Exploiting the Connections of Evolutionary Algorithms and Stochastic Approximation

G. Yin[*], G. Rudolph[†] and H.-P. Schwefel[‡]

August 1994

Abstract

This work is our first attempt in establishing the connections between evolutionary computation algorithms and stochastic approximation procedures. By treating evolutionary algorithms as recursive stochastic procedures, we study both constant gain and decreasing step size algorithms. We formulate the problem in a rather general form, supply the sufficient conditions for convergence (both with probability one, and in the weak sense). Among other things, our approach reveals the natural connection of the discrete iterations and the continuous dynamics (ordinary differential equations, and/or stochastic differential equations). We hope that this attempt will open up a new domain for further research and lead to in depth understanding of the underlying algorithms.

Key words: evolutionary computation, evolution strategy, stochastic approximation, convergence, rate of convergence.

Abbreviated title. Connections of EA and SA

[*] Department of Mathematics, Wayne State University, Detroit, MI 48202. This research was supported in part by the National Science Foundation under grant DMS-9224372, and in part by the Deutscher Akademischer Austauschdienst.

[†] Informatik Centrum Dortmund, Joseph-von-Fraunhofer-Str. 20, D-44227, Dortmund, Germany. The research of this author was supported by BMFT under grant 01 IB 403A, project EVOALG.

[‡] Universität Dortmund, FB Informatik 11, D-44221, Dortmund, Germany.

1 Introduction

The main objective of this work is to make effort to establish the connections between evolutionary algorithms (EA) and recursive algorithms of stochastic approximation type. One of our hopes is that with the help of the existing results in stochastic approximation and the bridge between evolutionary algorithms and stochastic approximation procedures, we will eventually be able to treat many interesting theoretical questions on asymptotic properties of evolutionary computation.

Evolutionary algorithms represent a class of stochastic optimization algorithms in which organic evolution is regarded as a set of rules for optimization. These algorithms have been applied to many problems in parameter optimization and related fields with great success. With simplification of biological reality, based on the collective learning process within a population of individuals, each of which is a search point of potential solutions for a given problem, the evolutionary algorithms carry out the designed computational task using randomized process of selection, mutation and recombination. The study of the evolutionary algorithms has witnessed rapid progress for nearly thirty years. For some of the important contributions, we mention the work of Rechenberg [21], Schwefel [25], [26], [27], Holland [11], De Jong [6], Fogel [8], Fogel [9] among others. For an extensive review of the recent advances, the readers are referred to Bäck and Schwefel [1], Bäck, Rudolph and Schwefel [2], [29] and the references therein.

The method of stochastic approximation was initiated in the early 50's to find the root of a function $f(\cdot)$ and/or to locate the maxima or minima of $f(\cdot)$, provided only noisy measurements or observations are available. Owing to its wide range of applicability, such algorithms have been studied extensively for years. We now have good understanding on the asymptotic behavior of the algorithms (see [19], [18], [14], [12] and [4] and the references therein). Early development via martingale approach is contained in Nevelson and Khasminskii [19]; the celebrated ODE (ordinary differential equation methods) are discussed in Ljung [18] and Kushner and Clark [14]; the method of weak convergence is due to Kushner and his associates and documented in [12]; a most recent book on stochastic approximation is the one by Benveniste, Métivier and Priouret [4]. It provides a comprehensive overview on the recent development of the subject and interesting applications in control and adaptive signal processing.

Both evolutionary algorithms (EA) and stochastic approximation are aiming at the objective–stochastic optimization. Nevertheless, surprisingly enough, until now, there has not been any attempt to connect these closely related fields, to the best of our knowledge. Taking this into account, our main effort in this paper is to apply some of the techniques in stochastic approximation to analyze the asymptotic properties of some recursive algorithms that have potential applications in evolutionary computation. We will make effort to establish the connection of these methods. We believe that the ideas to be presented below will be of interest to the EA community as well as to people working in the systems theory and related fields. By and large, the current work is served as a survey on convergence and rate of convergence issues.

The rest of the paper is arranged as follows. The precise formulation of the problem together with examples from evolutionary computation are given next. Both constant step size algorithms and decreasing step size schemes are given. Although not all the mathematical details are provided, appropriate references are given. Section 3 presents the convergence results and Section 4 focuses on the rate of convergence issues. In these sections, we will also state some of the mathematical background. Finally we close this paper with some concluding remarks in Section 5.

2 Problem formulation

We present the problem formulation in a rather general form so as to accommodate many potential applications in evolutionary computation.

Let x, $\xi \in \mathbb{R}^r$, $G(\cdot,\cdot) : \mathbb{R}^{r\times r} \mapsto \mathbb{R}^r$, where $G(x,\xi)$ denotes the noisy gradient estimate of a real-valued function $f(x)$. Our effort is to develop recursive algorithms to carry out the optimization task. Suppose the initial estimate x_0 is selected. We then generate a sequence of estimates $\{x_n\}$ by means of the following recursion:

$$x_{n+1} = x_n - a_n G(x_n, \xi_n), \tag{1}$$

or

$$x_{n+1} = x_n - aG(x_n, \xi_n), \tag{2}$$

where a_n and a are known as step size or gain sequences. In (1), we assume that

$$a_n > 0,\ a_n \xrightarrow{n} 0 \text{ and } \sum_{n=1}^{\infty} a_n = \infty,$$

whereas in (2), a is a constant step size. In the asymptotic analysis, however, we assume that $a \to 0$. To see the connection of the above algorithms with the evolutionary computation, we consider the following example.

Example 2.1. Suppose that we are employing a $(1,\lambda)$ strategy to solve an optimization problem. Select random vectors $z_n^{(i)}$, $1 \le i \le \lambda$. We then use the current estimate x_n to evaluate $f(x_n + z_n^{(i)})$, for $1 \le i \le \lambda$. After the evaluation, compare the corresponding values and select the vector $x_n + z_n^{(j)}$ such that $f(x_n + z_n^{(j)}) = \min f(x_n + z_n^{(i)})$, for $1 \le i \le \lambda$. In short,

$$x_{n+1} = \text{argmin}\{f(x_n + z_n^{(1)}), \ldots, f(x_n + z_n^{(\lambda)})\}. \tag{3}$$

Clearly, the algorithm can be thought of as a recursive procedure.

