-E_w(s)] \cdot \bar{p}_w(s)\mu(ds) = 1$. Under the conditions of Theorems 1 and 2 the limit distribution does not depend on the initial distribution, and it follows that

$$\bar{p}(s) = C[1-E_w(s)]\bar{p}_w(s),$$

μ-almost everywhere. Because of $1-E_w(x) \geqslant a > 0$ the sets on which $\bar{p}(s) = 0$ and $\bar{p}_w(s) = 0$ might differ at most by a μ-null-set. On the set

where $\overline{p}(s)$ and $\overline{p}_w(s)$ are positive we have

$$\frac{\overline{p}(s)}{\overline{p}_w(s)} = C[1 - E_w(s)]. \tag{12}$$

The densities $\overline{p}(s)$ and $\overline{p}_w(s)$ are non-decreasing functions of F. If the function w is not constant μ-almost everywhere (remember that F is not essentially constant by the general hypothesis formulated in Chapter 2), then $E_w(s)$ is not essentially constant. Then $1 - E_w(s)$ is a non-increasing function of F and by (12) so is $\overline{p}(s)/\overline{p}_w(s)$. It follows that there exists a constant $f \in F(\mathfrak{X})$ such that $\overline{p}_w(s) > \overline{p}(s)$ if and only if $F(s) > f$. We shall formulate all these results as the following theorem concerning the randomized process (9).

THEOREM 3. *Assume that* (i) P_1 *and* Q *are uniform distribution on* $(\mathfrak{X}, \mathfrak{B}, \mu)$; (ii) *the error r.v.'s* $\{Y_x, x \in \mathfrak{X}\}$ *are unbounded*; (iii) $\inf F > -\infty$; (iv) $w\colon R^1 \to [0, 1-a]$, $a \in (0, 1)$, *is a non-decreasing function. Then there exists a limit distribution* $\overline{P}_w(S) = \lim_{n\to\infty} P\{X_n \in S\}$, X_n *being given by* (9). *The limit distribution* $\overline{P}_w$ *is absolutely continuous and there exists such a variant* $\overline{p}_w$ *of its density which is constant on sets* $\{x\colon F(x) = \text{const}\}$ *and non-decreasing with respect to* $F(x)$. *If the function* w *is not constant, then there exists a constant* $f \in F(\mathfrak{X})$ *such that* $\overline{p}_w(s) > \overline{p}(s)$ *iff* $F(s) > f$, $\overline{p}(s)$ *being a limit density in the process* (9) *with* $w \equiv 0$, *i.e., in the process* (1). ■

The last theorem explains the sense in which the randomized g.s.a. process $(X_n, n = 1, 2, \ldots)$ given by (9) is better than the original process (1). This can be pointed out as follows. There exists a constant f such that if $A \subset C_f = \{x \in \mathfrak{X}\colon F(x) \geqslant F_0 - f\}$, $B \subset \mathfrak{X} - C_f$ $(A, B \in \mathfrak{B})$ and $\mu(A) = \mu(B)$, then $\overline{P}_w(A) \geqslant \overline{P}(A) \geqslant \overline{P}(B) \geqslant \overline{P}_w(B)$.

Further on we shall consider only the latter case of the randomized process (9), as a more general one.

The following case is interesting. Let $\{Y_x, x \in \mathfrak{X}\}$ be random variables concentrated at $F(x)$: $P\{Y_x = F(x)\} = 1$. Then obviously $E_w(x) = w(F(x))$. Assume $r(x, t) \equiv 1$; then $R(x) \equiv 0$ and

$$P(x, S) = w(F(x))\, 1_S(x) + [1 - w(F(x))]\, Q(S).$$

This is the case where at every point $x \in \mathfrak{X}$ the value $F(x)$ of the function F can be observed without any error. Such processes have been considered by Kharlamov [9].

4.5. Problem of optimal experimental design. Let $\overline{P}_1$ and $\overline{P}_2$ be limit distributions in two different g.s.a. processes. Following Kharlamov [9] and Perekrest [19], we shall call the former process better than the latter if

$$\int F(x)\overline{P}_1(dx) > \int F(x)\overline{P}_2(dx). \tag{13}$$

Similarly we will use "worse", "not better" or "not worse" if, respectively, $<$, $\leqslant$ or $\geqslant$ holds. Intuitively, we can expect that a better process will give a better solution (in the limit) to our problem.

