PAPER NO. WA2.2

ON THE PROPERTIES OF CONVERGENCE OF STATISTICAL SEARCH

Luc P. Devroye

Department of Electrical Engineering
The University of Texas at Austin
Austin, Texas 78712

ABSTRACT :The convergence of statistical (random) search for the minimization of an arbitrary function $Q(w)$ is treated.It is shown that random search can be regarded as a gradient algorithm in the q-domain.Using this gradient to define the minimum of the function,the convergence is discussed at length-including convergence WP1,convergence in the mean and ε-optimality.The proof of convergence is based upon the theorems of convergence of random processes of Braverman and Rozonoer.The relationship between random search and order statistics is explained.Finally,emphasis is put on the applicability of the theorems for the design of hierarchical search systems and statistical search with a mixture.

INTRODUCTION :The problem of the minimization of an unknown and perhaps multimodal function $Q(w)$, $w\in R^n$(n-dimensional Euclidean space), has been given considerable attention until now.Usually,multimodal functions are minimized by means of random search or stochastic automata or a combination of both eventually with other direct search techniques [1-6].

The properties of convergence of pure random search given by Matyas [8] are strengthened to convergence WP1.While in [7-9] the convergence of the value of $Q(w)$ at the best estimate of the minimum is studied,the convergence of several other random variables (such as the average measured performance etc.) can sometimes be more important.The asymptotical properties of the discussed class of procedures are studied both from the statistical viewpoint (i.e.,by using the asymptotical properties of order statistics [11-13]) and the machine learning viewpoint (i.e.by using the theorems of convergence of random processes in machine learning [14-15]) Concrete conditions of convergence are obtained and discussed.

It is emphasized that even the class of "anomalous" performance indices falls within the domain of application of the theorems proved in this paper (an anomalous function is such that the value of $Q(w)$ does not convey any information regarding the value at any other point $w^*\neq w$).This generalization is possible through the complete restatement of the problem in the $Q(w)$ domain.

The presented procedure includes random search [7-9] as a special case but has the additional attraction of asymptotically minimizing the average measured performance in analogy with probabilistic automata with a variable structure [10,17].

Though $W\subseteq R^n$(W is the range of w),the presented theorems also apply for arbitrary sets W in which a distribution can be defined.W may eventually be a finite set.

The organization of the search is left to the imagination of the designer within the limits dictated by the conditions of convergence and dependent upon a priori available data about the problem.A brief survey is given of where and how the theorems can be applied ranging from pure random search,creeping random search for local hill-climbing,and mixed random search to very complicated hierarchical search programs.

DESCRIPTION OF THE ALGORITHM :Let $W\subseteq R^n$ either be bounded by constraints or unbounded.It is supposed that no procedure can lead to points outside W.Further,let j denote the iteration counter and X_j denote the <u>state</u> of the search system at time j. The best estimate of the minimum at iteration j ,denoted by w_j,is usually called <u>basepoint</u>. The corresponding value of the <u>performance index</u> is denoted by $q_j=Q(w_j)$.The state X_j includes w_j,q_j and eventually other random elements or adaptive parameters to be specified by the designer (they are of no theoretical importance in the sequel and are,therefore,not explicitly given).X denotes the <u>state space</u>:$X_j\in X$.Notice that X_j is a random element on some (usually growing) probability space and as such,functions of $X_j:X\to R^1$ are random variables on this probability space.

Let the initial distribution of X_0 be such that w_0 is a random vector of W with pdf $g_0(w)$.The procedure consists of iterating the following operations:

1°)A <u>trial point</u> w^*_{j+1} is generated (computed) by means of a procedure that is completely determined by the state X_j.For example,w^*_{j+1} may be randomly generated according to a known pdf or w^*_{j+1} can be the result of a heuristic rule or other direct search operation.In any case,there exists a pdf $g(w/X_j)$ on W which is the pdf of w^*_{j+1} conditioned on X_j.If $\{p_{sj}\}$ is a sequence of probabilities,it is required that $w^*_{j+1}=w_j$ at least with probability p_{sj}.In mathematical terms,this means that there exists a pdf on W,say $g_0(w/X_j)$, such that :

$$(1)\qquad g(w/X_j)=p_{sj}\cdot g_0(w/X_j)+(1-p_{sj})\cdot\delta(w-w_j)$$

2°)$Q(w^*_{j+1})$ is observed (computed,measured,etc.).

3°)The new state X_{j+1} is computed.Since we are only interested in the couple (w_j,q_j),the way of updating other adaptive elements contained in X_j is not explicitly given here :

$$(2)\quad (w_{j+1},q_{j+1}) = \begin{cases} (w^*_{j+1},Q(w^*_{j+1})) & \text{if } Q(w^*_{j+1})<q_j \\ (w_j,q_j) & \text{otherwise} \end{cases}$$

This algorithm reduces to the random search scheme of Matyas [8],Gurin [9],etc. for $p_{sj}=1$ and special forms of $g_0(w/X_j)$.