Comparing (3) with that of (1) or (2), a first glance may lead to the conclusion that they do not have much in common. At least (3) does not involve step sizes. However, a closer examination reveals that there is a hidden step size in the algorithm. Suppose that $\{z_n^{(i)}\}$ is a sequence of independent and normally distributed random variables such that the mean of $z_n^{(i)}$ is zero and the covariance matrix is $\sigma_n^2 I$, where $\sigma_n^2 > 0$ (σ_n^2 can be either varying with n or equal to a constant σ). We note that σ_n^2 here is simply the scale factor of the distribution. Then we can rewrite $z_n^{(i)}$ as $z_n^{(i)} = \sigma_n \tilde{z}_n^{(i)}$. Now $\tilde{z}_n^{(i)}$ has a normal distribution with mean 0 and covariance I, the identity matrix. Thus (3) can further be written as:

$$x_{n+1} = x_n + \sigma_n \sum_{i=1}^{\lambda} \tilde{z}_n^{(i)} I_{\{f(x_n + z_n^{(i)}) = \min_{u \in \Lambda_n} f(u)\}},$$

where $\Lambda_n = \{x_n + z_n^{(i)};\ i = 1, \ldots, \lambda\}$, and I_A denotes the indicator function of the set A. In evolution strategy, one often chooses σ_n that is proportional to $\sim \frac{1}{r} H(\nabla f(x_n))$, where r is the dimension of the problem and $H(\cdot) : \mathbb{R}^r \mapsto \mathbb{R}$ is an appropriate real-valued function such that $H(0) = 0$ and the only root of $H(\cdot)$ is 0. For example, one may choose $H(\nabla f(x)) = |\nabla f(x)|$. With either a or a_n denoting the proportional constant (multiplied by $1/r$), the recursive formula can be written as

$$x_{n+1} = x_n + aH(\nabla f(x_n)) \sum_{i=1}^{\lambda} z_n^{(i)} I_{\{f(x_n + z_n^{(i)}) = \min_{u \in \Lambda_n} f(u)\}}.$$

or

$$x_{n+1} = x_n + a_n H(\nabla f(x_n)) \sum_{i=1}^{\lambda} z_n^{(i)} I_{\{f(x_n + z_n^{(i)}) = \min_{u \in \Lambda_n} f(u)\}}.$$

In the next section, we argue that for fixed x, the average of the random part in the iteration is not equal to zero. Algorithm (3) now becomes a constant step size or a decreasing step size recursive algorithm of stochastic approximation type. The constant a or a_n is the step size of the corresponding stochastic approximation algorithm.

To proceed, a word about the notation is in order. In the sequel, K denotes a generic positive constant. Its value may be different for different appearances. Thus, $K + K = K$, $KK = K$ are understood in an appropriate sense. z' denotes the transpose of z and f_x denotes the first partial derivatives of the function with respect to x. Similar notation is used for the second order derivatives.

3 Convergence of the algorithms

In this section, we study the convergence of the algorithms (1) and (2). We include the with probability one convergence and that of weak convergence in two subsections. In the third subsection, we discuss related problems in EA computation.

3.1 W.p.1 convergence

In general, dealing with discrete iterations is very hard and requires much more restrictive conditions. In the late 70's, an approach known as ODE (ordinary differential equation) methods was invented by Ljung [18] and further developed by Kushner and his colleagues [14]. The essence is that in lieu of examining the discrete iterations directly, one takes the continuous time interpolation of the estimate. Then combining the theory of analysis and probability, one shows that a suitably scaled sequence of functions is uniformly bounded and equicontinuous. Thus one may extract convergent subsequence in accordance with the Ascoli-Arzela's lemma, and identify the limit of the sequence as a solution of an ordinary differential equation. To give some heuristic argument, consider a special case, $G(x,\xi) = \nabla f(x) + \xi$, the additive noise setting. For large n, we expect the noise term to be averaged out owing to the law of large numbers type of conditions. Thus,

$$x_{n+k} \approx x_k - \sum_{i=k}^{n+k-1} a_i \nabla f(x_k),$$

or equivalently,

$$\frac{x_{n+k} - x_k}{\sum_{i=k}^{n+k-1} a_i} \approx \nabla f(x_k),$$

which leads to the connection to the ODE $\dot{x} = -\nabla f(x)$.

To proceed, we work with the decreasing step size algorithm, and define $t_n = \sum_{i=0}^{n-1} a_i$ and $m(t) = \max\{n; t_n \leq t\}$. Define the piecewise linear (denoted by $x^0(t)$) and piecewise constant (denoted by $\bar{x}(t)$) interpolations of the iterates as:

$$\begin{aligned} & x^0(t_n) = x_n, \\ & x^0(t) = \frac{t_{n+1} - t}{a_n} x_n + \frac{t - t_n}{a_n} x_{n+1} \text{ in } (t_n, t_{n+1}), \\ & \bar{x}(t) = x_n \text{ for } t \in [t_n, t_{n+1}). \end{aligned} \tag{4}$$

We also define a shifted sequence $x^n(\cdot)$ by $x^n(t) = x^0(t + t_n)$.

Now, we are in a position to give a set of conditions that yields the w.p.1 convergence of the algorithms.

(A.3.1) $\sum_n a_n^2 < \infty$, $\sum_n a_n = \infty$, $\{a_{n+1}/a_n\}$ is bounded. $G(x,\xi_n) = G_1(x,\alpha_n) + G_2(x)\beta_n$ such that $G_1(x,\xi)$ is bounded on bounded x-set and is continuous. $G_2(\cdot)$ is a continuous and bounded function. $\{\alpha_n\}$ is a sequence of uniformly bounded random variables and $\{\beta_n\}$ is a sequence of independent random variables with 0 mean and finite second moment.

(A.3.2) There is a twice continuously differentiable Liapunov function $0 \leq V(x)$ such that $V_{xx}(\cdot)$ is bounded, $V(x) \to \infty$ as $|x| \to \infty$. Let E_n denote the conditional expectation on the σ-algebra $\mathcal{F}_n$ generated by $\{x_0, \xi_i;\ i \leq n\}$. W.p.1,

$$\left|\sum_{i=n}^{\infty} a_i V_x'(x) E_n(G(x,\xi_i) - \nabla f(x))\right| \leq K a_n (1 + |V_x'(x)\nabla f(x)|),$$
$$\left|\sum_{i=n}^{\infty} a_i [V_x'(x) E_n(G_1(x,\alpha_i) - \nabla f(x)]_x\right| \leq K a_n (1 + |V_x'(x)\nabla f(x)|^{1/2}).$$

The bounds above also hold with $V(\cdot)$ replaced by a twice continuously differentiable function with a compact support. For some $\eta > 0$, some $\lambda_0 > 0$ and compact set $Q_0 = \{x;\ V(x) \leq \lambda_0\}$, $V_x'(x)\nabla f(x) > \eta$ for all $x \notin Q_0$.

