
Analyzing the $(1, \lambda)$ Evolution Strategy via Stochastic Approximation Methods

G. Yin*

Department of Mathematics
Wayne State University
Detroit, MI 48202

G. Rudolph†

Informatik Centrum Dortmund
Joseph-von-Fraunhofer-Str. 20
D-44227 Dortmund, Germany

H.-P. Schwefel

Universität Dortmund
FB Informatik 11
D-44221 Dortmund, Germany

Abstract

The main objective of this paper is to analyze the $(1, \lambda)$ evolution strategy by use of stochastic approximation methods. Both constant and decreasing step size algorithms are studied. Convergence and estimation error bounds for the $(1, \lambda)$ evolution strategy are developed. First the algorithm is converted to a recursively defined scheme of stochastic approximation type. Then the analysis is carried out by using the analytic tools from stochastic approximation. In lieu of examining the discrete iterates, suitably scaled sequences are defined. These interpolated sequences are then studied in detail. It is shown that the limits of the sequences have natural connections to certain continuous time dynamical systems.

Keywords

Evolutionary computation, Evolution strategy, stochastic approximation, convergence, rate of convergence.

1. Introduction

In this work, we develop asymptotic properties—in particular, convergence and upper bounds of estimation errors for the $(1, \lambda)$ evolution strategy. Both constant and decreasing step size algorithms are considered. We treat the problem as a recursive algorithm of stochastic approximation type and use the methods in stochastic approximation to analyze the algorithms and obtain weak convergence and with-probability-1 (w. p. 1) convergence results.

Based on collective processes with a population of individuals, which are search points for a given problem, the evolutionary algorithms carry out desired computing tasks through the use of randomized selection, mutation, and recombination. These algorithms have been applied to many problems in parameter optimization and related fields with great success. Significant progress has been made in the study of evolutionary algorithms for almost 30 years. Many interesting and useful results have been obtained. To mention just a few, we cite

* The research of this author was supported in part by the National Science Foundation under grant DMS-9224372, and in part by the Deutscher Akademischer Austauschdienst (DAAD).

† The research of this author was supported by BMBF under grant 01 IB 403A, project EVOALG.

the work of Rechenberg (1973), Schwefel (1965), Fogel, Owens, and Walsh (1966), Fogel (1992), Holland (1962), and De Jong (1975). For an extensive review of recent advances, the reader is referred to Bäck and Schwefel (1993), Bäck, Rudolph, and Schwefel (1993), Schwefel (1993), and the references in these papers.

The evolution strategies were first developed by Rechenberg and Schwefel in the mid-60s (Rechenberg, 1965; Schwefel, 1965). At that time, applications in hydrodynamics, such as optimizing the shape of a bent pipe and a flashing nozzle, were dealt with. Different versions of the strategy were simulated (Schwefel, 1965). The research in this subject has grown rapidly ever since. Nowadays, the (μ, λ) evolution strategies, introduced in Schwefel (1977) and Schwefel (1981) are state-of-the-art in evolution strategy research.

When the evolutionary algorithms (EAs) are examined closely, there appears to be a natural connection between EAs and stochastic approximation. However, such a connection has not been explored until very recently (Yin, Rudolph, & Schwefel, 1995). In Yin, Rudolph, and Schwefel (1995), the authors dealt with the “connection” question in a general setting, whereas in the current paper, we aim to derive asymptotic properties for a class of evolutionary algorithms.

Such a study is important. First, it will enhance our understanding of the intrinsic properties of the $(1, \lambda)$ strategy, which in turn will lead to further improvement of computation procedures. In addition, by formulating the problem as a stochastic approximation algorithm, many analytical tools can be employed to carry out the theoretical investigation.

In Yin, Rudolph, and Schwefel (1995), the hidden step size of the $(1, \lambda)$ strategy was discussed in an example. Here we take a closer look, and try to understand the basic properties of the scale parameter in the randomized sequence. It should be pointed out that EAs and stochastic approximation do have distinct features. The objective function $f(\cdot)$ under consideration in EAs is available through simulation, whereas the corresponding counterpart in stochastic approximation is observable and available only in the form of noisy measurements. Nevertheless, such a difference should not prevent us from employing stochastic approximation methods to analyze EA procedures.

Our plan is as follows: We formulate the problem and then convert it into a stochastic approximation algorithm in the next section. Section 3 deals with weak convergence issues. We show how the discrete iteration is related to a continuous time dynamical system. By taking appropriate interpolation, it is shown that an interpolated process converges weakly to a solution of an ordinary differential equation. Then, we proceed with obtaining upper bounds for the estimation errors in Section 4. This step is carried out via the use of the Liapunov function approach and stability of the dynamic system. Utilizing the upper bound as a bridge, we seek further development on a suitably scaled sequence for small a and large n . In the process of getting the asymptotic properties for the constant step size algorithms, our main technique is the weak convergence methods developed by Kushner (see Kushner 1984 and the references therein). Section 5 contains the analysis of algorithms with decreasing step size. W.p.1 convergence is derived by means of the ordinary differential equation (ODE) method. Some concluding remarks are issued in Section 6. Finally, an appendix containing the proof of a lemma is provided.

2. Problem Setup

2.1 $(1, \lambda)$ Evolution Strategy

We wish to minimize a function $f: \mathbb{R}^d \mapsto \mathbb{R}$. The plan is to employ the $(1, \lambda)$ evolution strategy for $\lambda \geq 2$. Loosely, the strategy can be described as follows: In each generation,

one parent produces λ offspring. Among the offspring, choose the best one (with respect to the evaluation of the objective function) to form the next estimate.

To be more specific, generate sequences of random vectors $\{z_n(i)\}$, for $1 \leq i \leq \lambda$, that are independent and identically distributed (i.i.d.) normal random variables with mean zero and covariance $\sigma^2 I_d$, where I_d denotes the $d \times d$ identity matrix such that for each n , $z_n(1), \dots, z_n(\lambda)$ are independent. To carry out the minimization task, choose an initial estimate $x_0 \in \mathbb{R}^d$. At iteration n , add the random vector $z_n(i)$ to the current content; that is, $x_n + z_n(i)$, for $i = 1, \dots, \lambda$. We evaluate the corresponding values $f(x_n + z_n(i))$. Next, choose the smallest among the λ values of $f(\cdot)$; that is,

$$f(x_n + z_n(j)) = \min_{\mu \in \Lambda_n} f(x_n + \mu)$$

where

$$\Lambda_n = \{z_n(i), i = 1, \dots, \lambda\} \tag{1}$$

Then assign $x_n + z_n(j)$ to x_{n+1} . In short

$$x_{n+1} = \operatorname{argmin}\{f(x_n + z_n(1)), \dots, f(x_n + z_n(\lambda))\} \tag{2}$$

This problem was studied in Rudolph (1994) by using the supermartingale convergence theorem developed by Doob (1967). Here we take a different approach. Our task now is to convert Equation 2 to a recursive algorithm of stochastic approximation type so that the techniques of analyzing the stochastic approximation type of algorithms can be applied.