STATE FUNCTIONS : $g(w/X_j)$ induces in the performance index domain R^1 a search pdf $p(q/X_j)$ defined by:

$$(3)\quad p(q/X_j) \triangleq \int_W \delta(q-Q(w)).g(w/X_j).dw$$

with corresponding search cdf (cumulative distribution function):

$$(4)\quad P(q/X_j) \triangleq \int_W \omega(Q(w),q).g(w/X_j).dw = \int_{-\infty}^{q]} p(u/X_j).du = \text{Prob}\{Q(w^*_{j+1})\leq q/X_j\}$$

$P(q/X_j)$ is thus the distribution function of $Q(w^*_{j+1})$ conditioned on X_j.The function $\omega(a,b)$ is a threshold function:

$$(5)\quad \omega(a,b) \triangleq \begin{cases} 1 & \text{if } a\leq b \\ 0 & \text{if } a>b \end{cases}$$

Notice that if W is an arbitrary set,a pdf $g(w/X_j)$ can not always be found but the search pdf corrects this discomfort and permits the theoretical study of optimization schemes in arbitrary sets.
Let there also be given a fixed pdf in W,$g_B(w)$, the domain of concentration of which defines W.For instance,if a minimum is looked for in a hypercube,then $g_B(w)$ might be the uniform pdf in this hypercube. If $g_B(w)$ is purely atomic with a countable number of atoms,W is a countable set of points in the n-dimensional Euclidean space.If the minimum can not be restricted a priori to a certain bounded region of R^n,but if,at the same time,there are strong suspicions that the minimum is close to $w^0\epsilon W$,then $g_B(w)$ may be gaussian centered at w^0 with fixed positive definite covariance matrix.In general,$g_B(w)$ is thus a pdf covering W according to which a point w* is generated in the absence of any information.The pdf and cdf induced by $g_B(w)$ in the q-domain are referred to as basic search pdf and basic search cdf:

$$(6)\quad p_B(q) = \int_W \delta(q-Q(w)).g_B(w).dw$$

$$(7)\quad P_B(q) = \int_W \omega(Q(w),q).g_B(w).dw = \int_{-\infty}^{q]} p_B(u).du$$

Thus,$p_B(q)$ counts in a certain sense the portion of points in W for which Q(w)=q.The existence of this basic search distribution is essential for the discussed class of procedures.In addition, $P_B(q)$ will be used to define the minimum q_{min} of Q(w) in W.
It is the purpose to study the asymptotical behaviour of state functions such as (8-10):

$$(8)\quad q_j=Q(w_j)$$

$$(9)\quad \omega(q_j,q_{min}+\varepsilon) = \text{Ind}\{q_j\leq q_{min}+\varepsilon\}$$

$$(10)\quad \int_{-\infty}^{q_j]} P_B(u).du$$

where $\varepsilon>0$; (9) represents the indicator function of the event $\{q_j\leq q_{min}+\varepsilon\}$.The properties of (10) are discussed in the next section.In general, (8-10) are random variables of interest in classical optimization.For some problems however (such as the problem of learning in a random environment [3,6,10,17],etc.) it is also desirable to minimize the average measured performance (11) or to maximize the degree of concentration on $(-\infty,q_{min}+\varepsilon]$ defined by (12).Both (11) and (12) are clearly random variables that fall under the category of "state functions".

$$(11)\quad \psi(X_j) \triangleq \int q.p(q/X_j).dq = E\{Q(w^*_{j+1})\ /X_j\}$$

$$(12)\quad D(X_j,\varepsilon) \triangleq P(q_{min}+\varepsilon/X_j) = \text{Prob}\{Q(w^*_{j+1})<q_{min}+\varepsilon\ /X_j\}$$

Finally,to relieve the notational burden,the bar operator $\bar{a}$ will sometimes be used to denote $\text{Max}\{a,q_{min}\}$.

DEFINITION OF THE MINIMUM :Although the minimum is well defined in optimization problems for continuous performance indices,it must be redefined here in view of the broad class of performance indices and sets W to be allowed.
Consider the goal function (13):

$$(13)\quad I(q) \triangleq \int_{-\infty}^{q]}\int_{-\infty}^{x]} P_B(u).du$$

whose gradient I'(q) is given in (14) if $u.P_B(u)\rightarrow 0$ for $u\rightarrow-\infty$.

$$(14)\quad I'(q)=\int_{-\infty}^{q]} P_B(u).du = \int_{-\infty}^{q]} (q-u).p_B(u).du = P_B(q).[q - \int_{-\infty}^{q]} u.\frac{p_B(u)}{P_B(q)}.du\ ]$$