(A.3.3) $E_n|G(x,\xi_n)|^2 \leq K(1 + |V_x'(x)\nabla f(x)|) \leq K(1 + V(x))$. For $0 \leq s \leq 1$, $E_n|V_x'(x + s a_n G(x,\xi_n))\nabla f(x + s a_n G(x,\xi_n))| \leq K(1 + |V'(x)\nabla f(x)|)$.

Theorem 3.1. *Suppose the conditions (A.3.1)–(A.3.3) are satisfied. Then $\{x_n\}$ is bounded w.p.1. If $-V_x'(x)\nabla f(x) \leq 0$ for all x, then $x_n \to \{x; V_x'(x)\nabla f(x) = 0\}$ w.p.1. In general $\{x_n\}$ converges to the largest bounded invariant set of*

$$\dot{x} = -\nabla f(x), \quad x(0) = x_0. \tag{5}$$

If x^0 is an asymptotically stable solution of (5) with domain of attraction $DA(x^0)$ and if $x \in A \subset DA(x^0)$ infinitely often, where A is a compact set, then $x_n \xrightarrow{n} x^0$ w.p.1.

The proof of this theorem uses the idea of perturbed Liapunov function methods (see [12]). The argument is analogues to [13]. Rather than going through all the technical details, we consider a simpler problem–the approximation scheme with additive structure.

3.1.1 Discussion on a simpler problem.

Consider the following simplified problem:

$$x_{n+1} = x_n - a_n(\nabla f(x_n) + \xi_n).$$

Define the interpolations as before, and define also

$$B^0(t_n) = \sum_{i=0}^{n-1} a_i \xi_i$$
$$B^0(t) = \frac{t_{n+1} - t}{a_n} B^0(t_n) + \frac{t - t_n}{a_n} B^0(t_{n+1}) \text{ in } (t_n, t_{n+1}).$$

Assume that:

- $\nabla f(\cdot)$ is a continuous function.
- $\lim_n P\left(\sup_{m\geq n}\left|\sum_{i=n}^{m} a_i\xi_i\right| \geq \varepsilon\right) = 0$ for each $\varepsilon > 0$ or simply $\sum_{i=1}^{n} a_i\xi_i$ converges w.p.1.
- $\{x_n\}$ is bounded w.p.1.
- There is a twice continuously differentiable Liapunov function $V(\cdot)$ such that

$$V_x'(x)\nabla f(x) > 0 \text{ for all } x \notin S = \{x;\ \nabla f(x) = 0\}.$$

Then $x_n \to S$ w.p.1, i.e., $\lim_n \rho(x_n, S) = 0$ w.p.1, where ρ denotes the usual distance function. In particular, if $S = \{x^*\}$ a singleton set, then $x_n \xrightarrow{n} x^*$ w.p.1.

The proof of the assertion goes as follows. By means of the boundedness of $\{x_n\}$, it can be verified that the sequence $\{x^n(\cdot)\}$ is uniformly bounded and equicontinuous. By virtue of Ascoli-Arzela's lemma, we can extract convergent subsequences. Select such a sequence but still denote the index by n. Using the recursive formulae, it is not difficult to see that

$$x^n = x^n(0) - \int_0^t \nabla f(x^n(s))ds - B^n(t) + e^n(t),$$

where $e^n(t) \xrightarrow{n} 0$ uniformly on finite time intervals. In addition, by virtue of the averaging condition on the noise sequence, $B^n(t)$ also goes to 0. As $n \to \infty$, the limit of the equation above gives us

$$x = x(0) - \int_0^t \nabla f(x(s))ds,$$

which is the desired equation. Finally the assertion follows from the LaSalle's invariance principal and some detailed probabilistic argument (see e.g., [14] Chapter 2).

Remark: The boundedness of $\{x_n\}$ above can be obtained via the use of perturbed Liapunov function methods. We assumed it for simplicity. The average condition of the noise or the summability of $\sum_i a_i\xi_i$ is a rather general condition. It is verified by a large class of random processes. For example, i.i.d. noise, martingale difference sequences, some ARMA models, mixing processes etc. can be shown to possess such properties (see Kushner and Clark [14], Yin [32] and the references therein). The conditions used here (even in the setting of Theorem 3.1) are not the most general one. Weaker conditions are possible. For ease of discussion, we selected the simple forms.

The significance of the limiting ODE (5) is that the stationary points of it corresponds to the stationary points we are searching for. The ODE method gives us an analytic way to convert the problem into one that can be relatively easily handled.

3.2 Weak convergence

First we recall the definition of weak convergence. A sequence of random variables $\{w_n\}$ is said to converge to w weakly, if for any bounded and continuous function $g(\cdot)$, $Eg(w_n) \to Eg(w)$ as $n \to \infty$. Weak convergence is a substantial generalization of the concept of convergence in distribution. It can be used not only for random variables living in a Euclidean space, but also for random processes taking values in function spaces as well. In the process of getting weak convergence result, one often needs to verify that the sequence involved is tight. A sequence $\{w_n\}$ is tight, if for any $\varepsilon > 0$, there is a compact set S_ε, such that $P(w_n \notin S_\varepsilon) \leq \varepsilon$ for all n. A well-known theorem due to Prokhorov states that, in a complete separable metric space, the tightness is equivalent to sequential compactness. In other words, once the tightness is verified, one may proceed to extract convergent subsequences.

There are reasons that weak convergence analysis is more preferable in many applications. First, it requires much less restrictive conditions than its with probability one convergence counter part. Secondly, dealing with the problem of rates of convergence, we often need to obtain results similar to that of the central limit theorem. In this regard, one is forced to treat convergence in the sense of convergence in distribution or convergence in the weak sense any way. Third, to analyze a constant step size algorithm, we need to use weak convergence tools since if a constant step size is used, almost sure (w.p.1) convergence results cannot generally be expected.

For technical purposes, it is easier to deal with paths than with measures. A device known as Skorokhod representation allows one to 'change' the weak convergence to w.p.1 convergence on a larger space. For the detailed account on the concept of weak convergence as well as many related materials, we refer the reader to the book of Ethier and Kurtz [7] and the references therein.

In our weak convergence analysis to follow, we often work with $D^r[0,\infty)$, the space of functions, that are right continuous, have left-hand limit endowed with some weak topology (Skorokhod topology). Our analysis requires that first the tightness is verified and then the limit process is characterized.

In what follows we provide some sufficient conditions that ensure the convergence in the sense of weak convergence. We work with the algorithm with constant step size a. The argument for that of the decreasing step size algorithms are virtually the same.

(B.3.1) The function $G(x,\xi)$ is bounded on bounded x-set,

$$\lim_{|x-y|\to 0} E|G(x,\xi_j) - G(y,\xi_j)| = 0,$$

and for each x belongs to a bounded set and each $T < \infty$, $\{|G(x,\xi_n)|;\ \ na \leq T\}$ is uniformly integrable.

