Convergence of Non–Elitist Strategies

Günter Rudolph

Abstract— **This paper offers sufficient conditions to prove global convergence of non–elitist evolutionary algorithms. If these conditions can be applied they yield bounds of the convergence rate as a by–product. This is demonstrated by an example that can be calculated exactly.**

KeyWords— **global convergence, non–elitist evolutionary algorithm, martingale theory**

I. Introduction

Evolutionary algorithms (EAs) represent a class of stochastic optimization algorithms in which principles of organic evolution are regarded as rules for optimization. They are often applied to parameter optimization problems [1] when specialized techniques are not available or standard methods fail to give satisfactory answers due to multimodality, nondifferentiability or discontinuities of the problem under consideration.

In general, evolutionary algorithms may be classified as elitist or non–elitist strategies. The characteristic feature of elitist strategies is that they always maintain the best solution (individual) in the population. Examples of such EAs are $(\mu+\lambda)$–evolution strategies (ES) [2][3], evolutionary programming (EP) methods for parameter optimization [4] and elitist genetic algorithms (GA) as introduced in [5]. Whenever the support of the invariant mutation distribution covers the feasible region of the optimization problem, it is easy to prove convergence to the global optimum [6][7][8][4] for these algorithms.

For non-elitist EAs the conditions for convergence are more delicate: The standard GA as introduced in [9] does not converge at all regardless of the objective function and the choice of the crossover operator [10]. But it can be shown that a standard GA is able to generate the global solution so that the policy to keep track of the best solution found so far guarantees global convergence [10][11]. This argument may be used to prove global convergence of non–elitist (μ,λ)–ES. In principle, those algorithms may be regarded as a kind of an elitist algorithm, because the best individual maintained can be viewed as a super individual in an extended population.

Therefore, the following question will be addressed here: Is it possible to prove convergence to the optimum for a non–elitist strategy that does not make use of the policy to keep track of the best solution ?

The author is with the Universität Dortmund, Fachbereich Informatik, Lehrstuhl für Systemanalyse, 44221 Dortmund, Germany. E-mail: Rudolph@LS11.Informatik.Uni-Dortmund.DE. Support of the Research Initiative 'Parallel Computing' of the Ministry of Science and Research of Northrhine-Westphalia, Germany, under project no. 107 004 91 is gratefully acknowledged.

The answer is: Yes — under certain conditions. To make this statement rigorous, we first introduce some basic definitions and collect some results from probability theory in section II, before we derive sufficient conditions for convergence to the global optimum in section III. In section IV these conditions are applied to an example that can be calculated exactly. Finally, the strengths and weaknesses of the conditions are discussed in section V.

II. Basic Definitions and Results

Let us consider the following conceptual optimization algorithm: $s_{t+1} = \mathrm{ALG}(s_t)$, where ALG denotes an operator, depending on the algorithm under consideration, that describes the transition of the algorithm from state s_t at step t to state s_{t+1} at step $t+1$. We suppose that there exists a real–valued mapping $best(s_t)$ that extracts the best objective function value known to the algorithm in state s_t. For a probabilistic algorithm this value is a random variable, say $B_t := best(s_t)$, and we require that the sequence $(B_t)_{t\geq 0}$ converges in some mode to the global optimum $f^* = \min\{\, f(x) \;:\; x \in M \,\}$ of the real–valued objective function $f(.)$ with the feasible region M. Equivalently, we may define a random variable $D_t := B_t - f^*$ to investigate whether the stochastic sequence $(D_t)_{t\geq 0}$ converges to 0. Here it is useful to distinguish between the different modes of convergence of random sequences:

Definition 1
If $\{\, X, X_t \;:\; t \geq 0 \,\}$ are random variables on a probability space $(\Omega, \mathcal{A}, P)$, then the random sequence $(X_t)_{t\geq 0}$ is said to

(a) *converge in probability* to X, denoted $X_t \xrightarrow{P} X$, if

$$\lim_{t\to\infty} P\{\, |X_t - X| \leq \epsilon \,\} = 1 \quad \forall \epsilon > 0 \;\; ;$$

(b) *converge almost surely* to X, denoted $X_t \xrightarrow{a.s.} X$, if

$$P\{\, \lim_{t\to\infty} X_t = X \,\} = 1 \;\; ;$$

(c) *converge in mean* to X, denoted $X_t \xrightarrow{\mathcal{L}_1} X$, if

$$\lim_{t\to\infty} \mathrm{E}[\,|X_t - X|\,] = 0 \;\; .$$

□

According to Definition 1 we say that the algorithm converges in probability, almost surely (a.s.) or in mean to the global optimum if $D_t \to D$, where D is a degenerate random variable with $P\{D = 0\} = 1$. The following Lemma collects some relations between the different concepts and a sufficient condition for uniform integrability.