· The problem arises whether in a given class of g.s.a. processes there exists the best process, i.e., a process with a limit distribution, say $\overline{P}_0$, such that $\int F(x)\overline{P}_0(dx) \geqslant \int F(x)\overline{P}(dx)$, $\overline{P}$ being the limit distribution of any other process in the class under consideration. We shall discuss the problem for the class of randomized g.s.a. processes given by formula (9) with w: $R^1 \to [0, 1-a]$, $a \in (0, 1]$ and Q assumed to be the uniform distribution on $(\mathfrak{X}, \mathfrak{B}, \mu)$ — this is just the case of Theorem 3. We shall show that for every function w there exists a function v belonging to the above class and such that

$$\int F(x)\overline{P}_v(dx) > \int F(x)\overline{P}_w(dx) \tag{14}$$

provided the function F is not essentially constant on $\mathfrak{X}$; $\overline{P}_v$ and $\overline{P}_w$ are the limit distributions in processes with v and w, respectively. Hence it follows that there is no optimal element in the discusseed class of processes.

Suppose that w is not constant and put $a = \sup_{x \in \mathfrak{X}} w(F(x))$. Take a number $b \in (a, 1)$ and define

$$v = \frac{b}{a} w.$$

Then we have $E_v(x) = \frac{b}{a} E_w(x)$. By Theorem 3 the limit distributions $\overline{P}_v$ and $\overline{P}_w$ are absolutely continuous with respect to μ; denote their densities by $\overline{p}_v$ and $\overline{p}_w$, respectively. By arguments similar to those used for establishing formula (12) we have

$$c_w[1 - E_w(x)]\overline{p}_w(x) = c_v\left[1 - \frac{b}{a} E_w(x)\right]\overline{p}_v(x),$$

μ-almost-everywhere, c_w and c_v being appropriate constants. This yields

$$\frac{\overline{p}_w(x)}{\overline{p}_v(x)} = \frac{c_v}{c_w} \cdot \frac{1 - \frac{b}{a} E_w(x)}{1 - E_w(x)}. \tag{15}$$

The densities $\overline{p}_w$ and $\overline{p}_v$ are non-decreasing functions of F and $\int \overline{p}_w(x)\mu(dx) = \int \overline{p}_v(x)\mu(dx) = 1$. The function $E_w(x)$ is a non-decreasing function with respect to F and takes on values exclusively from $[0, a]$. But then the right-hand term of (15) forms a non-increasing (and not constant) function of F. It follows that there exists a constant $f \in F(\mathfrak{X})$ such that

$\overline{p}_v(x) > \overline{p}_w(x)$ if and only if $F(x) > f$. Now the obvious equation

$$\int\limits_{\{F(x)>f\}} [\overline{p}_v(x) - \overline{p}_w(x)]\,\mu(dx) = \int\limits_{\{F(x)\leqslant f\}} [\overline{p}_w(x) - \overline{p}_v(x)]\,\mu(dx)$$

gives

$$\begin{aligned}\int\limits_{\{F(x)>f\}} F(x)[\overline{p}_v(x) - \overline{p}_w(x)]\,\mu(dx) &> \int\limits_{\{F(x)>f\}} f\cdot[\overline{p}_v(x) - \overline{p}_w(x)]\,\mu(dx)\\ &= \int\limits_{\{F(x)\leqslant f\}} f\cdot[\overline{p}_w(x) - \overline{p}_v(x)]\,\mu(dx)\\ &\geqslant \int\limits_{\{F(x)\leqslant f\}} F(x)[\overline{p}_w(x) - \overline{p}_v(x)]\,\mu(dx),\end{aligned}$$

and this is another form of (14).

The non-existence of optimal experimental design in the discussed class is of course due to the fact that a (in the definition of w) belongs to the interval which is open from below, but this is essential in the proof of the convergence Theorem 1 and seems to be irremovable. Suppose, however, that we have proved the convergence of the randomized g.s.a. process with $w\colon R^1 \to [0, 1)$. Then by arguments similar to those above we would be able to show that the process with $v = (1-a)w + a$ for a number $a \in (0, 1)$ is better than that with w and the optimal process still does not exist. On the other hand, if we allowed the function w to take on the value 1 for a finite value of the argument, then, because of the unboundness of the error variables, the process could stop with positive probability at any point $x \in \mathfrak{X}$, which is a highly undesirable case. For these reasons we will not examine the above extensions of the class of randomized g.s.a. processes. Other classes of g.s.a. processes in the case of unbounded error r.v.'s will not be discussed in the present paper.