(B.3.2) The following averaging condition holds: For each x,

$$\frac{1}{n}\sum_{i=m}^{m+n} E_m G(x,\xi_i) \xrightarrow{n} \nabla f(x) \text{ in probability.} \tag{6}$$

Remark: As can be seen that the conditions for the weak convergence are much weaker than that of the corresponding one for convergence in the sense of w.p.1. We do not even require that the function $G(\cdot)$ to be continuous. Only continuity in the weak sense is assumed. As far as the averaging condition is concerned, it is a law of large number type of condition. We only require the averaging take place in the sense of convergence in probability. Note that the condition is weaker with the conditional expectation added. In case of independent identically distributed and/or martingale difference type of noise ξ_n, it is averaged out even before taking the limit. We emphasize that the noise is averaged out in (6) while x is kept fixed. In fact, this is one of the main ingredients of the direct averaging procedures (see [12]). Keep in mind that we only average out the noise. The uniform integrability condition is verifiable for many applications. See for example Rudolph [22] on verification of the condition for problems in evolutionary computation. To analyze the algorithm, we take the piecewise constant interpolation defined by $x^a(t) = x_n$ for $t \in [na, na+a)$.

Clearly $x^a(\cdot)$ is in the $D^r[0,\infty)$. Now, we proceed to state the weak convergence theorem for the interpolated process.

Theorem 3.2. *Under the conditions of (B.3.1) and (B.3.2). Assume that there is a unique solution of (5) for each initial condition x_0, and $x_0^a \Rightarrow x_0$. Then the sequence $\{x^a(\cdot)\}$ is tight in $D^r[0,\infty)$ such that any weakly convergent subsequence has a limit $x(\cdot)$ that is a solution of the differential equation (5).*

Remark: Very often $x_0^a \equiv x_0$, i.e., it does not depend on the small parameter a. Here we are using a condition that is more general and can accommodate more complex situations.

Idea of proof: We divide the proof into several steps. First we need to show that the sequence $\{x^a(\cdot)\}$ is tight. We add a condition that the iterates x_n are bounded initially, and discuss how we can discard it afterward. It is easily seen that in this case

$$\lim_{A\to\infty}\limsup_{a} P\left\{\sup_{t\le T}|x^a(t)| \ge A\right\} = 0 \text{ for each } T<\infty. \tag{7}$$

Now by virtue of (B.3.2), $\{G(x_n,\xi_n)\}$ is uniformly integrable. Then Lemma 3.7 in Chapter 3 of [12] implies that $\{x^a(\cdot)\}$ is tight, and all limits have continuous paths with probability one.

Without the boundedness condition on the iterates $\{x_n\}$, we proceed by employing a technical device known as N-truncation (see [12] Page 43). For each $N < \infty$, define $S_N =$

$\{x; |x| \leq N\}$. $x^{a,N}(t)$ is said to be an N-truncation of $x^a(t)$ if $x^{a,N}(t) = x^a(t)$ up until first exit from S_N, and

$$\lim_{A\to\infty} \limsup_{a} P\left\{\sup_{t\leq T} |x^{a,N}(t)| \geq A\right\} = 0 \text{ for each } T < \infty. \tag{8}$$

In addition, the truncation for the discrete algorithm is defined as

$$x_{n+1}^N = x_n^N - aG(x_n^N, \xi_n)q_N(x_n^N),$$

where $q_N(\cdot)$ is known as a truncation function taking the form

$$q_N(x) = \begin{cases} 1, & x \in S_N; \\ 0, & x \in \mathbb{R}^r - S_N; \\ \text{smooth}, & \text{otherwise.} \end{cases}$$

We then proceed to obtain the tightness of the truncated process $\{x^{a,N}(\cdot)\}$, obtain its limit, and get the desired result by taking limit as $N \to \infty$ at the end. The details are omitted. We remark that without the boundedness, the verification of (7) is normally difficult, but the verification of (8) for the truncated process is relatively simpler.

In the second step, we characterize the limit process. In the traditional approach of weak convergence analysis, after proving the tightness, one needs to identify the limit process and also show that the finite dimensional distributions of the interpolated process converge. Such an approach is simplified by the direct averaging methods developed by Kushner (see [12] and the references therein). The direct averaging requires to characterize the limit process only by use of the martingale problem formulation of Stroock and Varadhan (see [7]). A process $x(\cdot)$ is said to be a solution of a martingale problem if for any function $g(\cdot)$, that is twice continuously differentiable with compact support,

$$g(x(t)) - g(x(0)) - \int_0^t \mathcal{L}g(x(s))ds$$

is a martingale, where $\mathcal{L}$ is an elliptic operator of the form

$$\mathcal{L} = \sum_i b^i(x)\partial/\partial x^i + (1/2)\sum_{i,j} a^{ij}(x)\partial/\partial x^i \partial x^j$$

corresponding to the stochastic differential equation $dx = b(x)dt + \sigma(x)dw(t)$ such that $\sigma(x)\sigma'(x) = a(x)$.

For ease of presentation, in what follows, we will not use the function $g(\cdot)$ in our analysis. Carrying it in the discussion makes no essential changes.

We extract a convergent subsequence and without change of notation still denote the sequence by $\{x^a(\cdot)\}$, and denote the limit by $x(\cdot)$. By virtue of the Skorokhod representation, (without changing notations), it may be assumed that $x^a(\cdot)$ converges to $x(\cdot)$ w.p.1 and the convergence is uniform on any bounded time interval.

We claim that $x(\cdot)$ is a solution of (5) or what is equivalent that $x(\cdot)$ is a solution of the martingale problem with an degenerate operator (that is the part corresponding to the Brownian motion term disappears or equivalently, $a(x) = 0$). Define

$$M(t) = x(t) - x(0) - \int_0^t (-\nabla f(x(u))du.$$

To prove this assertion, we need only show that $M(\cdot)$ is a continuous martingale. Since it can be verified that $M(\cdot)$ is Lipschitz continuous, it then follows from [12], $M(t)$ =constant. However, $M(0) = 0$. Therefore, $M(t) \equiv 0$. As a result, $x(\cdot)$ is a solution of the equation (5) as claimed.