REMARK: In Equation 1 above, without loss of generality, we have assumed that there is only one j satisfying Equation 1. If there are more indices than one satisfying Equation 1, choose j to be the smallest one among them; that is,

$$j = \min\{1 \leq l \leq \lambda; f(x_n + z_n(l)) = \min_{\mu \in \Lambda_n} f(x_n + \mu)\}$$

2.2 A Recursive Algorithm

It is well known that the standard deviation σ is a scale factor in the problem. Since $z_n(i)$ are i.i.d. random vectors and $z_n(i) \sim N(0, \sigma I_d)$, we can rescale the sequence $z_n(i)$ or, equivalently, define another sequence $\{\tilde{z}_n(i)\}$ by setting $z_n(i) = \sigma \tilde{z}_n(i)$ such that $\tilde{z}_n(i) \sim N(0, I_d)$. That is, $\tilde{z}_n(i)$ follows the standard normal distribution. Now Equation 2 can be rewritten as

$$x_{n+1} = x_n + \sigma \sum_{i=1}^{\lambda} \tilde{z}_n(i) I_{\{f(x_n + z_n(i)) = \min_{\mu \in \Lambda_n} f(x_n + \mu)\}} \tag{3}$$

where I is an indicator function.

Again, in the equation above, we have assumed that there is only one i satisfying Equation 1. For the case of multiple indices leading to the minimal, choose i to be the smallest of them and as a result

$$x_{n+1} = x_n + \sigma \sum_{i=1}^{\lambda} \tilde{z}_n(i) I_{\{f(x_n + z_n(i)) = \min_{\mu \in \Lambda_n} f(x_n + \mu); i = \min_{1 \leq l \leq \lambda} f(x_n + z_n(l)) = \min_{\mu \in \Lambda_n} f(x_n + \mu)\}}$$

For reasons of notational simplicity, we will use Equation 3 throughout. As mentioned before, there is no loss in generality in doing so.

In evolution strategy, one often chooses σ so that it is proportional to $(1/d)H(f_x(x_n))$, where $f_x(\cdot)$ denotes the gradient of $f(\cdot)$, d is the dimension of the problem, and $H(\cdot): \mathbb{R}^d \mapsto [0, \infty)$ is an appropriate real-valued function such that $H(0) = 0$ and the only root of $H(\cdot)$ is 0. With a denoting the proportional constant multiplied by $1/d$, the recursive formula can be written as

$$x_{n+1} = x_n + aH(f_x(x_n)) \sum_{i=1}^{\lambda} z_n(i) I_{\{f(x_n+z_n(i))=\min_{\mu \in \Lambda_n} f(x_n+\mu)\}} \tag{4}$$

Equation 4, in fact, is a constant step size stochastic approximation algorithm with step size a . Since normally the problems we treat are many-dimensional ones, a is relatively small. Our interest lies in obtaining convergence and rate of convergence results for the limit case $a \rightarrow 0$. We wish to emphasize the following point: In the actual computation, we neither change the evolution algorithm nor modify it in any way. The equivalent expression, Equation 4, is simply a convenient form that allows us to analyze the algorithm by using methods of stochastic approximation.

2.3 An Assumption

In the sequel, K denotes a generic positive constant. By convention, $K + K = K$ and $KK = K$. We make the following assumption throughout the rest of the paper:

(A) The function $f(\cdot)$ is convex and is twice continuously differentiable such that $f_{xx}(\cdot)$ is bounded; that is, for all $x \in \mathbb{R}^d$, $|f_{xx}(x)| \leq K$, where $f_{xx}(\cdot)$ denotes the second derivative (or Hessian) of $f(\cdot)$. $\{\tilde{z}_n(1)\}, \dots, \{\tilde{z}_n(\lambda)\}$ are sequences of i.i.d. normal random vectors such that $\tilde{z}_n(i) \sim N(0, I_d)$ for $i = 1, \dots, \lambda$.

REMARK: Since we can generate $\{\tilde{z}_n(i)\}$, they are at our disposal, so the i.i.d. condition is not a restriction. From the basic assumption, we know that the components of $\tilde{z}_n(i)$ denoted by $\tilde{z}_{n,j}(i)$ for $j = 1, 2, \dots, d$ are $N(0, 1)$ and that $\tilde{z}_{n,1}(i), \dots, \tilde{z}_{n,d}(\lambda)$ are also independent. Since $f_{xx}(\cdot)$ is continuous, it is bounded on bounded sets. Here, we require a slightly stronger condition. When we carry out the EA computation, we are normally interested in the solution on bounded sets only. Furthermore, it is possible to design algorithms with projections and/or truncations so that the boundedness of the iterates is fulfilled.

3. Convergence

Under very natural conditions, we derive the convergence theorem and relate the discrete iteration to a continuous time ordinary differential equation. Note that the iterates x_n in Equation 4 should really have been written as x_n^a . We have suppressed the a -dependence to keep the notation simple. If it is necessary later, we may retain it as needed.

We recall the definition of weak convergence first. A sequence of random variables $\{w_n\}$ is said to converge to w weakly if for any bounded and continuous function $g(\cdot)$, $Eg(w_n) \rightarrow Eg(w)$ as $n \rightarrow \infty$. Weak convergence is a substantial generalization of the notion of convergence in distribution. It implies much more than the simple convergence of multidimensional distributions, since the function $g(\cdot)$ can be chosen in various ways. Note that $g(\cdot)$ is an arbitrary bounded and continuous function and is not related to the objective function $f(\cdot)$ in any way. The concept of weak convergence can be employed not only to random variables in a Euclidean space, but also to random processes taking values in function spaces as well.

In the process of getting a weak convergence result, one often needs to verify that the sequence involved is tight. A sequence $\{w_n\}$ of \mathbb{R}^d -valued process is tight if for any $\varepsilon > 0$, there is a compact set S_ε in \mathbb{R}^d such that $P(w_n \notin S_\varepsilon) \leq \varepsilon$ for all n . The definition of tightness carries over to the more general metric space-valued sequences. A well-known theorem, due to Prohorov (see Ethier & Kurtz, 1986), states that in a complete separable metric space, tightness is equivalent to sequential compactness. In other words, once tightness is verified, one may proceed to extract convergent subsequences.

There are reasons that weak convergence analysis is preferable in many applications. First, it requires much weaker conditions than its with-probability-1 convergence counterpart. Second, dealing with the problem of rates of convergence, we often need to obtain weak limit results. Therefore, one is forced to treat convergence in distribution or convergence in the weak sense anyway. 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. In addition, the constant step size algorithms are known to have the ability of tracking small parameter variations and are rather robust with respect to the noise processes.

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 a detailed account of the concept of weak convergence as well as many related materials, we refer the reader to Ethier and Kurtz (1986) and the references therein.

In our weak convergence analysis, below, we often work with $D^d[0, \infty)$, which is the space of functions that are right continuous and have left-hand limits endowed with a certain weak topology (Skorokhod topology). Our analysis requires that first the tightness of the underlying processes be verified and then the limit process be characterized.