LEMMA 1

(a) $X_t \xrightarrow{a.s.} X$ as well as $X_t \xrightarrow{\mathcal{L}_1} X$ imply $X_t \xrightarrow{P} X$. The converse is not true in general.

(b) $X_t \xrightarrow{\mathcal{L}_1} X$ iff $X_t \xrightarrow{P} X$ and (X_t) is *uniformly integrable*, i.e.,

$$\lim_{c\to\infty} \sup_{t\geq 0} \int_{\{|X_t|>c\}} |X_t|\, dP = 0 \; .$$

(c) If $(X_t)_{t\geq 0}$ are random variables bounded in $\mathcal{L}_p$, i.e., $\sup\{\mathrm{E}|X_t|^p : t \geq 0\} < \infty$ for some $p > 1$, then the sequence $(|X_t|^q)$ is uniformly integrable for $0 < q < p$.

PROOF: For (a) see [12, pp. 33-37], for (b) and (c) see [13] p. 131 and p. 127, respectively. □

In the following we use the notation (X_t) to denote the process $(X_t)_{t\geq 0}$. The convergence condition to be derived below relies on martingale theory:

DEFINITION 2

Let $(\Omega, \mathcal{F}, P)$ be a probability space and $\{\mathcal{F}_t : t \geq 0\}$ the natural *filtration* $\mathcal{F}_t = \sigma(X_0, X_1, \ldots, X_t)$ of $\mathcal{F}$ of some stochastic process (X_t). A process (X_t) is called a *supermartingale* if $\mathrm{E}[\,|X_t|\,] < \infty$ and $\mathrm{E}[\,X_{t+1} \,|\, \mathcal{F}_t\,] \leq X_t$ a.s. for all $t \in \mathbb{N}_0$. □

A filtration $\{\mathcal{F}_t : t \geq 0\}$ is an increasing family of sun–σ–algebras of $\mathcal{F}$: $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \ldots \subseteq \mathcal{F}$. The natural filtration $\mathcal{F}_t = \sigma(X_0, X_1, \ldots, X_t)$ of a stochastic process (X_t) may be regarded as its observable history up to time t. Definition 2 states, roughly speaking, that a supermartingale decreases on average. More precisely:

THEOREM 1 ([14, p. 26])

Every nonnegative supermartingale (X_t) converges almost surely to the limit $X_\infty := \lim_{t\to\infty} X_t$. Moreover, X_∞ satisfies the inequalities

$$\mathrm{E}[\,X_\infty \,|\, \mathcal{F}_t\,] \leq X_t \quad \text{a.s.} \tag{1}$$

for all $t \in \mathbb{N}_0$. □

Note that we may not conclude that $X_t \xrightarrow{\mathcal{L}_1} X_\infty$ although (X_t) is bounded in $\mathcal{L}_1$. But it follows from Lemma 1 that a nonnegative supermartingale converges in mean to X_∞ if it is uniformly integrable. Then it remains to show that $\mathrm{E}[X_\infty] = 0$ to conclude that the algorithm converges a.s. and in mean to the global optimum.

III. CONVERGENCE CONDITION

At first, we derive a *sufficient* condition that a nonnegative supermartingale converges a.s. and in mean to a random variable X_∞.

THEOREM 2

Let (X_t) denote a nonnegative supermartingale. If $\mathrm{E}[X_t^2] < \infty$ and $\mathrm{E}[X_{t+1}^2 \,|\, \mathcal{F}_t] \leq X_t^2$ for all $t \geq 0$, then $X_t \xrightarrow{a.s.} X_\infty$ and $X_t \xrightarrow{\mathcal{L}_1} X_\infty$.