5. Almost sure convergence to global maximum

When r.v.'s $\{Y_x, x \in \mathfrak{X}\}$ are bounded the g.s.a. process converges with probability 1 to the global maximum of F. The result is formulated in Theorem 4 below, preceded by two definitions.

DEFINITION 3. R.v.'s $\{Y_x, x \in \mathfrak{X}\}$ are said to be *essentially bounded* from above μ-almost everywhere if $l = \operatorname*{esssup}_{(\mathfrak{X},\mathfrak{B},\mu)} l(x) < \infty$, where $l(x) = \operatorname*{esssup}_{(\Omega,\mathfrak{U},P)} Y_x(\omega)$. We shall call the above r.v.'s shortly *bounded.*

DEFINITION 4. The set $\mathfrak{X}_0 = \{x \in \mathfrak{X}\colon F(x) \geqslant F_0 - \varepsilon\}$ will be called the *ε-optimal set.*

THEOREM 4. *If* (i) *the r.v.'s* $\{Y_x, x \in \mathfrak{X}\}$ *are bounded*; (ii) *the experimental design* Q *dominates* μ; (iii) $w\colon R^1 \to (0,1)$ *is a non-decreasing function*; *then for every* $\varepsilon > 0$ *the process* $(X_n, n = 1, 2, \ldots)$ *defined by* (9) *converges in the discrete topology to the* ε*-optimal set* ${}_\varepsilon\mathfrak{X}_0$ *with probability one.*

Proof. Let $\varepsilon > 0$ be a given number. To prove the theorem it is enough to show that with probability 1 there exists an integer N such that $X_N \in {}_\varepsilon\mathfrak{X}_0$ and $X_{N+k} \in {}_\varepsilon\mathfrak{X}_0$ for all $k = 1, 2, \ldots$

Consider a new "two-dimensional" discrete-time Markov process $\big((X_n, Z_n), n = 1, 2, \ldots\big)$, $X_n \in \mathfrak{X}$, $Z_n \in R^1$, such that

$$P\{X_1 \in S\} = P_1(S) \quad \text{for every } S \in \mathfrak{B},$$

$$P\{Z_1 < z \mid X_1 = x\} = P\{Y_x < z\}$$

and (X_{n+1}, Z_{n+1}), $n = 1, 2, \ldots$ is defined as

$$(16)\qquad (X_{n+1}, Z_{n+1}) = \begin{cases} (X_n, Z_n) & \text{if } \{U_n \leqslant w(Z_n)\} \text{ or} \\ & \{U_n > w(Z_n) \text{ and } Y_{\xi_n} \leqslant Z_n\}, \\ (\xi_n, Y_{\xi_n}) & \text{otherwise,} \end{cases}$$

where U_n, ξ_n, Y_{ξ_n} are defined as before. It is easy to see that the coordinates X_n, $n = 1, 2, \ldots$, form the process given by formula (9) and considered in the previous theorems.

Define the following sets in the product-space $\mathfrak{X} \times R^1$:

$$\Xi = \{(x, z)\colon x \in \mathfrak{X},\ z \in R^1,\ z \leqslant l(x)\},$$

$$\Xi_\varepsilon = \{(x, z) \in \Xi\colon l - \varepsilon \leqslant z \leqslant l\}.$$

It is obvious that $\mu \times P$-almost all points (X_n, Z_n) lie in the set Ξ. The ε-optimal set ${}_\varepsilon\mathfrak{X}_0$ can be obtained as the projection of Ξ_ε on $\mathfrak{X}$.