To proceed, with step size parameter a , we define a process $x^a(\cdot)$ by a piecewise constant interpolation as follows:

$$x^a(t) = x_n \text{ for } t \in [na, na + a)$$

Thus, in lieu of examining the discrete iterates, we treat the process in continuous time, which gives us a better description of the dynamic behavior of the system involved.

In what follows, we apply the direct averaging methods (see Kushner, 1984, Chapter 5) to study the process $x^a(\cdot)$. Notice that because of the distinct features of the evolutionary algorithms, the argument in Kushner (1984) must be modified for our needs.

In the weak convergence approach, normally one needs to have an averaging condition of law-of-large-numbers type. Such a condition now holds for our case based on the basic assumption (A). The essence of the direct averaging approach is to treat the variable x as fixed and only average out the noise processes. Keeping this in mind, we first derive a preparatory result below.

THEOREM 3.1: *Let E_m denote the conditional expectation with respect to the σ -algebra \mathcal{F}_m , generated by $\{x_0, \tilde{z}_j(i), j < m, i = 1, 2, \dots, \lambda\}$. Under Condition (A),*

1. *For each x , denote*

$$E \sum_{i=1}^{\lambda} \tilde{z}_n(i) I_{\{f(x+\tilde{z}_n(i))=\min_{\mu \in \Lambda_n} f(x+\mu)\}} = \zeta(x)$$

Then for any $n \geq m$,

$$E_m \sum_{i=1}^{\lambda} \tilde{z}_n(i) I_{\{f(x+\tilde{z}_n(i))=\min_{\mu \in \Lambda_n} f(x+\mu)\}} = \zeta(x) \tag{5}$$

- For each $x \neq x_{\min}$, (that is, $f(x_{\min}) = \min_x f(x)$), $\zeta(x)$ is in the descent (or “downhill”) direction; that is, $f'_x(x)\zeta(x) < 0$. Moreover, there exists a function $\hat{H}(\cdot)$ such that for all r with $0 < r < \hat{H}(f'_x(x))$, we have $f(x+r\zeta(x)) < f(x)$.

REMARK: We notice that $\zeta(\cdot)$ is a function of x . It also implicitly depends on the function $f(\cdot)$. The second assertion indicates that the algorithm is stable. (We will be more precise about this later.) In view of the first assertion in the above theorem, for each m , each n , and each x ,

$$\frac{1}{n} \sum_{k=m}^{n+m-1} E_m \sum_{i=1}^{\lambda} \tilde{z}_k(i) I_{\{f(x+\tilde{z}_k(i))=\min_{\mu \in \Lambda_k} f(x+\mu)\}} = \zeta(x)$$

which will be needed later. Thus a law-of-large-numbers type of condition holds for the underlying sequence.

PROOF OF THEOREM 3.1: The verification of the first assertion above is almost obvious. We simply note that although $\sum_{i=1}^{\lambda} \tilde{z}_n(i) I_{\{f(x_n+\tilde{z}_n(i))=\min_{\mu \in \Lambda_n} f(x_n+\mu)\}}$ is used, in any given instance, only one of the random vectors is nonzero. For any $n \geq m$, since $\tilde{z}_n(i)$ is independent of \mathcal{F}_m ,

$$E_m \sum_{i=1}^{\lambda} \tilde{z}_n(i) I_{\{f(x+\tilde{z}_n(i))=\min_{\mu \in \Lambda_n} f(x+\mu)\}} = E \sum_{i=1}^{\lambda} \tilde{z}_n(i) I_{\{f(x+\tilde{z}_n(i))=\min_{\mu \in \Lambda} f(x+\mu)\}} = \zeta(x)$$

As for the second part of the theorem, for simplicity we suppress the subscript in z_n and write it as z instead in what follows. We note that the $(1, \lambda)$ -ES generates λ i.i.d. Gaussian random variables, which can be decomposed as $z \stackrel{D}{=} ru$, where u denotes a random vector distributed uniformly on the surface of a unit hypersphere and r a nonnegative random variable stochastically independent from u .

Let x_n be the current position with $x_n \neq x_{\min}$. Then the algorithm compares the values

$$f(x_n + z(1)), f(x_n + z(2)), \dots, f(x_n + z(\lambda))$$

The new position x_{n+1} is the argument $x_n + z(b)$, $b \in \{1, \dots, \lambda\}$ that offers the lowest objective function value. Owing to Condition (A), for $L > 0$

$$f(x + b) \leq f(x) + f'_x(x) b + Lb' b \tag{6}$$

the algorithm compares the values $f(x) - f(x + z(i))$, which can be bounded by

$$f(x) - f(x + z) \geq -f'_x(x) z - Lz' z = -rf'_x(x) u - r^2 L \tag{7}$$

with the use of Inequality 6. If we can show that the expectation of the maximum of the right-hand side of Inequality 7 within λ trials is greater than zero, it is clear that the expectation of the selected step $z(b)$ is a step of descent.

Assuming that r is a constant for the moment, Inequality 7 may be used to formulate the condition

$$\max_i \{ -f'_x(x) u_i \} \geq rL$$

where $f'_x(x)$ denotes the transpose of $f_x(x)$. This leads to

$$\max_i \{ |f_x(x)| \cos \omega_i \} \geq rL$$

where we used the relation

$$\frac{-f'_x(x) u_i}{|f_x(x)|} = \cos \omega_i$$

Here, ω_i denotes the random angle between the negative gradient and the random direction u_i . Finally, taking expectations, we obtain

$$M_\lambda := E[\max_i \{ \cos \omega_i \}] \geq \frac{rL}{|f_x(x)|}$$

so that

$$0 \leq H(f_x(x)) := r \leq \frac{M_\lambda}{L} |f_x(x)| := \hat{H}(f_x(x)) \tag{8}$$

To proceed, we state a lemma. Its proof is presented in the appendix to this article.

LEMMA 3.2: *The random variable $c = \cos \omega$ possesses a beta distribution with parameters $p = q = d - 1/2$ and distribution function*

$$F_c(x) = \frac{1}{B((d-1)/2, (d-1)/2)} \int_0^{(x+1)/2} y^{(d-3)/2} (1-y)^{(d-3)/2} dy \tag{9}$$

where d denotes the dimensionality of random vector u .