The goal function I(q) is convex since its second derivative equals $P_B(q)>0$.Further,the gradient is zero only if either $P_B(q)=0$ (in which case one can say that there is probability zero that the random variable Q(w) with search pdf (6) is less than or equal to q) or the mean of the conditional random variable $Q(w)/Q(w)\leq q$ equals q.
The minimal value (minimum) q_{min} of Q(w) with respect to $g_B(w)$ is the greatest q for which gradient (14) is zero:

$$(15)\quad \int_{-\infty}^{q_{min}]} P_B(u).du = 0$$

$$(16)\quad \varepsilon>0,\ \delta(\varepsilon)>0\ :\ \int_{-\infty}^{q_{min}+\varepsilon]} P_B(u).du\geq\delta(\varepsilon)>0$$

Notice here that $P_B(q_{min})$ is not necessarily 0.One can easily imagine a staircase function Q(w) with a set of points $w\epsilon W:Q(w)=q_{min}$,this set having nonzero measure with respect to $g_B(w)$.In that case however,the second factor in the right-hand term of (14) is zero.Conversely,it can be proved that to each $\varepsilon>0$,there exists a number $\theta(\varepsilon)>0$ such that $P_B(q_{min}+\varepsilon)\geq\theta(\varepsilon)>0$ .Otherwise,a contradiction with relations (15-16) would be obtained.
In the sequel it will be proved that the presented procedure can be organized in such a way that

$\bar{q}_j$ and even $\bar{\psi}(X_j)$ converge to q_{min} in the appropriate way, i.e. WP1, in probability, in the mean, etc..

RANDOM SEARCH AS A GRADIENT ALGORITHM : If $g_0(w/X_j)$ (1) is such that at least with probability α_j ($\alpha_j \in [0,1]$), w^*_{j+1} is a random variable with pdf $g_B^j(w)$, then there exists a pdf $g_1(w/X_j)$ such that

(17) $g_0(w/X_j) = \alpha_j . g_B(w) + (1-\alpha_j).g_1(w/X_j)$

Using definitions (3-7), the following inequalities are obtained:

(18) $p_0(q/X_j) \geq \alpha_j . p_B(q)$; $P_0(q/X_j) \geq \alpha_j . P_B(q)$

where $p_0(q/X_j)$ and $P_0(q/X_j)$ are the density, respectively distribution function induced in the q-domain by $g_0(w/X_j)$ (see (3-4)).

Theorem 1: If $E\{q_0\}$ exists, if $|q_{min}|<\infty$ and if $p_{sj} \in [0,1]$ and $\alpha_j \in [0,1]$ are numbers only dependent upon j:

(19) $\sum_{j=1}^{\infty} p_{sj}.\alpha_j = \infty$

then the state sequence $\{X_j\}_{j>0}$ generated through the described procedure is such that

(20) $\bar{q}_j \to q_{min}$ WP1 as $j\to\infty$

(21) $\omega(\bar{q}_j, q_{min}+\varepsilon) \to 1$ WP1 as $j\to\infty$ (for every $\varepsilon>0$)

(22) $\int_{-\infty}^{q_j]} P_B(u).du \to 0$ WP1 as $j\to\infty$

Proof: Introduce the nonnegative state functions $U(X_j)$ and $V(X_j)$:

(23) $U(X_j) = \bar{q}_j - q_{min}$

(24) $V(X_j) = (\bar{q}_j - q_{min}).P_B(q_{min}) + \int_{-\infty}^{q_j]} P_B(u).du$

It will first be proved that:

(25) $E\{U(X_{j+1})/X_j\} \leq U(X_j) - p_{sj}.\alpha_j.V(X_j)$

From the description of the process 1°)-3°), it follows that:

$$q_{j+1} = \begin{cases} q_j & \text{w.p. } 1 - p_{sj}.P_0(q_j/X_j) \\ u & \text{w.p. } p_{sj}.\omega(u,q_j).p_0(u/X_j) \end{cases}$$

and thus:

(26) $E\{\bar{q}_{j+1}/X_j\} = \bar{q}_j - p_{sj}.\int_{-\infty}^{q_j]} (\bar{q}_j - \bar{u}).p_0(u/X_j).du$

Since in the integration interval $(-\infty, q_j]$, $\bar{q}_j > \bar{u}$, inequality (18) may be applied to yield:

(27) $E\{\bar{q}_{j+1}/X_j\} \leq \bar{q}_j - p_{sj}.\alpha_j.\int_{-\infty}^{q_j]} (\bar{q}_j - \bar{u}).p_B(u).du$

For $u \in (-\infty, q_{min}]$: $\bar{q}_j - \bar{u} = \bar{q}_j - q_{min}$ and for $u \in (q_{min}, q_j]$: $\bar{q}_j - \bar{u} = q_j - u$. Then:

$$\int_{-\infty}^{q_j]} (\bar{q}_j - \bar{u}).p_B(u).du = \int_{-\infty}^{q_{min}]} (\bar{q}_j - q_{min}).p_B(u)du + \int_{(q_{min},}^{q_j]} (q_j - u).p_B(u).du = V(X_j)$$

in view of property (14) (which holds if $|q_{min}|<\infty$) and definition (24). Combining the last expression with (27) gives (25).