Now the theorem will follow from the fact that with probability one there exists an integer N such that $(X_N, Z_N) \in \Xi_\varepsilon$ and Ξ_ε is an "absorbing" set, i.e.,

$$P\{(X_{n+1}, Z_{n+1}) \in \Xi_\varepsilon \mid X_n = x, Z_n = z\} = 1$$

whenever $(x, z) \in \Xi_\varepsilon$. The latter part of this statement is obvious by the very definition (16) of the process. To prove the first part of the statement consider the probability $P\{(X_{n+1}, Z_{n+1}) \in \Xi_\varepsilon \mid X_n = x, Z_n = z\}$ for $(x, z) \in \Xi_\varepsilon^c$, where Ξ_ε^c denotes the set $\Xi - \Xi_\varepsilon$. Under the hypothesis $(x, z) \in \Xi_\varepsilon^c$ the event $\{(X_{n+1}, Z_{n+1}) \in \Xi_\varepsilon\}$ may occur if and only if $U_n > w(z)$, $\xi_n \in {}_\varepsilon\mathfrak{X}_0$ and $Y_{\xi_n} > z$. The events $\{U_n > w(z)\}$ and $\{\xi_n \in {}_\varepsilon\mathfrak{X}_0 \text{ and } Y_{\xi_n} > z\}$ are conditionally (given (X_n, Z_n)) independent. For $P\{U_n > w(z)\}$ we

have

$$P\{U_n > w(z)\} = 1 - w(z) \geqslant 1 - w(l-\varepsilon),$$

which is positive. For the probability of the event $\{\xi_n \in {}_\varepsilon\mathfrak{X}_0 \text{ and } Y_{\xi_n} > z\}$ we have

$$P\{\xi_n \in {}_\varepsilon\mathfrak{X}_0 \text{ and } Y_{\xi_n} > z\} = \int_{{}_\varepsilon\mathfrak{X}_0} P\{Y_{\xi_n} > z \mid \xi_n = x\} Q(dx)$$

$$= \int_{{}_\varepsilon\mathfrak{X}_0} P\{Y_x > z\} Q(dx) \geqslant \int_{{}_\varepsilon\mathfrak{X}_0} P\{Y_x \geqslant l-\varepsilon\} Q(dx).$$

From the definition of the set ${}_\varepsilon\mathfrak{X}_0$ we have $\mu({}_\varepsilon\mathfrak{X}_0) > 0$, and because of the dominance of Q we have $Q({}_\varepsilon\mathfrak{X}_0) > 0$ for any $\varepsilon > 0$.

Let $x \in {}_{\varepsilon/2}\mathfrak{X}_0$. Then $l(x) \geqslant l - \varepsilon/2$. But $l(x) = \operatorname{ess\,sup} Y_x$, so that $P\{Y_x \geqslant l(x) - \varepsilon/2\} > 0$. It follows that $P\{Y_x \geqslant l-\varepsilon\}$ is positive for $x \in {}_{\varepsilon/2}\mathfrak{X}_0$, which in turn is a Q-positive subset of ${}_\varepsilon\mathfrak{X}_0$. Thus the integral $\int_{{}_\varepsilon\mathfrak{X}_0} P\{Y_x \geqslant l-\varepsilon\} Q(dx)$ is positive. Denote the positive number $[1 - w(l-\varepsilon)] \cdot \int_{{}_\varepsilon\mathfrak{X}_0} P\{Y_x \geqslant l-\varepsilon\} Q(dx)$ by η; so we have proved that

$$P\{(X_{n+1}, Z_{n+1}) \in \Xi_\varepsilon \mid X_n = x, Z_n = z\} \geqslant \eta > 0$$

for $\mu \times P$-almost all $(x, z) \in \Xi_\varepsilon^c$.

The event $\{(X_n, Z_n) \in \Xi_\varepsilon^c\}$ may occur if and only if $(X_k, Z_k) \in \Xi_\varepsilon^c$ for all $k = 1, 2, \ldots, n-1$. Hence we have

$$P\{(X_n, Z_n) \in \Xi_\varepsilon^c\} = \int_{\Xi_\varepsilon^c} P\{(X_n, Z_n) \in \Xi_\varepsilon^c \mid X_{n-1} = x, Z_{n-1} = z\} \cdot P(dx \times dz)$$

$$\leqslant (1-\eta) P\{(X_{n-1}, Z_{n-1}) \in \Xi_\varepsilon^c\},$$

and by iteration

$$P\{(X_n, Z_n) \in \Xi_\varepsilon^c\} \leqslant (1-\eta)^n.$$

Now

$$P\{\limsup [(X_n, Z_n) \in \Xi_\varepsilon^c]\} = \lim_{n\to\infty} P\Big\{\bigcup_{n=m}^{\infty} [(X_n, Z_n) \in \Xi_\varepsilon^c]\Big\}$$

$$\leqslant \lim_{m\to\infty} \sum_{n=m}^{\infty} P\{(X_n, Z_n) \in \Xi_\varepsilon^c\}$$

$$\leqslant \lim_{m\to\infty} \sum_{n=m}^{\infty} (1-\eta)^n,$$

so that $P\{\limsup [(X_n, Z_n) \in \Xi_\varepsilon^c]\} = 0$, which proves the theorem. ■