Now come back to the proof of Theorem 3.1. Using Lemma 3.2, it remains to be shown that $M_\lambda > 0$. Note that $M_\lambda = 2 \cdot E[\max\{B_i; i = 1, \dots, \lambda\}] - 1$, where the random variables B_i possess beta distribution with parameters as in Equation 9 but with support $(0, 1)$. Let $m_\lambda = E[\max\{B_i; i = 1, \dots, \lambda\}]$. Since the probability density function of B_i is unimodal and symmetric with respect to $1/2$, we may use the inequality (see David, 1970, p. 63)

$$F(m_\lambda) \geq \frac{\lambda}{\lambda + 1} > \frac{1}{2} \quad (\lambda \geq 2)$$

where $F(\cdot)$ denotes the distribution function of B_i . Since $F(\cdot)$ is convex-concave, the inverse $F^{-1}(\cdot)$ is concave-convex, so $m_\lambda > F^{-1}(1/2)$. Moreover, $F(1/2) = 1/2$ owing to symmetry. Thus, $m_\lambda > 1/2$ for $\lambda \geq 2$ and $M_\lambda = 2 m_\lambda - 1 > 0$, and the proof of Theorem 1 is completed. \square

REMARK: The second part of Theorem 3.1 was shown for the case in which r is a deterministic choice, whereas $z \stackrel{D}{=} ru$ was assumed to be a normal random vector. We argue that this result reflects the situation in large parameter spaces ($d \gg 1$). Since $r \sim \chi_n(\sigma)$, we have $E(r) \approx \sigma\sqrt{d-1/2}$ and $Var(r) \approx \sigma^2/2$. Therefore, the realized step size r has only small

variation, so we can regard $r = \sigma \sqrt{d - 1/2} \approx \sigma \sqrt{d}$ as a good approximation. Then relation (8) becomes

$$\sigma < \frac{M_\lambda}{L\sqrt{d}} |f_x(x)|$$

Now we are in a position to present the weak convergence theorem which links the discrete iteration with a continuous time dynamic system.

THEOREM 3.3: *Suppose that Condition (A) holds and $H(\cdot)$ and $\zeta(\cdot)$ are continuous. Then $\{x^a(\cdot)\}$ is tight in $D^d[0, \infty)$, such that as $a \rightarrow 0$, every weakly convergent subsequence has a limit $x(\cdot)$ satisfying the differential equation*

$$\dot{x} = \frac{dx}{dt} = H(f_x(x))\zeta(x), \quad x(0) = x_0 \tag{10}$$

provided that the equation has a unique solution for each initial condition x_0 .

REMARK: Before going through the proof, let us make the following remarks: The conditions indicate that $H: \mathbb{R}^d \mapsto \mathbb{R}$ is a continuous function. Therefore, the composite function $H(f_x(\cdot))$ is continuous. We also require that $\zeta: \mathbb{R}^d \mapsto \mathbb{R}^d$ be continuous.

In the theorem above, we assumed that the initial condition x_0 is the same as that of the discrete iteration and does not depend on a . More complex situations, —for example, $x_0 = x_0^a$ —can be treated. The analysis is about the same. The only difference is that we have to add another condition: x_0^a converges weakly to x_0 as $a \rightarrow 0$. The continuity of $\zeta(x)$ is not a restriction. Taking expectation is a smoothing process. Even indicator functions after such a process become continuous.

PROOF OF THEOREM 3.3: For clarity, we divide the proof into several steps.

Step 1 (the use of N -truncation): Since it is not known a priori whether the iterates $\{x_n\}$ are bounded, we utilize the N -truncation technique (see Kushner, 1984, p. 43). For each $N < \infty$, define $S_N = \{x; |x| \leq N\}$ —that is, the ball with radius N . Recall that $x^{a,N}(t)$ is an N -truncation of $x^a(t)$ if $x^{a,N}(t) = x^a(t)$ until the first exit from S_N and that

$$\lim_{A \rightarrow \infty} \limsup_{a \rightarrow 0} P \left\{ \sup_{t \leq T} |x^{a,N}(t)| \geq A \right\} = 0 \text{ for each } T < \infty \tag{11}$$

For the discrete algorithm, we use

$$x_{n+1}^N = x_n^N + aH(f_x(x_n^N)) \sum_{i=1}^{\lambda} \tilde{z}_n^{(i)} I_{\{f(x_n^N + z_n^{(i)}) = \min_{\mu \in \Lambda_n} f(x_n^N + \mu)\}} q_N(x_n^N) \tag{12}$$

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

$$q_N(x) = \begin{cases} 1, & x \in S_N \\ 0, & x \in \mathbb{R}^d - S_{N+1} \\ \text{smooth}, & \text{otherwise} \end{cases}$$

Step 2 (tightness of the truncated process $\{x_n^{a,N}(\cdot)\}$): By the definition (12), $\{x_n^N\}$ is bounded.

Since $H(\cdot)$ is a continuous function and $f(\cdot)$ is C^2 , $H(f_x(x_n^N))$ is bounded. As a result,

$$\begin{aligned} & E \left| H(f_x(x_n^N)) \sum_{i=1}^{\lambda} \tilde{z}_n(i) I_{\{f(x_n^N + z_n(i)) = \min_{\mu \in \Lambda_n} f(x_n^N + \mu)\}} \right|^2 \\ & \leq KE \sum_{i=1}^{\lambda} |\tilde{z}_n(i)|^2 I_{\{f(x_n^N + z_n(i)) = \min_{\mu \in \Lambda_n} f(x_n^N + \mu)\}} \\ & \leq K \sum_{i=1}^{\lambda} E |\tilde{z}_n(i)|^2 < \infty \end{aligned}$$

It follows that

$$\left\{ H(f_x(x_n^N)) \sum_{i=1}^{\lambda} \tilde{z}_n(i) I_{\{f(x_n^N + z_n(i)) = \min_{\mu \in \Lambda_n} f(x + \mu)\}} \right\}$$

is uniformly integrable. Lemma 7 in Chapter 3 of Kushner (1984) then yields that $\{x^{a,N}(\cdot)\}$ is tight and that the limit of any convergent subsequence has continuous paths w.p.1. Now extract a convergent subsequence. For notational simplicity, continue to use a as the index of the subsequence and denote the limit by $x^N(\cdot)$. By using the Skorokhod representation, without loss of generality, we may assume that $x^{a,N}(\cdot)$ converges to $x^N(\cdot)$ w.p.1, and that the convergence is uniform on any bounded interval. Our next task is to characterize the limit process.

Step 3 (characterization of the limit process $x^N(\cdot)$): Our objective here is to show that the limit process $x^N(\cdot)$ satisfies a truncated version of Equation 10. Introduce the statement

$$M^N(t) = x^N(t) - x^N(0) - \int_0^t H(f_x(x^N(\tau))) \zeta(x^N(\tau)) q_N(x^N(\tau)) d\tau$$

It is easily seen that $M^N(\cdot)$ is Lipschitz-continuous. If it is a martingale, it must be a constant (see Kushner, 1984). However, $M^N(0) = 0$. As a result, $M^N(t)$ must be 0 identically, or, equivalently, $x^N(\cdot)$ is a solution of the ordinary differential equation 10.

Thus, the problem can be reduced to verifying the martingale property of $M^N(\cdot)$. To prove this, we need only show that for any bounded and continuous function $b(\cdot)$, any integer ν and $j \leq \nu$, with $t_j \leq t < t + s$,

$$Eb(x^N(t_j), j \leq \nu)[M^N(t + s) - M^N(t)] = 0$$

We begin with the process $x^{a,N}(\cdot)$. In what follows, $(t + s)/a$, t/a , and so on are all meant to be integers for notational convenience. (If they are not integers, we can always take the integral parts anyway.)