To verify the martingale property, we need only prove that for any bounded and continuous function $h(\cdot)$, any integer k, $j \leq k$ and $t_j \leq t < t + s$,

$$Eh(x(t_j), j \leq k)(x(t+s) - x(t)) = -Eh(x(t_j), j \leq k) \int_t^{t+s} \nabla f(x(u))du.$$

To this end, we work with the pre-limit process $x^a(\cdot)$. Choose a sequence of real numbers $\{n_a\}$ such that $n_a \to \infty$ as $a \to 0$, but $\delta_a = an_a \to 0$. Detailed computation leads to

$$\begin{aligned} & Eh(x^a(t_j), j \leq k)(x^a(t+s) - x^2(t)) \\ = \quad & -Eh(x^a(t_j), j \leq k) \sum_{t/a}^{(t+s)/a} aG(x_i, \xi_i) \\ = \quad & -Eh(x^a(t_j), j \leq k) \sum_{ln_a = t/a}^{(t+s)/a} \delta_a \left\{ \frac{1}{n_a} \sum_{i \in L^a} E_{ln_a} G(x_i, \xi_i) \right\}, \end{aligned}$$

where $L^a = \{i; ln_a \leq i \leq ln_a + n_a - 1\}$. Notice that the conditioning is inserted since $t_j \leq t$, $h(x(t_i))$ is $\mathcal{F}_{ln_a}$ measurable.

Loosely, the outer summation in the above formula is replaced by $\int_t^{t+s}$ whereas the term inside the curly bracket gives us the integrand in the limit (in the sense of in probability). To obtain the desired result, it now suffices to consider the term inside the curly bracket. Sending $l\delta_a \to u$, we need only show that

$$\frac{1}{n_a} \sum_{i \in L^a} E_{ln_a} G(x_i, \xi_i) \xrightarrow{a} \nabla f(x(u)) \text{ in probability.}$$

Now by using condition (B.3.1), the limit of

$$\frac{1}{n_a} \sum_{i \in L^a} E_{ln_a} G(x_i, \xi_i)$$

is the same as that of

$$\frac{1}{n_a}\sum_{i\in L^a} E_{ln_a} G(x_{ln_a}, \xi_i).$$

In fact, we can prove that

$$\frac{1}{n_a}\sum_{i\in L^a} E_{ln_a} G(x_i, \xi_i) = \frac{1}{n_a}\sum_{i\in L^a} E_{ln_a} G(x^a(l\delta_a), \xi_i) + o(1),$$

where $o(1)\xrightarrow{a}0$ in probability. Since $l\delta_a \to u$, by the weak convergence of $x^a(\cdot)$ and the Skorokhod representation,

$$\frac{1}{n_a}\sum_{i\in L^a} E_{ln_a} G(x^a(l\delta_a), \xi_i) = \frac{1}{n_a}\sum_{i\in L^a} E_{ln_a} G(x(u), \xi_i) + o(1),$$

where $o(1)\xrightarrow{a}0$ in probability.

Suppose for the moment that $x(u)$ takes finitely many values, e.g., $\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_\kappa$. We then have

$$\begin{aligned}
\frac{1}{n_a}\sum_{i\in L^a} E_{ln_a} G(x(u), \xi_i) &= \sum_{\nu=1}^{\kappa} \frac{1}{n_a}\sum_{i\in L^a} E_{ln_a} G(\hat{x}_\nu, \xi_i) I_{\{x(u)=\hat{x}_\nu\}} \\
\xrightarrow{a} \quad & \sum_{\nu=1}^{\kappa} \nabla f(\hat{x}_\nu) I_{\{\nabla x(u)=\hat{x}_\nu\}} = \nabla f(x(u)) \text{ in probability.}
\end{aligned}$$

as desired. In general, what we need to do is to approximate $x(u)$ by a function that takes only finitely many values, i.e., for any $\varepsilon > 0$, choose $x_\varepsilon(u)$ that takes only finitely many values such that

$$\lim_{a\to 0, ia\to u} E|G(x(u), \xi_i) - G(x_\varepsilon(u), \xi_i)| = 0,$$

and then work out the convergence for the approximation function $G(x_\varepsilon(u), \xi_i)$. Notice that owing to the choice of n_a, when $l\delta_a \to u$, $ia \to u$ for all $i \in L^a$. The details are omitted.

The theorem above gives us a result on arbitrarily large but still bounded time intervals. It is of particular interest to us to see what happens when $a \to 0$ and $n \to \infty$. A result concerning such a problem is recorded in the following theorem.

Theorem 3.3. *Suppose that θ is a stationary point of the equation (5), i.e., $\nabla f(\theta) = 0$, and suppose that θ is globally attracting (in the sense of Liapunov stability). Assume the conditions of Theorem 3.2 are satisfied and $\{x_n, n < \infty, a > 0\}$ is tight in $\mathbb{R}^r$. Let $\{t_a\}$ be such that $t_a \to \infty$ as $a \to 0$. Then $x^a(\cdot + t_a)$ converges weakly to θ.*

The proof of the theorem is very similar to that of Theorem 3.2. Consider the joint pair $(x^a(\cdot + t_a), x^a(\cdot - T + t_a))$ for each $T < \infty$. Extract a convergent subsequence and denote the limit by $(x(\cdot), x_T(\cdot))$. We realize that $x(T) = x_T(0)$. By virtue of the assumption $x_T(0)$ belongs to a set which is tight. We then proceed to use the stability argument to finish up the proof. For more details on this matter, one may wish to see a corresponding theorem in [15]. We point out that the tightness of $\{x_n\}$ can be proved. Since the proof uses the techniques of perturbed Liapunov function methods and is similar to the error bound estimate to be derived in the sequel, we simply assumed this condition holds at this point.

3.3 Discussion on EA related algorithms

Similar limit theorems can be obtained for the example given in the previous section. The convergence theorems hold if $\nabla f(\cdot)$ is replaced by a function of $\nabla f(\cdot)$. For the example discussed in the previous section, the limiting ODE reads as:

$$\dot{x} = -H(\nabla f(x))\bar{v}(x),$$

where $\bar{v}(x) \neq 0$ is a vector (depending on the function form of $f(\cdot)$) resulting from the average of the sequence

$$\tilde{z}_n^{(i)} I_{\{f(x_n+z_n^{(i)})=\min_{u\in\Lambda_n} f(u)\}}.$$

Note that setting the right-hand side to be 0 leads to the equation $H(\nabla f(x))\bar{v}(x) = 0$ or equivalently, $H(\nabla f(x)) = 0$. This in turn implies that $\nabla f(x) = 0$ as desired. The solutions of this gives us the stationary points of the function $f(\cdot)$.

Next we illustrate why $\bar{v}(x) \neq 0$ should hold in many cases. First, let us consider a very simple example. Suppose $f(\cdot)$, the function to be minimized is linear and suppose and $\{z_n^{(i)}\}$ is a sequence of i.i.d. normal random vectors. Suppose also that the components of the random vector are independent, i.e., the covariance matrix is a diagonal matrix. Then essentially, we are dealing with a scalar problem. Let us consider one component of the vector, but suppress the dependence (index) of the vector. Thus, we treat x_n, $z_n^{(i)}$ as scalars. Using elementary statistics, the algorithm is of the form

$$x_{n+1} = x_n + z_{n,\{1\}},$$

where $z_{n,\{1\}}$ denotes the (minimum) order statistics.