By definition (15-16) of q_{min} and by definition of the bar operator, there follows that $U(X_j) \geq 0$, $V(X_j) \geq 0$ and only for $\bar{q}_j = q_{min}$: $U(X_j) = V(X_j) = 0$. Further, both are continuous functions of $\bar{q}_j$. Then, if $\{X_j\}_{j>0}$ is such that $V(X_j) \to 0$ (deterministically, in prob., WP1), then also: $U(X_j) \to 0$ in the same way (deterministically, in probab., WP1) and vice versa. (for instance, see lemma 1 of Braverman and Rozonoer [15]).

This latter property plus the existence of $E\{\bar{q}_0 - q_{min}\}$, $p_{sj}.\alpha_j \geq 0$ and (19) is sufficient for the theorems of convergence of random processes of Braverman and Rozonoer ([14], th.3, th.5) to be applicable to the described process for which inequality (25) holds. Thus, $U(X_j) \to 0$ WP1 as $j\to\infty$ and $V(X_j) \to 0$ WP1 as $j\to\infty$. This implies (20) and (22).

The meaning of (20) is that:

(28) $\lim_{J\to\infty} \text{Prob}\{\text{Max}_{j>J}\{\bar{q}_j\} \leq q_{min}+\varepsilon\} = 1$ for every $\varepsilon>0$

Thus, for every $\varepsilon>0$:

(29) $\lim_{J\to\infty} \text{Prob}\{\text{Min}_{j>J}\ \omega(\bar{q}_j, q_{min}+\varepsilon) = 1\} = 1$

which corresponds with property (21). Notice that $\omega(\bar{q}_j, q_{min}+\varepsilon) = \omega(q_j, q_{min}+\varepsilon)$. (21) can also be directly proved. From the description of the process, there follows:

(30) $E\{(1-\omega(\bar{q}_{j+1}, q_{min}+\varepsilon))/X_j\} = (1-\omega(\bar{q}_j, q_{min}+\varepsilon)).$
$(1-p_{sj}.P_0(q_{min}+\varepsilon/X_j))$
$\leq (1-\omega(q_j, q_{min}+\varepsilon)).(1-p_{sj}.\alpha_j.P_B(q_{min}+\varepsilon))$

Since $E\{1-\omega(\bar{q}_0, q_{min}+\varepsilon)\}$ always exists and since $P_B(q_{min}+\varepsilon) \geq \theta(\varepsilon) > 0$ and $\Sigma p_{sj}.\alpha_j = \infty$: $\omega(\bar{q}_j, q_{min}+\varepsilon) \to 1$ WP1 as $j\to\infty$.

Since this holds for every $\varepsilon>0$, (21) is valid. ■

Corollary: The convergence in probability of the state functions (8-10) follows from (20-22). If, in addition :

(31) $Q(w) \leq q_{max} < \infty$ for all $w \in W$

then:

(32) $\lim_{j\to\infty} E\{\bar{q}_j\} = q_{min}$

which implies convergence in the mean of $\bar{q}_j$ to q_{min} since $\bar{q}_j \geq q_{min}$.

Remark: Note that proving the convergence in probability for state functions (8) and (9) is equivalent since

(33) $\text{Prob}\{\bar{q}_j < q_{min}+\varepsilon\} = E\{\omega(\bar{q}_j, q_{min}+\varepsilon)\} =$
$\text{Prob}\{\omega(\bar{q}_j, q_{min}+\varepsilon) = 1\}$

RELATED PROPERTIES OF CONVERGENCE : Note that from (1), (3), there holds:

(34) $p(q/X_j) = p_{sj}.(\alpha_j.p_B(q) + (1-\alpha_j).p_1(q/X_j))$
$+ (1-p_{sj}).\delta(q-q_j)$

where $p_1(q/X_j)$ is the pdf in the q-domain induced by $g_1(w/X_j)$ (17).Expression (34) links q_j directly with the statistics of $Q(w^*_{j+1})$ through the search pdf $p(q/X_j)$.This enables us to prove theorem 2:

Theorem 2:If the conditions of Theorem 1 are fulfilled and (36) holds,then:

$$(35) \quad D(X_j,\varepsilon) \to 1 \quad \text{WP1 as } j\to\infty \quad (\text{for every } \varepsilon>0)$$

$$(36) \quad \lim_{j\to\infty} p_{sj} = 0$$

If,in addition,(31) holds,then also:

$$(37) \quad \bar{\psi}(X_j)\to q_{min} \quad \text{WP1 as } j\to\infty$$

Proof:Let μ_B denote the mean of the basic search pdf:

$$(38) \quad \mu_B = \int q.p_B(q).dq$$

Multiplying (34) by q and taking expectations at both sides leads to inequality (39) by definition of the average measured performance (11) and $Q(w)\leqslant q_{max}<\infty$:

$$(39) \quad q_{min}\leqslant\bar{\psi}(X_j)\leqslant p_{sj}.(\alpha_j.\mu_B+(1-\alpha_j).q_{max}) +(1-p_{sj}).\bar{q}_j$$

Note that the upper limit tends to q_{min} WP1 as $j\to\infty$ .Indeed,$p_{sj}\to 0$ (36) and $\bar{q}_j\to q_{min}$ WP1 as proved in theorem 1 (20).Thus, $\bar{\psi}(X_j)\to q_{min}$ WP1. as $j\to\infty$ (37).