PROOF: Note that (X_t^2) is a nonnegative supermartingale. It follows that (X_t) is bounded in $\mathcal{L}_2$, since inequality (1) implies $\mathrm{E}[X_t^2] \leq \mathrm{E}[X_0^2] < \infty$. With Lemma 1(c) we may conclude that (X_t) is uniformly integrable, so that we have established convergence in mean. Almost surely convergence follows directly from Theorem 1. □

Finally, we need a condition to guarantee that the expectation of X_∞ is zero.

THEOREM 3

Let (X_t) denote a uniformly integrable nonnegative supermartingale. If $\mathrm{E}[X_{t+1} \,|\, \mathcal{F}_t] = (1-c_t)\cdot X_t$ *a.s.* with $c_t \in (0, 1]$ and

$$\sum_{t=1}^{\infty} c_t = \infty \tag{2}$$

then $X_t \xrightarrow{\mathcal{L}_1} 0$ and $X_t \xrightarrow{a.s.} 0$.

PROOF: Since (X_t) is a uniformly integrable nonnegative supermartingale it follows from Theorem 1 and Lemma 1 that X_t converges in mean and a.s. to a random variable X_∞. Therefore, we may take the expectation on both sides of the equation

$$\mathrm{E}[X_{t+1} \,|\, \mathcal{F}_t] = (1 - c_t) \cdot X_t \tag{3}$$

to obtain $\mathrm{E}[X_{t+1}] = (1 - c_t) \cdot \mathrm{E}[X_t]$. Iterated application of the expectation operator yields

$$\mathrm{E}[X_{t+1}] = \mathrm{E}[X_0] \prod_{\tau=1}^{t} (1 - c_\tau) \; .$$

Since

$$\prod_{t=1}^{\infty} (1 - c_t) \leq \exp\left(-\sum_{t=1}^{\infty} c_t\right) = 0$$

by (2) and $\mathrm{E}[X_0] < \infty$ by Definition 2 we have established $\mathrm{E}[X_\infty] = 0$. Consequently, X_t converges a.s. and in mean to 0 as $t \to \infty$. □

IV. EXAMPLE

Consider the following non–elitist $(1, \lambda)$–ES: In each iteration sample λ offspring by mutation and select the best among them to be the parent of the next generation. Note that the new parent may be worse than the old one. In particular:

```
Initialize x_0, l_0; set t = 0
repeat
    for i = 1 to λ do
        y_{t,i} = x_t · l_t · u
    endfor
    x_{t+1} = y_{t,b} with f(y_{t,b}) = min{f(y_{t,i}) : i = 1, ..., λ}
    adjust l_{t+1}
    increment t
until termination criterion satisfied
```

Here, u is a random vector uniformly distributed on a unit hypersphere surface of dimension n and l_t is the step length or the radius of the mutation hypersphere. The adjustment of l_t plays a crucial role and depends on the problem. Here, we consider the minimization of the function $f(x) = \sum_{i=1}^{n} x_i^2 = \|x\|^2$. This is not a challenging problem, because there is only one local/global optimum, but it allows an easy mathematical treatment.

This example was analyzed in [15] for an algorithm in the spirit of a $(1+1)$–ES, so that we follow their approach until equation (4).

Suppose that the algorithm has reached a point $x_t \in M$ with $f(x_t) = R^2$ at step t. A new point is sampled on the surface of a hypersphere with radius l_t. Since both the isolines of the problem and the mutation hypersphere are invariant under rotation it is feasible to analyze the projection into the plane as illustrated in Fig. 1. The large circle with radius R sketches an isoline of the problem with $f(x) = R^2$, whose center is the location of the optimum. The small circle with radius l is the surface of the hypersphere representing all possible locations accessible by mutation. For symmetry reasons we may restrict the analysis to the case with $\omega \in [0, \pi]$. The difference $R^2 - r^2$ determines whether the mutated point is worse or better than the old point. The value of r^2 depends on the angle ω and the step length l: Simple trigonometric considerations lead to $r^2 = R^2 - 2lR\cos\omega + l^2$, so that $R^2 - r^2 = 2lR\cos\omega - l^2$. For the remainder of the analysis it is useful to define the *relative improvement* V by