Using the decomposition outlined in Example 2.1, with $\lambda \geq 2$,

$$x_{n+1} = x_n + \sigma\tilde{z}_{n,\{1\}}.$$

The density function of $\tilde{z}_{n,\{1\}}$ is given by

$$\hat{f}_{\tilde{z}_{n,\{1\}}}(z) = \lambda\hat{f}(z)(1-\hat{F}(z))^{\lambda-1}dz,$$

where $\hat{f}(z)$ and $\hat{F}(z)$ are the density and distribution functions of a standard normal random variable, respectively.

By virtue of an integration by parts, we have that

$$\begin{aligned} E\tilde{z}_{n,\{1\}} &= \lambda \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} z \exp(-z^2/2)(1-\hat{F}(z))^{\lambda-1} dz \\ &= -\frac{\lambda(\lambda-1)}{2\pi} \int_{-\infty}^{\infty} \exp(-z^2)(1-\hat{F}(z))^{\lambda-2} dz \neq 0. \end{aligned}$$

For, suppose not, i.e., the integral above is 0. Since the integrand is non-negative, the integrand must be equal to 0 identically, which is a contradiction. In fact, the discussion above shows that $E\tilde{z}_{n,\{1\}} < 0$.

This example may seem to be over simplified, but it illustrates the reason that the limit vector $\bar{v}$ is non-zero. In general, the situation becomes more complex, we are effectively dealing with functions of order statistics, but the main idea remains the same.

If the function $f(\cdot)$ is smooth enough, say C^2, then we may wish to take a Taylor expansion. This leads to

$$f(x_n + z_n^{(i)}) = f(x_n) + \sigma f_x'(x_n)\tilde{z}_n^{(i)} + O(\sigma^2)(|\tilde{z}_n^{(i)}|^2),$$

provided if $f_{xx}(\cdot)$ is bounded. When we compare the values of $f(x_n + z_n^{(i)})$ in the $(1,\lambda)$ strategy, we are basically comparing the term $f_x'(x_n)\tilde{z}_n^{(i)} + O(\sigma^2)(|\tilde{z}_n^{(i)}|^2)$. Now for fixed x, we can treat the corresponding order statistics for $1 \leq i \leq \lambda$ (by using the weak convergence theory). As in the linear case, it can be shown that the expectation is non-zero for many practically interesting functions.

Finally, we point out that for the i.i.d. sequence $\{z_n^{(i)}\}$, the average conditions in Section 3.1 and 3.2 hold. The verification can be done readily. This paper deals with a somewhat more general setup. The specific problem related to the convergence of the $(1,\lambda)$ strategy will be studied elsewhere.

4 Rate of convergence

This section is divided into two subsections. The first of them gives an order of magnitude estimate on the estimation error (or an error bound), and the second one derives a local limit result similar in spirit to the well-known central limit theorem or rather functional central limit theorem. We shall concentrate on the constant step size algorithms. As for the decreasing step size procedures, using essentially the same techniques, we get similar results. We mention these results at the end.

4.1 An error bound on $x_n - \theta$

The analysis to follow uses the perturbed Liapunov function methods (see [12] and the references therein). For notational simplicity, we assume $\theta = 0$ henceforth. This is no loss of generality at all since we can always translate the origin as needed.

To proceed, we list the conditions to be used in the sequel.

(A.4.1) There is a Liapunov function $V(\cdot) : \mathbb{R}^r \mapsto \mathbb{R}$ such that the function together with its first and second partial derivatives are continuous. $V(x) \geq 0$ for all x, $V(x) \overset{|x|\to\infty}{\longrightarrow} \infty$, $V_{xx}(\cdot)$ is bounded, and $V_x'(x)\nabla f(x) > \eta V(x)$ for all $x \neq 0$ and for some $\eta > 0$.

(A.4.2) $G(x, \xi_n) = G_1(x, \alpha_n) + G_2(x)\beta_n$ such that $G_1(\cdot)$ is bounded on bounded x-sets, and $G_1(\cdot, \xi)$ is continuous. $G_2(\cdot)$ is a continuous and bounded function. $\{\alpha_n\}$ is a stationary sequence of uniformly bounded random variables satisfying $EG_1(x, \alpha_n) = \nabla f(x)$ for each x, and $\{\beta_n\}$ is a sequence of independent random variables with zero mean and finite second moment. $E_n|G(x, \xi_n)|^2 \leq K(1 + V(x))$.

(A.4.3) The following inequalities hold:

$$\begin{aligned}
&\left|\sum_{i=n}^{\infty} V_x'(x) E_n(G(x, \xi_i) - \nabla f(x))\right| \leq K(1 + V(x)),\\
&\left|\sum_{i=n}^{\infty} [V_x'(x) E_n(G_1(x, \alpha_i) - \nabla f(x)]_x\right| \leq K a_n (1 + V^{1/2}(x)).
\end{aligned}$$

Theorem 4.1. *Under the conditions of (A.4.1)-(A.4.3), for sufficiently large n, (i.e., there is an N_a such that for all $n \geq N_a$),*

$$EV(x_n) = O(a) \textit{ for sufficiently small } a > 0. \tag{9}$$

Since x_n is $\mathcal{F}_n$ measurable, and β_n has mean 0,

$$E_n V_x'(x_n) G_2(x_n) \beta_n = V_x'(x_n) G_2(x_n) E_n \beta_n.$$

By direct computation, we get

$$\begin{aligned}
E_n V(x_{n+1}) - V(x_n) \ &= -a V_x'(x_n) \nabla f(x_n)\\
&\quad - a V_x'(x_n) E_n [G_1(x_n, \alpha_n) - \nabla f(x_n)]\\
&\quad\quad + O(a^2)(1 + V(x_n))\\
&\leq -a\eta V(x_n)\\
&\quad - a V_x'(x_n) E_n [G_1(x_n, \alpha_n) - \nabla f(x_n)]\\
&\quad\quad + O(a^2)(1 + V(x_n)).
\end{aligned} \tag{10}$$

The second term on the right side of the inequality sign is an extraneous term. To obtain the desired result, it needs to be eliminated. To overcome the difficulties, we introduce a perturbation term as

$$V_1(x, n) = -a \sum_{i=n}^{\infty} V_x'(x) E_n [G_1(x, \alpha_i) - \nabla f(x)].$$