Integrating both sides of (34) from $-\infty$ to $q_{min}+\varepsilon$ and using definitions (7),(12) and assuming the worst case $(P_1(q_{min}+\varepsilon/X_j)=0)$ gives the following inequality:

$$(40) \quad 1\geqslant D(X_j,\varepsilon)\geqslant p_{sj}.(\alpha_j.P_B(q_{min}+\varepsilon)+(1-\alpha_j).0) +(1-p_{sj}).\omega(q_j,q_{min}+\varepsilon)$$

(35) holds since $\omega(q_j,q_{min}+\varepsilon) \to 1$ WP1 (Th.1) and $p_{sj}\to 0$ (36).This completes the proof. ∎

Remark 1:The convergence in probability of $D(X_j,\varepsilon)$ to 1 follows directly from (35) or from the convergence in probability of q_j to q_{min},property (33),condition (36) and inequality (40). The convergence in probability of $\bar{\psi}(X_j)$ to q_{min} follows directly from (37) or from the convergence in probability of $\bar{q}_j$ to q_{min},(31),(33), (36) and inequality (39).

Remark 2:From $Q(w)\leqslant q_{max}<\infty$ (31) and $\bar{\psi}(X_j)\to q_{min}$ in probability,it follows that:

$$(41) \quad \lim_{j\to\infty} E\{\bar{\psi}(X_j)\} = q_{min}$$

Convergence in the mean follows by noting that $\bar{\psi}(X_j)\to q_{min}$ and by applying Markov's inequality.

Remark 3:Sometimes,one is interested in the behaviour of the random variable $Q(w^*_{j+1})$ which is defined on a bigger probability space than the state X_j and therefore not a state function.Taking the expectation over all the realizations $X_j\in X$ in inequality (40) yields, for every $\varepsilon>0$:

$$(42) \quad 1\geqslant E\{D(X_j,\varepsilon)\}=\text{Prob}\{Q(w^*_{j+1})\leqslant q_{min}+\varepsilon\} \geqslant p_{sj}.\ \alpha_j.P_B(q_{min}+\varepsilon)+(1-p_{sj}).\text{Prob}\{\bar{q}_j\leqslant q_{min}+\varepsilon\}\geqslant(1-p_{sj}).\text{Prob}\{\bar{q}_j\leqslant q_{min}+\varepsilon\}$$

From (20)* it follows that $\text{Prob}\{\bar{q}_j\leqslant q_{min}+\varepsilon\}\to 1$.If $p_{sj}\to 0$,$\bar{Q}(w^*_{j+1})$ is a random variable converging in probability to q_{min} as $j\to\infty$ .However convergence WP1 can not be proved ,due to $\Sigma p_{sj}.\alpha_j=\infty$,an infinite number of points will be generated according to $g_B(w)$.But ,if the sequence $\{Q(w^*_j)\}_{j>0}$ contains an infinite subset of iid random variables with pdf $p_B(q)$,convergence WP1 is excluded.Indeed,this would mean that the sequence $\{Q(w^*_j)\}_{j>0}$ can only finitely often be above every level $q_{min}+\varepsilon,\varepsilon>0$.

RANDOM SEARCH AND ORDER STATISTICS :The state function q_j can also be studied through the properties of order statistics [11-13].Indeed,notice that

$$q_j=\text{Min}\{q_0,Q(w^*_1),Q(w^*_2),\ldots,Q(w^*_j)\}$$

where the pdf $p(q/X_j)$ of $Q(w^*_{j+1})$ is such that,independent of X_j and for all $j=1,2,\ldots$:

$$(43) \quad \text{Prob}\{Q(w^*_{j+1})\leqslant q_{min}+\varepsilon\}\geqslant\alpha_j.p_{sj}.P_B(q_{min}+\varepsilon)$$

which is obtained by integrating (34).Accordingly,

$$(44) \quad \text{Prob}\{q_j\leqslant q_{min}+\varepsilon\}\geqslant 1-\prod_{i=1}^{j}(1-p_{si}.\alpha_i.P_B(q_{min}+\varepsilon))$$

which,in view of $P_B(q_{min}+\varepsilon)\geqslant\theta(\varepsilon)>0$ and $\Sigma p_{sj}.\alpha_j=\infty$ tends to 1 as $j\to\infty$.

Some properties of the indicator function $\omega(q_j,q_{min}+\varepsilon)$ (9) are also obtainable from generalizations of the lemma of Borel [18,24] but the same condition (19) inevitably appears as the condition of convergence.