6. A Monte Carlo method

Hitherto we have been constructing convergent processes $(X_n, n = 1, 2, \ldots)$ with limit distributions which are in a sense concentrated at the set $\mathfrak{X}_0$. An idea adapted from Monte Carlo methods enables us to construct the process $(X_n, n = 1, 2, \ldots)$ as a sequence of independent identically distributed r.v.'s, the distribution $P\{X_n \in S\}$, $S \in \mathfrak{B}$, being in the sense of Theorems 2 and 3 concentrated "near" the higher values of F.

Let Q be the uniform distribution on $(\mathfrak{X}, \mathfrak{B}, \mu)$ and let $(V_n, n = 1, 2, \ldots)$ be the sequence of independent identically distributed real-valued r.v.'s. Consider the following construction: sample a point $\xi_n \in \mathfrak{X}$ according to the distribution Q and observe the r.v. Y_{ξ_n}. Sample the r.v. V_n. If $V_n \leqslant Y_{\xi_n}$, put $X_n = \xi_n$; otherwise repeat the sampling of ξ_n and V_n.

The above procedure is a generalization of the well-known rejection technique suggested by J. von Neumann [12] in connection with random number generators and used for calculation of the global maximum by Jermakov [8] and independently by Zieliński [19]; the current generalization consists in the fact that $F(x)$ is observed by the r.v. Y_x. An appropriate result is given in Theorem 5 below.

THEOREM 5. *Suppose that* (i) *ξ is a r.v. with the distribution Q which is uniform on $(\mathfrak{X}, \mathfrak{B}, \mu)$*; (ii) *$\{Y_x, x \in \mathfrak{X}\}$ is a family of r.v.'s defined as in* (A3, a *and* b); (iii) *the distribution function G of the r.v. $Y_x - F(x)$ is strictly increasing in an interval containing zero*; (iv) *V is a real valued r.v. with distribution function $G_V(v) = P\{V \leqslant v\}$ which is strictly increasing in an interval $(F_0 - \delta, F_0)$, $\delta > 0$. Let $\Omega_0 = \{\omega\colon Y_{\xi(\omega)}(\omega) \geqslant V(\omega)\}$ and suppose $P(\Omega_0) > 0$. Consider the probability space $(\Omega_0, \mathfrak{A}_0, P_0)$ constructed as the truncation of $(\Omega, \mathfrak{A}, P)$. Then the distribution of the r.v. $X = \xi \mid \Omega_0$ is absolutely continuous with respect to μ and its mode coincides with the global maximum of the function F.*

Proof. Write $P(\Omega_0) = \lambda^{-1}$. Then for $A \in \mathfrak{A}_0$

$$P_0\{X \in A\} = \lambda P\{\xi \in A,\ Y_\xi \geqslant V\}$$

$$= \lambda \int_A P\{Y_\xi \geqslant V \mid \xi = x\} Q(dx)$$

$$= \lambda \int_A Q(dx) \int P\{Y_\xi \geqslant V \mid \xi = x,\ V = v\}\, dG_V.$$

By (A3, b) we have $P\{Y_\xi \geqslant V \mid \xi = x,\ V = v\} = 1 - G\big(v - F(x)\big)$ and hence

$$P_0\{X \in A\} = \lambda \int_A Q(dx) \int \big[1 - G\big(v - F(x)\big)\big]\, dG_V.$$