By using the interpolation,

$$\begin{aligned} & \lim_a Eb(x^{a,N}(t_j), j \leq \nu)[x^{a,N}(t + s) - x^a(t)] \\ & = \lim_a Eb(x^{a,N}(t_j), j \leq \nu)[x_{(t+s)/a}^N - x_{t/a}^N] \end{aligned}$$

$$= \lim_a Eb(x^{a,N}(t_j), j \leq \nu) \left(a \sum_{k=t/a}^{(t+s)/a} H(f_x(x_k^N)) \times \sum_{i=1}^{\lambda} \tilde{z}_k(i) I_{\{f(x_k^N+z_k(i))=\min_{\mu \in \Lambda_k} f(x_k^N+\mu)\}} q_N(x_k^N) \right) \quad (13)$$

Now choose a sequence of integers $\{n_a\}$ satisfying $n_a \rightarrow \infty$ as $a \rightarrow 0$ but $\delta_a = an_a \rightarrow 0$ as $a \rightarrow 0$. Subdivide the interval $[t/a, (t+s)/a]$ into intervals with length n_a . The term on the right side of the last equality sign of Equation 13 then can be rewritten as

$$\begin{aligned} & \lim_a Eb(x^{a,N}(t_j), j \leq \nu) \\ & \times \left(\sum_{k=t/a}^{(t+s)/a} aH(f_x(x_k^N)) \sum_{i=1}^{\lambda} \tilde{z}_n(i) I_{\{f(x^N+z_n(i))=\min_{\mu \in \Lambda_k} f(x_k^N+\mu)\}} q_N(x_k^N) \right) \\ = & \lim_a Eb(x^{a,N}(t_j), j \leq \nu) \\ & \times \left(\sum_{l\delta_a=t}^{(t+s)} \delta_a \frac{1}{n_a} \sum_{k \in L_a} E_{l n_a} H(f_x(x_k^N)) \sum_{i=1}^{\lambda} \tilde{z}_k(i) I_{\{f(x^N+z_n(i))=\min_{\mu \in \Lambda_k} f(x_k^N+\mu)\}} q_N(x_k^N) \right) \end{aligned}$$

where L_a denotes the interval $l n_a \leq k \leq l n_a + n_a - 1$, and $E_{l n_a}$ denotes the conditioning on $\mathcal{F}_{l n_a}$, the σ -algebra generated by $\{x_0, z_k(i), k < l n_a, i = 1, 2, \dots, \lambda\}$. Since $x^N(t_j)$ for $j \leq \nu$ are $\mathcal{F}_{l n_a}$ -measurable, this conditioning can be inserted.

Define the piecewise constant interpolation of

$$\frac{1}{n_a} \sum_{k \in L_a} E_{l n_a} H(f_x(x_k^N)) \sum_{i=1}^{\lambda} \tilde{z}_k(i) I_{\{f(x_k^N+z_k(i))=\min_{\mu \in \Lambda_k} f(x_k^N+\mu)\}} q_N(x_k^N)$$

on $[l\delta_a, l\delta_a + \delta_a)$ as $\tilde{H}^a(\tau)$. Then

$$\begin{aligned} \lim_n Eb(x^{a,N}(t_j), j \leq \nu) & \left(\int_t^{t+s} \tilde{H}^a(\tau) du - \sum_{l\delta_a=t}^{t+s} \delta_a \frac{1}{n_a} \sum_{k \in L_a} E_{l n_a} H(f_x(x_k^N)) \right. \\ & \left. \times \sum_{i=1}^{\lambda} \tilde{z}_k(i) I_{\{f(x_k^N+z_k(i))=\min_{\mu \in \Lambda_k} f(x_k^N+\mu)\}} q_N(x_k^N) \right) = 0 \end{aligned}$$

In view of the above equation, we need only consider the limit of the function $\tilde{H}^a(\cdot)$ as $a \rightarrow 0$.

By using the nested conditional expectation, for $k \geq l n_a$,

$$\begin{aligned} & E_{l n_a} H(f_x(x_k^N)) \sum_{i=1}^{\lambda} \tilde{z}_k(i) I_{\{f(x^N+z_k(i))=\min_{\mu \in \Lambda_k} f(x_k^N+\mu)\}} q_N(x_k^N) \\ = & E_{l n_a} H(f_x(x_k^N)) \sum_{i=1}^{\lambda} E_k \tilde{z}_k(i) I_{\{f(x_k^N+z_k(i))=\min_{\mu \in \Lambda_k} f(x_k^N+\mu)\}} q_N(x_k^N) \\ = & E_{l n_a} H(f_x(x_k^N)) \zeta(x_k^N) q_N(x_k^N) \end{aligned}$$

by Theorem 3.1.

In view of the continuity of the functions $H(\cdot)$, $f(\cdot)$, $\zeta(\cdot)$, and $q_N(\cdot)$, we have

$$\begin{aligned} \frac{1}{n_a} \sum_{k \in L_a} H(f_x(x_k^N)) \zeta(x_k^N) q_N(x_k^N) &= \frac{1}{n_a} \sum_{k \in L_a} H(f_x(x_{m_a}^N)) \zeta(x_{m_a}^N) q_N(x_{m_a}^N) + o(1) \\ &= H(f_x(x^{a,N}(\delta_a))) \zeta(x^{a,N}(\delta_a)) q_N(x_{m_a}^N) + o(1) \end{aligned}$$

where $o(1) \xrightarrow{a} 0$ in probability. Letting $\delta_a \rightarrow \tau$, we can further replace

$$\begin{aligned} &H(f_x(x^{a,N}(\delta_a))) \zeta(x^{a,N}(\delta_a)) q_N(x_{m_a}^N) \\ &H(f_x(x^N(\tau))) \zeta(x^N(\tau)) q_N(x^N(\tau)) \end{aligned}$$

The limit for the truncated process is thus proved.

Step 4 (the result for the untruncated process): The proof is similar to that of Theorem 2 of Chapter 3 in Kushner (1984). Let $P_{x(0)}(\cdot)$ and $P^N(\cdot)$ be the measures induced by $x(\cdot)$ and $x^N(\cdot)$, respectively. Owing to the uniqueness of Equation 10, $P_{x(0)}(\cdot)$ is unique. For each $T < \infty$, $P_{x(0)}(B) = P^N(B)$ for each Borel subset $B \subset \mathcal{B}$, where

$$B = \{x(\cdot) \in D^d[0, \infty); x(t) \in S_N \text{ for each } t \leq T\}$$

Observe that

$$P_{x(0)}\{\sup_{t \leq T} |x(t)| \leq N\} \xrightarrow{N} 1 \quad \text{as } N \rightarrow \infty$$

Thus, the desired result follows. The proof of the theorem is completed. \square

4. Further Asymptotic Results

We derive further asymptotic properties in this section. In the first subsection, we derive an upper bound of the estimation errors; in the second subsection, we study the case in which a is small and n is large and obtain limit results.

4.1 An Upper Bound on the Estimation Error

The result is recorded in Theorem 4.1, below. It indicates how the estimation errors depend on the step size, and gives us a way to assess the rate of convergence.