EXAMPLES :Theorems 1,2 do still hold if p_{sj},α_j are adapted between upper and lower boundaries that are number sequences satisfying (19) and (36),in which case p_{sj} and α_j are "components" of the state X_j.If the convergence of the average measured performance is not needed,there is no reason for not taking $p_{sj}=1$.

If $p_{sj}\to 0$ (36), the search virtually stops after some time.In slightly nonstationary environments however -i.e.Q(w) is subjected to small changes as time passes-,a minimal amount of basic search is desired at any moment to make up for the "aging" of the basepoint.It is therefore of definite interest to study constant search parameter algorithms:$p_{sj}=C_p$;$\alpha_j=C_\alpha$ for all $j>0$.It can be proved that

$$(45) \quad \lim_{C_p\to 0}\lim_{j\to\infty} E\{D(X_j,\varepsilon)\} = 1 \text{ for every } \varepsilon>0$$

if $C_p>0$ and $C_\alpha>0$.Indeed,theorem 1 holds and extending (42) yields:

$$(46) \quad E\{D(X_j,\varepsilon)\}\geqslant C_p.C_\alpha.P_B(q_{min}+\varepsilon)+(1-C_p).E\{\omega(q_j,q_{min}+\varepsilon)\}$$

(45) follows from (21) and (33).Expression (45) is in complete agreement with the definition of ε-optimality formulated in connection with finite set stochastic automata [6,17] where there also exists a positive function of the search parameters (here: C_p) whose convergence to zero implies asymptotical concentration of the search pdf on the minimum (3,12,45).If the desired asymptotical fault is given,say ε',then,if $C_p=\varepsilon'$,

$$\lim_{j\to\infty} E\{D(X_j,\varepsilon)\} \geqslant 1 - \varepsilon' \text{ for every } \varepsilon>0$$

For such constant search parameter algorithms it is also possible to extend the idea of expediency first defined by Tsetlin [22] for finite set automata. In terms of the average measured performance (11) and the mean of the basic search pdf μ_B (38), the definition of expediency reads:

$$(47)\quad \lim_{j\to\infty} E\{\bar{\psi}(X_j)\} \leq \mu_B \triangleq \int q.p_B(q).dq$$

For the presented technique, if $C_p>0$, $C_\alpha>0$, theorem 1 holds. If (31) holds, inequality (39) can be written. Replace p_{sj} and α_j in (39) by C_p and C_α, take the expectations at both sides of (39) and substitute $E\{\bar{q}_j\}$ by q_{min} for $j\to\infty$. Then, the condition of expediency (47) reduces to:

$$(48)\quad C_p \leq \frac{(q_{max}-q_{min}) - (q_{max}-\mu_B)}{(q_{max}-q_{min})-C_\alpha.(q_{max}-\mu_B)}$$

which is always fulfilled for C_p small enough or for $C_\alpha=1$.

The specific field of application of random search and probabilistic automata is the optimization of "difficult" performance indices [1], i.e. multimodal, non-continuous, non-differentiable, etc. performance indices. In the limit such probabilistic procedures can be used for anomalous $Q(w)$ for which, by definition, no hypothesis concerning the behaviour of $Q(w)$ holds. Hence, no other search strategy can be followed than to take $g_j(w/X_j)=g_B(w)$. Thus, $\alpha_j=1$. Such procedures are called "pure probabilistic search" procedures. To illustrate the slowness of these procedures in high-dimensional Euclidean spaces for non-anomalous functions $Q(w)$, consider the following example. $W\subseteq R^n$ is a hypersphere with radius 1 centered in the origin and $Q(w)=|w|^\beta$, $\beta>0$. Let $g_B(w)$ denote the uniform pdf in W and $\alpha_j=p_{sj}=1$. Noticing that $q_{min}=0$ and $P_B(q)=q^{n/\beta}$, (44) reads:

$$\text{Prob}\{\,Q(w_j)\leq q\} = 1 - (1-q^{n/\beta})^{j+1}$$

This result was first obtained by Brooks [7].

In reality, most $Q(w)$ are non-anomalous and pure probabilistic search procedures are seldom used. The option that α_j might tend to 0 for $j\to\infty$ and that nothing is required concerning $g_1(w/X_j)$ permits a gradual switch to non-random direct search techniques as the search proceeds. It was mentioned in [19] that "random search" is not particularly fastly converging near the minimum and therefore, other direct search techniques (gradient, stochastic gradient, conjugate gradient, simplex, etc.) seem desirable at the end of the optimization. Condition (19) concerns the rate with which such a gradual switch can be made. This class of methods, very popular for multimodal functions, is called "mixed search" class of procedures.