By $Q(dx) = \mu(dx)/\mu(\mathfrak{X})$ it follows that $P_0 \ll \mu$ and the density P_0 is proportional μ-almost everywhere to

$$C(x) = \int [1 - G(v - F(x))]\, dG_V.$$

Let δ be a positive number such that G_V is strictly increasing in the interval $(F_0 - \delta, F_0)$ and G is strictly increasing in $(-\delta, \delta)$. Denote $\int [1 - G(v - F_0)]\, dG_V$ by C_0. To see that $C(x) < C_0$ whenever $F(x) < F_0$ it is obviously sufficient to prove that $C(x) < C_0$ for x satisfying the inequality $0 < F_0 - F(x) < \delta$. Take an x satisfying this inequality. For every $v \in (F_0 - \delta, F_0)$ we have $-\delta < v - F_0 < v - F(x) < \delta$, and consequently $G(v - F_0) < G(v - F(x))$. Hence $C(x) < C_0$ which proves the theorem. ■

The theorem just proved enables us to construct the sequence of r.v.'s X_n distributed as X above. Then the problem of estimating the global maximum of F reduces to a purely statistical (although rather difficult) problem of estimating the mode of the r.v. X by using the sequence $(X_n, n = 1, 2, \ldots)$ as a random sample.

References

[1] S. H. Brooks, *A discussion of random methods for seeking maxima*, Opns. 6 (1958) pp. 244–251.

[2] J. L. Doob, *Stochastic processes*, Wiley and Sons 1953.

[3] M. Driml and O. Hanš, *On a randomized optimization procedure*, Trans. 4-th Praque Conf. on Information Theory, Statistical Decision Functions, Random Processes, Prague 1967.

[4] V. Fabian, *Stochastic approximation of minima with improved asymptotic speed*, Ann. Math. Statist. 38 (1967), pp. 191–200.

[5] — *Stochastic approximation*, In: *Optimizing methods in statistics*, Ed. J. S. Rustagi. Academic Press, 1971.

[6] Б. П. Харламов, *Об одном алгоритме стохастического поиска максимума в детерминированном поле*, Труды матем. института им. В. А. Стеклова LXXIX (1965), pp. 71–75.

[7] C. C. Heyde, *On martingale limit theory and strong convergence results for stochastic approximation procedures*, Stochastic Processes and their Applications 2 (1974), pp. 359–370.

[8] S. M. Jermakow, *Metoda Monte Carlo i zagadnienia pokrewne*, PWN, Warszawa 1976.

[9] В. Г. Карманов, *О сходимости метода случайного поиска в выпуклых задачах минимизации*, Теория вероят. и ее примен., XIX, 4 (1974) pp. 817–824.

[10] J. Kiefer and J. Wolfowitz, *Stochastic estimation of the maximum of a regression function*, Ann. Math. Statist. 23 (1952), pp. 462–466.

[11] P. Major, *A law of the iterated logarithm for the Robbins–Monro method*, Stud. sci. math. Hung. 8, 1–2 (1973), pp. 95–102.

[12] J. von Neumann, *Various technique used in connection with random digits*, Nat. Bur. Stand. Appl. Math. Ser. 12 (1951), pp. 36–38.

[13] В. Т. Перекрест, *Об одной адаптивной схеме глобального поиска*, УМН 29, 3 (1974), pp. 223–224.

[14] Л. А. Растригин, *Системы экстремального управления*, „Наука", Москва 1974.

[15] H. Robbins and S. Monro, *A stochastic approximation method*, Ann. Math. Statist. 22 (1951), pp. 400–407.

[16] Э. М. Вайсборд, Д. Б. Юдин, *Многоэкстремальная стохастическая аппроксимация*, Изв. АН СССР, Сер. техн. кибернетика. 5 (1968), pp. 3–13.

[17] Э. М. Вайсборд, Д. Б. Юдин, *Стохастическая аппроксимация для многоэкстремальных задач в гильбертовом пространстве*, ДАН СССР 181, 5 (1968), pp. 1034–1037.

[18] M. T. Wasan, *Stochastic approximation*, Cambridge 1969.

[19] R. Zieliński, *On convergence of a randomized optimization procedure*, Algorytmy 7, 12 (1970), pp. 29–32.

[20] — *A Monte Carlo estimation of the maximum of a function*, ibidem 7, 13 (1970), pp. 5–7.

[21] — *A randomized Kiefer–Wolfowitz procedure*, European Meeting of Statisticians and 7-th Prague Conference on Information Theory, Statistical Decision Functions and Random Processes, Prague 1974.