THEOREM 4.1: *Let the conditions of Theorem 3.3 be satisfied. Suppose that θ is an asymptotically stable point of Equation 10, and that there is a twice continuously differentiable Liapunov function $V(\cdot): \mathbb{R}^d \mapsto \mathbb{R}$ such that $V(x) \geq 0$ for all x , $V(x) \xrightarrow{|x| \rightarrow \infty} \infty$, $V_{xx}(\cdot)$ is bounded, and $V'_x(x)H(f_x(x))\zeta(x) < -\eta V(x)$ for all $x \neq \theta$ and for some $\eta > 0$. Furthermore, let $|H(x)| \leq K(1 + |x|)$ and $|f_x(x)|^2 \leq K(1 + V(x))$. Then the following statements hold:*

1. $EV(x_n) < \infty$ and hence $\{x_n\}$ is tight in \mathbb{R}^d .
2. There is an $N_a > 0$ and an $a_0 > 0$ such that for all $n \geq N_a$ and all $a \leq a_0$, $EV(x_n) = O(a)$.

REMARK: It has been shown in Theorem 3.1 that the vector field $\zeta(\cdot)$ is always in the downhill direction. Our analysis is based on the Liapunov stability theory. Although we have assumed the existence of a Liapunov function, its actual form need not be known. As far as the theoretical development is concerned, there is no loss of generality in assuming that $\theta = 0$,

since we can always translate the coordinate axes by subtracting θ . Henceforth, we work with the case $\theta = 0$. This is rather convenient for notational concern.

PROOF OF THEOREM 4.1: We shall only prove the second part of the theorem. The proof of the first part is easier.

Using the recursive formula (4), and owing to the fact that x_n is \mathcal{F}_n -measurable, together with the continuity of $H(\cdot)$ and $f(\cdot)$,

$$\begin{aligned}
 E_n V(x_{n+1}) - V(x_n) &= aV'_x(x_n)H(f_x(x_n)) \sum_{i=1}^{\lambda} \tilde{z}_n(i) I_{\{f(x_n+z_n(i))=\min_{\mu \in \Lambda_n} f(x_n+\mu)\}} \\
 &\quad + a^2 H^2(f_x(x_n)) \sum_{i=1}^{\lambda} \tilde{z}'_n(i) I_{\{f(x_n+z_n(i))=\min_{\mu \in \Lambda_n} f(x_n+\mu)\}} \\
 &\quad \times V_{xx}(x_n^*) \sum_{i=1}^{\lambda} \tilde{z}_n(i) I_{\{f(x_n+z_n(i))=\min_{\mu \in \Lambda_n} f(x_n+\mu)\}} \\
 &\leq aV'_x(x_n)H(f_x(x_n))\zeta(x_n) + Ka^2 H^2(f_x(x_n))E_n |\tilde{z}_n|^2 \\
 &\leq -\eta aV(x_n) + Ka^2(1 + V(x_n))
 \end{aligned} \tag{14}$$

where x_n^* is a point on the line segment joining x_n and x_{n+1} .

For sufficiently small $a > 0$ (that is, there exists an $a_0 > 0$ such that for all $0 < a \leq a_0$), $Ka^2 - \eta a \leq -\eta/2a$, and as a result

$$E_n V(x_{n+1}) \leq (1 - \eta/2a)V(x_n) + Ka^2$$

Taking expectation and iterating on the above inequality leads to

$$\begin{aligned}
 EV(x_{n+1}) &\leq (1 - \eta a/2)^n EV(x_0) + Ka^2 \sum_{i=0}^n (1 - \eta a/2)^i \\
 &\leq (1 - \eta a/2)^n EV(x_0) + Ka
 \end{aligned} \tag{15}$$

Choose N_a such that for all $n \geq N_a$, $(1 - \eta a/2)^n \leq Ka$. Then the desired estimate follows from Inequality 15. This concludes the proof of the theorem. \square

To proceed, we note that Theorem 3.3 gives us a convergence result that is on an arbitrarily large but still bounded time interval. One of our goals is to figure out what happens when a is small and n is large. This problem is treated in Section 4.2. As before, we assume that $\theta = 0$ throughout.

4.2 Convergence for $a \downarrow 0$ and $n \uparrow \infty$

For all problems, it is important to have a convergence result that is uniform in t . The way to study the problem is similar to what has been done in Theorem 3.3. However, we wish to have the time variable tend to infinity. Introduce another sequence $\{t_a\}$ such that $t_a \rightarrow \infty$ as $a \rightarrow 0$. The significance of the differential equation 10 is that its stationary points correspond to the points $f_x(x) = 0$ we are searching for. In lieu of $x^a(\cdot)$ as defined in Section 3, consider $x^a(\cdot + t_a)$. Note that the weak convergence result alone does not imply that $x^a(\cdot + t_a)$ converges weakly to a stationary solution to Equation 10. In what follows, we establish the convergence of this sequence.

THEOREM 4.2: *Under the conditions of Theorem 4.1, $x^a(\cdot + t_a)$ converges to $\theta = 0$ weakly.*

PROOF: The proof is quite similar to that of Theorem 3.3 and to a corresponding result in Kushner and Yin (1987), so we will be very brief. For each $T < \infty$, consider the pair $(x^a(\cdot + t_a), x^a(\cdot - T + t_a))$. The tightness can be obtained as in the previous case. Therefore, we can extract a convergent subsequence, still use a as the index, and denote the limit by $(x(\cdot), x_T(\cdot))$. It is clear that $x(0) = x_T(T)$ owing to the construction. The value of $x_T(0)$ may not be known. However, in accordance with Theorem 4.1, the value of it belongs to a set that is tight. As a result, by the stability argument, for any $\hat{\varepsilon} > 0$, there is a $T_{\hat{\varepsilon}}$ such that for all $T \geq T_{\hat{\varepsilon}}$, $P(|x_T(T)| > \hat{\varepsilon}) \leq \hat{\varepsilon}$. This completes the proof of the theorem. \square

5. Algorithms with Decreasing Step Size

There are times when we may wish to use decreasing step size algorithms. This section concentrates on the study of such algorithms related to the $(1, \lambda)$ strategy. Consider

$$x_{n+1} = x_n + a_n H(f_x(x_n)) \sum_{i=1}^{\lambda} \tilde{z}_n(i) I_{\{f(x_n+z_n(i))=\min_{\mu \in \Lambda_n} f(x_n+\mu)\}} \tag{16}$$

where $\{a_n\}$ is a sequence of nonnegative real numbers such that

$$a_n \xrightarrow{n \rightarrow \infty} 0, \quad \sum_{n=1}^{\infty} a_n = \infty, \quad \sum_{n=1}^{\infty} a_n^2 < \infty$$

A typical choice of $\{a_n\}$ is $a_n = O(1/n^\gamma)$, with $1/2 < \gamma \leq 1$. In what follows, we derive the w.p.1 convergence result and then establish an upper bound on the estimation errors.