Recently, more sophisticated 100% probabilistic techniques are being studied for use in multimodal optimization [2,5,21]. $g_0(w/X_j)$ is a mixture of different pdfs with adaptive weights and adaptive parameters in each mixture component pdf. Commonly, one component is the uniform distribution in W and the other components either gaussian with adaptive statistics [2,8] or uniform in subdomains of W. The parameters are adapted in order to concentrate the search effort on promising regions [2,5] or in order to avoid sampling in high-loss regions [21]. If $g_B(w)$ is the uniform pdf in W and α_j is the weight of $g_B(w)$ in the mixture, the above given discussion of the asymptotical properties of statistical search procedures can be applied to this group of procedures, sometimes referred to as "random search with a mixture".

The presented procedure has even applications in local hill-climbing in order to avoid absorption on "defects" of $Q(w)$. Assume that $p_{sj}=1$ and $g_0(w/X_j)$ represents the pdf in W used for local hill-climbing. As a rule, $g_0(w/X_j)$ is a pdf centered at w_j and uniformly concentrated on a hypersphere, hypercube, etc. or concentrated on the corners of a hypercube or uniformly in a hypercone, hypersphere etc. [5,8,16,19,23] (for a survey, see [1]). Expression (26) evolves in:

$$(49)\quad E\{q_{j+1}/X_j\} = q_j - \int_{-\infty}^{q_j]} P_0(u/X_j).du$$

(49) is a kind of inequality dealt with in Theorem 5 of Braverman and Rozonoer [14] and, accordingly,

$$\int_{-\infty}^{q_j]} P_0(q/X_j).dq \to 0 \quad \text{WP1 as } j\to\infty$$

Thus, stationary points w^* of the search are satisfying:

$$(50)\quad \int_{-\infty}^{Q(w^*)} \int_{-\infty}^{+\infty} \omega(Q(w),u).g_0(w/X^*).dw.du = 0$$

where X^* is a stationary state for the process.

It is not necessarily true that all the states for which (50) holds are stationary points of the search process. For instance, if w/X_j is uniformly distributed on a hypersphere with radius R centered at w_j, then (50) holds for all w^* for which there are no points on the hypersphere with $Q(w)<Q(w^*)$. Thus, w^* is either on a flat level or near a local minimum (if $Q(w)$ is continuous) or near the boundary of a constraint region. In the second case, the local minimum cannot be further away from w^* than $2R$, with which value the asymptotical accuracy of the solution can be controlled. On the contrary, if $g_0(w/X_j)$ is the uniform pdf in the hypersphere or a gaussian density $N(w_j,\sigma_j^2 I)$ exact local minimization will be achieved by the same reasoning. The situation is worse if $g_0(w/X_j)$ does not "enclose" w_j, such as, for instance for purely atomic densities. The stability condition (50) might be satisfied for many points including some that are not even close to any local minimum. It is clear that in creeping random search it is preferable to search with densities $g_0(w/X_j)$ that are dense for all small $|w-w_j|$. Flat levels and boundaries can be viewed as traps for the basepoint even if $Q(w)$ is unimodal and continuous. If $g_0(w/X_j)$ depends upon "second level" adaptive parameters (such as the variance of a gaussian pdf), the set of stationary states X^* can only grow and extra preventions should be made against a too rapid contraction of $g_0(w/X_j)$ on w_j. In gaussian creeping random search with adaptive variance ([5,8] etc.) absorption on one of the "traps" can usually not be avoided unless the variance is prohibited to decrease too rapidly. If W is bounded, d is the maximum distance between two points of W, $g_B(w)$ defining q_{min} (15-16) is uniform in W then α_j=Min $g_0(w/X_j)$ over all $w\in W$. Condition (19) reads:

$$(51)\quad \sum_{j=1}^{\infty} p_{sj}\cdot\sigma_j^{-n}\cdot\exp\left(-\frac{d^2}{\sigma_j^2}\right) = \infty$$

Slightly superior experimental results have been obtained by combining creeping random search with linear search in analogy with direct search algorithms of Rosenbrock's type [20] (see [1] for a survey). Each iteration with random search is followed by some number (say, M) of iterations, consisting of linear search in a specially determined direction. Notice that during these linear search steps $\alpha_j=0$ and consequently, the value of p_{sj} is irrelevant (19) whenever such linear search steps are made.

Large scale optimization programs usually consist of a combination of several direct search techniques. Complex systems in which several "levels" can be distinguished (or: hierarchical search systems) were the object of several experimental research projects [2-6] and hopeful results have been obtained by combining stochastic automata, random search and other direct search techniques in multimodal optimization problems. Most of the programs are heuristically derived, however, and the proof of convergence is usually lacking. For a large group of hierarchical search systems, an elegant proof of convergence can be given using the theorems developed in this paper. As an example, let W be partitioned into S domains and let each domain be partitioned into S subdomains etc.. If there are M such levels, the number of blocks equals S^M. In each block search can be performed according to a simple uniform random law. The selection of one block is done through M probabilistic automata each with S probabilities. If $g_B(w)$ is uniform in W, then condition (19) can be translated in terms of the maximal rate of decrease of the S.M selection probabilities of the automata. In view of its importance this topic deserves special separate treatment.