By virtue of (A.4.3), $V_1(\cdot)$ is well defined and $|V_1(x,n)| \le Ka(1+V(x))$. In addition,

$$\begin{aligned} &E_n V_1(x_{n+1}, n+1) - V_1(x_n, n) \\ &\qquad = aV_x'(x_n)E_n[G_1(x_n,\xi_n) - \nabla f(x_n)] + O(a^2)(1+V(x_n)). \end{aligned}$$

Now define the perturbed Liapunov function $V(\cdot)$ by

$$V^a(x,n) = V(x) + V_1(x,n).$$

As a result,

$$E_n V^a(x_{n+1}, n+1) - V^a(x_n,n) \le -a\eta V(x_n) + Ka^2(1+V(x_n)).$$

Using the bound on $V_1(\cdot)$, we can show that the above inequality holds with $V(x_n)$ replaced by $V^a(x_n,n)$. For sufficiently small a, $Ka - \eta \le -\eta_0$ for some $0 < \eta_0 < \eta$ with $\eta_0 a < 1$, and hence

$$E_n V^a(x_{n+1}, n+1) \le (1-\eta_0 a)V^a(x_n,n) + Ka^2.$$

Iterating on the above inequality and taking expectation yields

$$\begin{aligned} EV^a(x_{n+1}, n+1) &\le (1-\eta_0 a)^n EV^a(x_0, 0) + K\sum_{i=0}^{n}(1-\eta_0 a)^i a^2 \\ &\le (1-\eta_0 a)^n EV^0(x_0,0) + Ka. \end{aligned}$$

Using the bound on $V_1(\cdot)$ again, we also have

$$EV(x_{n+1}) \le (1-\eta_0 a)^2 EV(x_0) + Ka.$$

Select N_a such that for all $n \ge N_a$, $(1-\eta_0 a)^n \le Ka$. The desired result then follows.
Remark: In fact, even more general conditions can be used. In the assumption on the function $G(\cdot)$, we could put it as

$$G(x,\xi_n) = G_1(x,\alpha_n) + G_2(x)\beta_n + G_3(x) + \gamma_n.$$

In this way, we deal with both additive noise and non-additive noise. Some more details can be found in [34] for instance.

If the Liapunov function is locally quadratic, i.e.,

$$V(x) = x'Qx + o(|x|^2),$$

where Q is a symmetric positive definite matrix, then we obtain

$$\{x_n/\sqrt{a}, \text{ for } n \ge N_a\} \text{ is tight..} \tag{11}$$

If we are dealing with decreasing step size, and if $a_n = 1/n^\gamma$, for $0 < \gamma \le 1$, then under similar conditions, with slight modification of the proof, we obtain that $EV(x_n) = O(n^{-\gamma})$ for sufficiently large n. Corresponding to the remark just made above, we have the tightness of $\{n^{\gamma/2}x_n\}$.

4.2 Asymptotic normality

Theorem 4.1 above exploits the dependence of the iterates on a by giving an upper bound on the estimation error. In this subsection, we shall derive another local limit theorem that is similar to the functional central limit theorem.

The idea is that we linearize the function $G(\cdot)$ around its stable point, and obtain a suitably scaled sequence. Owing to (11), the appropriate scaling here is $\sqrt{a}$. For simplicity, we will treat $G(x, \xi)$ as one term without separating it as $G_1(\cdot)$, $G_2(\cdot)$ etc. Starting with (2), and assuming that $G_x(\cdot, \xi)$ and $G_{xx}(\cdot, \xi)$ exist and are continuous, and $G_{xx}(\cdot, \xi)$ is bounded, we arrive at

$$u_{n+1} = u_n - aG_x(0, \xi_n)u_n - \sqrt{a}G(0, \xi_n) + O(a^{3/2}|u_n|^2). \tag{12}$$

To obtain the asymptotic normality, again, we take a continuous time interpolation as follows. For $n \geq N_a$, define $u^a(\cdot)$ by $u^a(t) = u_n$ for $t \in [a(n - N_a), a(n - N_a + 1))$. As in Section 3.2, $u^a(\cdot)$ lives in $D^r[0, \infty)$. Notice that the last term in (12) is asymptotically negligible, so we discard it henceforth. Suppose that

$$\begin{aligned} &\sum_{i=N_a}^{t/\varepsilon} \sqrt{a}G(0, \xi_i) \Rightarrow w(t) \text{ a Brownian motion with covariance } \Sigma t; \\ &\frac{1}{n_a} \sum_{i \in L^a} E_{ln_a} G_x(0, \xi_i) \xrightarrow{a} f_{xx}(0) \text{ in probability.} \end{aligned} \tag{13}$$

Using the weak convergence methods as described in the previous sections, we can shown that $u^a(\cdot)$ converges weakly to $u(\cdot)$ such that $u(\cdot)$ is a solution of the stochastic differential equation

$$du = -f_{xx}(0)udt + dw. \tag{14}$$

Remark: Eq. (14) has a unique solution for each initial condition since it is linear. The assumption of the convergence to a Brownian motion can be verified in a wide variety of cases. Suppose the noise is a sequence of i.i.d. random variables with 0 mean and finite variance. Then this condition is verified by the well-known result of Donsker's invariance principle (see [7]). It also holds for more general noise structure such as ϕ-mixing type of random processes which allow correlated noise with the correlation diminishing asymptotically. Many forms of sufficient conditions guarantee the existence of the limit can be found in [7] and the references therein. For stochastic approximation related problems see Kushner [12], Yin [32], Yin and Yin [34] among others.

5 Concluding remarks

In this work, we exploited the connection of evolutionary computation and stochastic approximation. As it is explained that both of them have the objective of carrying out stochastic optimization tasks. By studying some appropriate stochastic recursive algorithms, we reviewed some of the recent developments in stochastic approximation. We also investigated the possible applications to evolutionary algorithms. Limit theorems are obtained by taking suitable scaling and continuous time interpolations.

For the problems studied in this paper, we assumed that the noisy gradient estimate is available. If one has to use, for example a finite difference method to get the gradient estimates, then the convergence rate will be slower as is the case for the classical KW procedures. Nevertheless, there are some recent advances to speed up the convergence for the gradient estimates. We refer the readers to Ho and Cao [10] for further details. A survey on the recent progress in this direction in conjunction with stochastic approximation can be found in Kushner and Vázquez-Abad [17].

It should be mentioned that the evolutionary algorithms can deal with non-smooth objective functions. For stochastic approximation, the corresponding part is the use of non-smooth analysis via differential inclusion.