5.1 Wp.1 Convergence

In this section, we obtain a w.p.1 convergence result by using the ordinary differential equation method (see Benveniste, Métivier, & Priouret (1990), Kushner & Clark (1978), Ljung (1977), and the references in each). For future use define

$$\hat{S} = \{x; H(f_x(x))\zeta(x) = 0\}$$

$$\psi_n(x) = \sum_{i=1}^{\lambda} \tilde{z}_n(i) I_{\{f(x+z_n(i))=\min_{\mu \in \Lambda_n} f(x+\mu)\}}$$

THEOREM 5.1: *Let (A) hold. Suppose that there is a twice continuously differentiable Liapunov function $V(\cdot): \mathbb{R}^d \mapsto \mathbb{R}$ such that $V(x) \geq 0$ for all x , $V(x) \xrightarrow{|x| \rightarrow \infty} \infty$, $V_{xx}(\cdot)$ is bounded, and $V'_x(x)H(f_x(x))\zeta(x) < -\eta_0$ for all x and for some $\eta_0 > 0$, and $|V'_x(x)H(f_x(x))\zeta(x)| \leq K(1 + V(x))$. Furthermore, $|H(x)| \leq K(1 + |x|)$ and $|f_x(x)|^2 \leq K(1 + |V'_x(x)H(f_x(x))\zeta(x)|)$. Then $\{x_n\}$ is bounded w.p.1. In addition, x_n converges to \hat{S} w.p.1; that is,*

$$P\left(\lim_{n \rightarrow \infty} \rho(x_n, \hat{S}) = 0\right) = 1$$

where $\rho(\cdot)$ denotes the usual distance function such that $\rho(x, \hat{S}) = \inf_{y \in \hat{S}} \rho(x, y)$. In particular, if $\hat{S} = \{\theta\}$, a singleton set, then $x_n \xrightarrow{n \rightarrow \infty} \theta$ w.p.1.

PROOF: We first note owing to the assumptions of the theorem,

$$H^2(f_x(x)) \leq K(1 + |V'_x(x)H(f_x(x))\zeta(x)|)$$

Direct computation yields

$$\begin{aligned} E_n V(x_{n+1}) - V(x_n) &= a_n V'_x(x_n)H(f_x(x_n))E_n \psi_n(x_n) \\ &\quad + a_n^2 H^2(f_x(x_n))E_n \psi'_n(x_n)V_{xx}(x_n^*)\psi_n(x_n) \\ &= a_n(1 + a_n \bar{\varepsilon}_n)V'_x(x_n)H(f_x(x_n))\zeta(x_n) + a_n^2 \varepsilon_n \\ &\leq -Ka_n \eta_0 + a_n^2 \varepsilon_n < 0 \end{aligned} \tag{17}$$

for some N_0 and all $n \geq N_0$, where x_n^* is a point with all of its components located between x_n and x_{n+1} ; $\{\bar{\varepsilon}_n\}$ and $\{\varepsilon_n\}$ are sequences of uniformly bounded random variables.

In view of (17), for $n \geq N_0$, $V(x_n)$ is a supermartingale. Since $V(x_n) \geq 0$, the limit of $V(x_n)$ exists w.p.1. Since $V(x) \xrightarrow{|x| \rightarrow \infty} \infty$, the boundedness of $\{x_n\}$ (in the sense of w.p.1) follows.

To complete the proof of the theorem, we apply a result from Kushner and Clark (1978). First rewrite the recursion as

$$x_{n+1} = x_n + a_n H(f_x(x_n))\zeta(x_n) + a_n H(f_x(x_n))[\psi_n(x_n) - \zeta(x_n)] \tag{18}$$

In accordance with Theorem 2.4.2 of Kushner and Clark (1978), we need only verify that

$$m_n = \sum_{i=1}^n a_i H(f_x(x_i))[\psi_i(x_i) - \zeta(x_i)] \tag{19}$$

converges w.p.1. It is readily seen that m_n is a martingale. Since $\{x_n\}$ is bounded w.p.1, $H(f_x(x_n))$ is also bounded w.p.1 by the continuity of $H(\cdot)$ and $f_x(\cdot)$. Consequently, since x_i is \mathcal{F}_i -measurable,

$$\sum_{i=1}^{\infty} a_i^2 H^2(f_x(x_i))E_i |\psi_i(x_i) - \zeta(x_i)|^2 \leq K \sum_{i=1}^{\infty} a_i^2 [E|\tilde{z}_n(i)|^2 + E|\zeta(x_i)|^2] < \infty$$

since $\{x_i\}$ is bounded w.p.1 and $\zeta(\cdot)$ is continuous. Owing to the local martingale convergence theorem in Chow (1965), Equation 19 holds. Now applying Theorem 2.4.2 of Kushner and Clark (1978), the desired results follow. \square

5.2 $EV(x_n) = O(1/n^\gamma)$ for $a_n = 1/n^\gamma$

Next, we obtain an upper bound for the estimation errors. Since the techniques and details are similar to those of Theorem 4.1, we shall omit the proof.

THEOREM 5.2: *Let the conditions of Theorem 5.1 be satisfied and $a_n = 1/n^\gamma$ for some $1/2 < \gamma \leq 1$. Suppose that θ is an asymptotically stable point of (10), and suppose that in addition to the conditions of Theorem 5.1, the Liapunov function $V(\cdot)$ satisfies $V'_x(x)H(f_x(x))\zeta(x) < -\eta V(x)$ for all $x \neq \theta$ and for some $\eta > 0$. Then there is an $N > 0$ such that for all $n \geq N$, $EV(x_n) = O(n^{-\gamma})$.*

6. Concluding Remarks

In this work, asymptotic properties of the $(1, \lambda)$ evolution strategy were developed through the use of stochastic approximation methods. The evolutionary algorithm was rewritten in

a recursive form and then the analytic tools in stochastic approximation were employed to carry out the investigation. We considered both constant step size and decreasing step size algorithms. Under suitable conditions, we have obtained the convergence and the error bounds of the underlying algorithm. Our current effort is to study more complex situations and to extend the results to evolution strategies with noisy evolution.

This paper is our first effort to use a stochastic approximation method to analyze the evolution strategies. One of the immediate questions is whether we can extend the results to nonsmooth functions $f(\cdot)$. We believe we can. It seems that we can deal with nonsmooth functions by using nonsmooth analysis techniques in conjunction with stochastic approximation methods. The investigation of the corresponding recursive procedures then requires techniques from set-valued analysis, and the differential equations become differential inclusions. Many more details deserve further study and in-depth investigation.

Recently, several modifications on the standard stochastic approximation procedures were proposed by Polyak, Ruppert, and Bather (see Yin & Yin (1994) or Yin (1995) for more details and references). These result in asymptotical optimality. These attempts soon attracted much attention. One of the elements of their approach is the use of arithmetic averaging. It is conceivable that such an idea may be well suited to the study of the evolutionary algorithms.

A further promising observation is that stochastic approximation methods provide a route to convert discrete time EAs to *equivalent* continuous time evolutionary processes so that we can apply the well-developed mathematical apparatus for analyzing continuous time stochastic processes to the investigation of evolutionary algorithms that are more complex than the $(1, \lambda)$ EA considered here.