CONCLUSION: There is a lack of theoretical study of the properties of statistical search procedures when compared to other search techniques. In this paper, important concepts for the analytical treatment of random search systems are brought together, redefined or newly introduced. The asymptotical features of the technique are dealt with in the q-domain. Based upon the connection between random search and probabilistic automata, the definitions of expediency, optimality and ε-optimality are extended for use in random search. Convergence of the average measured performance is proved for the class of presented algorithms.

In a similar but somewhat more complicated way, stochastic performance indices can be dealt with.

The basic theorems for proving the convergence are the theorems of Braverman and Rozonoer [14-15] that are more general than the theorem of Dvoretzky. As this way of dealing with random search is new, further research seems necessary to develop direct search techniques for use in stochastic optimization problems that are constructed similarly to their counterparts for use in deterministic optimzation. Finally, it is shown how the approach presented here can be used as a framework for devising and proving convergence for hierarchical search systems and random search with a mixture.

BIBLIOGRAPHY :

[1] G.J.MCMURTRY:"Adaptive optimization procedures" in:"Adaptive,Learning And Pattern Recognition Systems",J,M.Mendel,K.S.Fu,Ed.,Ac.Press,1970

[2] R.A.JARVIS:"Adaptive global search in a time-variant environment using a probabilistic automaton with pattern recognition supervision",IEEE,Trans.on SSC,vol.SSC-6,No.3,pp.209-217,1970

[3] G.J.MCMURTRY,K.S.FU:"A variable structure automaton used as a multimodal searching technique" IEEE,Trans.on AC,vol.AC-11,No.3,pp.379-387,1966

[4] J.D.HILL:"A search technique for multimodal surfaces",IEEE,Trans.on SSC,vol.SSC-5,No.1,pp.1-11,1969

[5] L.D.COCKRELL,K.S.FU:"On search techniques in adaptive systems",Techn.Rept.TR-EE-70-01,Purdue Univ.,Lafayette,Ind.,1970

[6] R.VISWANATHAN,K.S.NARENDRA:"Application of stochastic automata models to learning systems with multimodal performance criteria",Becton Center,Techn.Rept. CT-40,Yale Univ.,New Haven, Conn.,1972

[7] S.H.BROOKS:"A discussion of random methods for seeking maxima",Op.Res.,vol.6,No.2,pp.244-251, 1958

[8] J.MATYAS:"Random optimization",Aut.and Rem.Control,vol.26,No.2,pp.244-251,1965

[9] L.S.GURIN:"Random search in the presence of noise",Eng.Cybern.,vol.4,No.3,pp.252-260,1966

[10] I.J.SHAPIRO,K.S.NARENDRA:"The use of stochastic automata for parameter self-optimization with multimodal performance criteria",IEEE,Trans.on SSC,vol.SSC-5,No.4,pp.352-360,1969

[11] H.A.DAVID:"Order Statistics",John Wiley,1970

[12] S.S.WILKS:"Mathematical Statistics",John Wiley, 1962

[13] M.FISZ:"Probability Theory And Mathematical Statistics",John Wiley,1963

[14] E.M.BRAVERMAN,L.I.ROZONOER:"Convergence of random processes in learning machines theory,part 1",Aut.and Rem.Control,vol.30,No.1,pp.44-64,1969

[15] ——:"Convergence of random processes in learning machines theory,part 2",Aut.and Rem.Control,vol.30,No.3,pp.386-402,1969

[16] B.T.POLYAK,Ya.Z.TSYPKIN:"Pseudogradient adaptation and training algorithms",Aut.and Rem.Control,vol.34,No.3,pp.377-397,1973

[17] R.VISWANATHAN,K.S.NARENDRA:"A note on the linear reinforcement scheme for variable structure stochastic automata",IEEE,Trans.on SMC,vol.SMC-2,No.2,pp.292-294,1972

[18] S.W.NASH:"An extension of the Borel Cantelli lemma",Ann.Math.Stat.,vol.25,No.1,pp.165-166,1954

[19] L.S.GURIN.L.A.RASTRIGIN:"Convergence of the random search method in the presence of noise",Aut. and Rem.Control,vol.26,No.9,pp.1505-1511,1965

[20] M.A.SCHUMER,K.STEIGLITZ:"Adaptive step size random search",IEEE,Trans.on AC,vol.AC-13,No.3,pp. 270-276,1968

[21] A.N.MUCCIARDI:"A new class of search algorithms for adaptive computation",Proc.12th Symp.on Ad. Processes,San Diego,Cal.,pp.94-100,1973

[22] M.TSETLIN:"On the behaviour of finite automata in random media",Aut.and Rem.Control,vol.22, No.10,pp.1345-1354,1961

[23] A.S.POZNYAK:"Use of learning automata for the control of random search",Aut.and Rem.Control, vol.33,No.12,pp.1992-2000,1972

[24] M.LOEVE:"On almost sure convergence",Proc.Sec. Berkeley Symp.,vol.2,pp.177-180,1956