Recently, there are renewed interests in improving the rate of convergence of stochastic approximation type algorithms by utilizing post-averages of the iterates or by taking averages of the iterates as well as the observations (see Bather [3], Kushner and Yang [16], Polyak [20], Ruppert [23], Schwabe [24], Yin [32], Yin and Gupta [33], Yin and Yin [34] and the references therein). It is conceivable that such an attempt will be beneficial for the EA related procedures. In addition to the algorithms considered in this paper, various variants of the recursive algorithms such as projection and other modifications (see Chen and Zhu [5], Kushner and Clark [14], Kusner [12], Kushner and Yin [15], Yin and Zhu [30], Yin [31] and the references therein) can also be studied.

This paper deals with a somewhat more general setup. As was mentioned, our main objective is to see the connection of the EA's and SA's. Although they have many similarities, they also have very distinguished features. In the SA setting, the function under consideration is normally either not known explicitly or the form is very complex. For the EA algorithms, however, the function $f(\cdot)$ under consideration is known, i.e., the computed output of a simulation model. In a subsequent work, we shall treat the $(1, \lambda)$ strategy in detail and obtain the desired asymptotic properties by using the stochastic approximation approach.

At this point, the study is only preliminary in nature with respect to the applications to evolutionary algorithms. Our current effort lies in carrying out in depth study further, and gain a basic understanding of the asymptotic properties of evolutionary algorithms.

References

[1] T. Bäck and H.-P. Schwefel, An overview of evolutionary algorithms for parameter optimization, *Evolutionary Computation* **1** (1993), 1-23.

[2] T. Bäck, G. Rudolph and H.-P. Schwefel, Evolutionary programming and evolution strategies: similarities and differences, in *Proc. The Second Annual Conf. Evolutionary Programming*, 1993, 11-22.

[3] J.A. Bather, Stochastic approximation: A generalization of the Robbins-Monro procedure, in *Proc. Fourth Prague Symp. Asymptotic Statist.*, P. Mandl and M. Hušková Eds., 1989, 13-27.

[4] A. Benveniste, M. Métivier and P. Priouret, *Adaptive Algorithms and Stochastic Approximation*, Springer-Verlag, Berlin, 1990.

[5] H.F. Chen and Y.M. Zhu, Stochastic approximation procedures with randomly varying truncations, *Scientia Sinica* **29** (1986), 914-926.

[6] K. De Jong, An analysis of the behavior of a class of genetic adaptive systems, Doctoral dissertation, U. of Mich., 1975.

[7] S.N. Ethier and T.G. Kurtz, *Markov Processes, Characterization and Convergence*, Wiley, New York, 1986.

[8] L.J. Fogel, A.J. Owens and M.J. Walsh, *Artificial Intelligence Through Simulated Evolution*, J. Wiley, New York, 1966.

[9] D.B., Fogel, An analysis of evolutionary programming, *Proc. First Annual Conf. on Evolutionary Programming*, D.B. Fogel and J.W. Atmar Eds., 43-51, 1992.

[10] Y.C. Ho and X. Cao, *Perturbation Analysis of Discrete Event Dynamic Systems*, Kluwer Academic, Boston, MA, 1991.

[11] J.H. Holland, Outline for a logical theory of adaptive systems, *J. Assoc. Comp. Machinery*, **3** (1962), 297-314.

[12] H.J. Kushner, *Approximation and Weak Convergence Methods for Random Processes, with applications to Stochastic Systems Theory*, MIT Press, Cambridge, MA, 1984.

[13] H.J. Kushner, Stochastic approximation with discontinuous dynamics and state dependent noise; w.p. and weak convergence, *J. Math. Anal. Appl.* **82** (1981), 527-542.

[14] H.J. Kushner and D.S. Clark, *Stochastic Approximation Methods for Constrained and Unconstrained Systems*, Springer-Verlag, 1978.

[15] H.J. Kushner and G. Yin, Asymptotic properties of distributed and communicating stochastic approximation algorithms, *SIAM J. Control Optim.* **25** (1987), 1266-1290.

[16] H.J. Kushner and J. Yang, Stochastic approximation with averaging of the iterates: optimal asymptotic rate of convergence for general processes, *SIAM J. Control Optim.* **31** (1993), 1045-1062.

[17] H.J. Kushner and F.J. Vázquez-Abad, Brown U. LCDS#94-4, 1994.

[18] L. Ljung, Analysis of recursive stochastic algorithms, *IEEE Trans. Automatic Control* **22** (1977), 551-575.

[19] Nevelson and R. Khasminskii, *Stochastic Approximation and Recursive Estimation*, AMS Translation, V. 47, 1973.

[20] B.T. Polyak, New method of stochastic approximation type, *Automat. Remote Control* **51** (1990), 937-946.

[21] I. Rechenberg, *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*, Stuttgart, Frommann-Holzboog, 1973.

[22] G. Rudolph, Convergence of Non-Elist strategies, in *Proc. The First IEEE Conf. Evolutionary Comp.*, Vol. I, 63-66, 1994.

[23] D. Ruppert, Stochastic approximation, in *Handbook in Sequential Analysis*, B.K. Ghosh and P.K. Sen Eds., Marcel Dekker, 503-529, New York, 1991.

[24] R. Schwabe, Stability results for smoothed stochastic approximation procedures, Fachbereich Mathematik, Series A, Mathematik, Preprint Nr. A-92-14, Freie Universität Berlin, 1992.

[25] H.-P. Schwefel, *Kybernetische Evolution als Strategie der experimentellen Forschung in der Strömungstechnik*, Diploma thesis, Technical University of Berlin, 1965.

[26] H.-P. Schwefel, Binäre optimierung durch somatische Mutation, Technical report, Technical University of Berlin and Midical University of Hannover, 1975.

[27] H.-P. Schwefel, *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie*, Vol. 26 of Interdisciplinary systems research, Basel, Birkhäuser, 1977.

[28] H.-P. Schwefel, *Numerical Optimization of Computer Models*, Wiley, Chichester, 1981.

[29] H.-P. Schwefel and R. Männer Eds., *Parallel Problem Solving from nature*, Lect. Notes. in Computer Sci. Vol. 496, Springer, 1991.

[30] G. Yin and Y.M. Zhu, On H-valued Robbins-Monro processes, *J. Multivariate Anal.* **34** (1990), 116-140.

[31] G. Yin, A stopping rule for the Robbins-Monro method, *J. Optim. Theory Appl.* **67** (1990), 151-173.

[32] G. Yin, On extensions of Polyak's averaging approach to stochastic approximation, *Stochastics Stochastic Rep.* **36** (1991), 245-264.

[33] G. Yin and I. Gupta, On a continuous time stochastic approximation problem, *Acta Appl. Math.* **33** (1993), 3-20.

[34] G. Yin and K. Yin, Asymptotically optimal rate of convergence of smoothed stochastic recursive algorithms, *Stochastics Stochastic Rep.* **47** (1994), 21-46.