Appendix: Distribution of $\cos \omega$

PROOF OF LEMMA 3.2: Let $X \sim N(0, I_d)$. Then $U = X/|X|$ is uniformly distributed on a hypersphere surface of dimension d and X and $|X|$ are stochastically independent (see Fang, Kotz, & Ng (1990)). We wish to answer the question: What is the distribution of $\cos \omega$, where ω is an angle between a vector y and u ? Without loss of generality, choose $y = e_1 = (1, 0, \dots, 0)'$. Then

$$\cos(\omega) = e_1' u = \frac{e_1' X}{|X|} = \frac{X_1}{|X|}$$

and

$$Z = \cos^2 \omega = \frac{X_1^2}{\sum_{i=1}^d X_i^2} = \frac{X_1^2}{X_1^2 + \sum_{i=2}^d X_i^2}$$

Thus, $X_1^2 \sim \chi_1^2$ and $\sum_{i=2}^d X_i^2 \sim \chi_{d-1}^2$, so $Z \sim \text{Beta}(\frac{1}{2}, \frac{d-1}{2})$ with probability density function (p.d.f.)

$$f(z) = \frac{z^{-1/2} (1-z)^{(d-3)/2}}{B(1/2, (d-1)/2)} \cdot I_{[0,1]}(z)$$

Transformation of the above density with $C = Z^{1/2}$ leads to

$$f(c) = \frac{2}{B(1/2, (d-1)/2)} (1-c)^{(d-3)/2} (1+c)^{(d-3)/2} \cdot I_{[0,1]}(c)$$

which is the p.d.f. of $|\cos \omega|$. Owing to the symmetry of the cosine, we may modify the above p.d.f. to obtain

$$f(c) = \frac{1}{B(1/2, (d-1)/2)} (1-c)^{(d-3)/2} (1+c)^{(d-3)/2} \cdot I_{[-1,1]}(c)$$

so that the distribution function of $\cos \omega$ is

$$F_c(x) = \frac{1}{B(1/2, (d-1)/2)} \int_{-1}^x (1-c)^{(d-3)/2} (1+c)^{(d-3)/2} dc$$

As a result, direct substitutions yield

$$\begin{aligned} F_c(x) &= \frac{2^{d-2}}{B(1/2, (d-1)/2)} \int_0^{(x+1)/2} y^{(d-3)/2} (1-y)^{(d-3)/2} dy \\ &= \frac{1}{B((d-1)/2, (d-1)/2)} \int_0^{(x+1)/2} y^{(d-3)/2} (1-y)^{(d-3)/2} dy \end{aligned}$$

where we have used the relation $B(1/2, (d-1)/2) = 2^{d-2} B((d-1)/2, (d-1)/2)$. Thus, $\cos \omega$ possesses a beta distribution on $[-1, 1]$. \square

References

- Bäck, T., Rudolph, G., & Schwefel, H.-P. (1993). Evolutionary programming and evolution strategies: Similarities and differences. In Fogel, D. & Atmar, W. (Eds.), *Proceedings of the Second Annual Conference on Evolutionary Programming* (pp. 11–22). La Jolla, CA: Evolutionary Programming Society.
- Bäck, T., & Schwefel, H.-P. (1993). An overview of evolutionary algorithms for parameter optimization. *Evolutionary Computation*, 1(1), 1–23.
- Benveniste, A., Métivier, M., & Priouret, P. (1990). *Adaptive algorithms and stochastic approximation*. Berlin: Springer.
- Chow, Y. (1965). Local convergence of martingales and the law of large numbers. *Annals of Mathematical Statistics*, 36, 552–558.
- David, H. (1970). *Order statistics*. New York: Wiley.
- De Jong, K. (1975). *An analysis of the behavior of a class of genetic adaptive systems*. Ph.D. dissertation, University of Michigan.
- Doob, J. L. (1967). *Stochastic processes* (7th ed.). New York: Wiley.
- Ethier, S., & Kurtz, T. (1986). *Markov processes, characterization and convergence*. New York: Wiley.
- Fang, K.-T., Kotz, S., & Ng, K.-W. (1990). *Symmetric multivariate and related distributions*. London and New York: Chapman and Hall.
- Fogel, D. B. (1992). An analysis of evolutionary programming. In Fogel, D. B., & Atmar, W. (Eds.), *Proceedings of the First Annual Conference on Evolutionary Programming* (pp. 43–51). San Diego, CA: Evolutionary Programming Society.
- Fogel, L. J., Owens, A. J., & Walsh, M. J. (1966). *Artificial intelligence through simulated evolution*. New York: Wiley.
- Holland, J. (1962). Outline for a logical theory of adaptive systems. *Journal of the Association for Computing Machinery*, 3, 297–314.
- Kushner, H. (1984). *Approximation and weak convergence methods for random processes, with applications to stochastic systems theory*. Cambridge, MA: MIT Press.

- Kushner, H., & Clark, D. (1978). *Stochastic approximation methods for constrained and unconstrained systems*. New York: Springer.
- Kushner, H., & Yin, G. (1987). Asymptotic properties of distributed and communicating stochastic approximation algorithms. *SIAM Journal on Control and Optimization*, 25 (pp. 1266–1290).
- Ljung, L. (1977). Analysis of recursive stochastic algorithms. *IEEE Trans. on Automatic Control*, 22, 551–575.
- Rechenberg, I. (1965, August). *Cybernetic solution path of an experimental problem*. Royal Aircraft Establishment, Library translation No. 1122, Farnborough, Hants., UK.
- Rechenberg, I. (1973). *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Stuttgart: Frommann–Holzboog Verlag.
- Rudolph, G. (1994). Convergence of non-elitist strategies. In *Proceedings of the First IEEE Conference on Computational Intelligence*, 1 (pp. 63–66). IEEE Press.
- Schwefel, H.-P. (1965). *Kybernetische Evolution als Strategie der experimentellen Forschung in der Strömungstechnik*. Diploma thesis, Technische Universität Berlin.
- Schwefel, H.-P. (1977). *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie. Interdisciplinary systems research*, 26. Basel: Birkhäuser.
- Schwefel, H.-P. (1981). *Numerical optimization of computer models*. Chichester: Wiley.
- Schwefel, H.-P. (1993, August). *Evolving evolutionary algorithms*. Invited opening lecture, SIAM Conference on Simulation and Monte Carlo Methods.
- Yin, G. (1995). Adaptive filtering with averaging. In K. J. Astrom, G. C. Goodwin, & P. R. Kumar (Eds.), *Adaptive control filtering and signal processing* (pp. 375–396). New York: Springer-Verlag.
- Yin, G., Rudolph, G., & Schwefel, H.-P. (1995). Establishing connections between evolutionary algorithms and stochastic approximation. *Informatica*, 6(1), 93–116.
- Yin, G., & Yin, K. (1994). Asymptotically optimal rate of convergence of smoothed stochastic recursive algorithms. *Stochastics and Stochastic Reports*, 47, 21–46.