$$V = \frac{R^2 - r^2}{R^2} = 2a\cos\omega - a^2 \ , \tag{4}$$

where $a = l/R$. Using the notation of section II we may write $V_t = (D_t - D_{t+1})/D_t$ with $D_t = R^2$ and $D_{t+1} = r^2$. It follows that

$$D_{t+1} = D_t \cdot (1 - V_t) \tag{5}$$

and a.s. $\mathrm{E}[D_{t+1}|\mathcal{F}_t] = (1 - \mathrm{E}[V_t]) \cdot D_t$. If $\mathrm{E}[V_t] \in (0, 1]$ then the process (D_t) qualifies a nonnegative supermartingale, provided that $\mathrm{E}[D_t] < \infty$ for all $t \geq 0$.

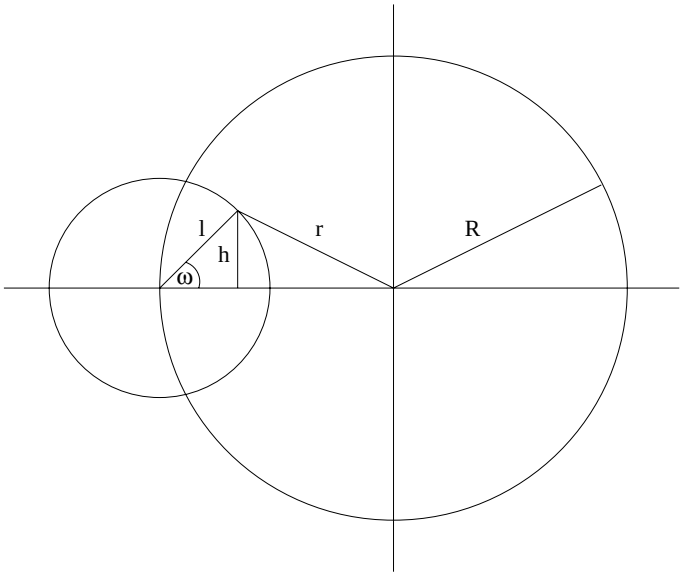

Fig. 1: A cross section of the parameter space.

The distribution of the angle ω has the density [15]

$$p(\omega) = \frac{\sin^{n-2}\omega}{B(\frac{1}{2}, \frac{n-1}{2})} \cdot 1_{[0,\pi)}(\omega) \ , \tag{6}$$

where $B(.,.)$ denotes the Beta function and $1_A(x)$ the indicator function of set A, which is the support of the distribution. To obtain the distribution of the relative improvement $V = 2a\cos\omega - a^2$ the density (6) is transformed to

$$p(v) = \frac{\left[1 - (\frac{v+a^2}{2a})^2\right]^{(n-3)/2}}{2a\, B(\frac{1}{2}, \frac{n-1}{2})} \cdot 1_{S_a}(v) \tag{7}$$

with support $S_a = [-a(2+a), a(2-a)]$. Straightforward calculations reveal that V possesses a Beta distribution:

$$P(V < x) = \int_0^{h_a(x)} \frac{(1-v)^{\frac{n-3}{2}} v^{\frac{n-3}{2}}}{B(\frac{n-1}{2}, \frac{n-1}{2})}\, dv \ , \tag{8}$$

where $h_a(x) = (x + a(2+a))/(4a)$. For $n = 3$ the distribution reduces to a special case, namely a uniform distribution with density

$$p(v) = \frac{1}{4a} \cdot 1_{S_a}(v) \ . \tag{9}$$

The $(1, \lambda)$–ES as described here samples λ times with the same distribution and selects the best sample. Therefore, we are interested in the distribution of the maximum of λ independent samples of random variable V. Let $V(\lambda)$ be this random variable. According to [16] its density is given by

$$p_{V(\lambda)}(x) = \lambda \cdot p(x) \cdot P^{\lambda-1}(V < x) \cdot 1_{S_a}(x) \ . \tag{10}$$

Using (8) with $n = 3$ and (9) in (10) the kth moment of $V(\lambda)$ is

$$\mathrm{E}[V^k(\lambda)] = \lambda \cdot \int_{-a(2+a)}^{a(2-a)} \frac{x^k}{4a} \left(\frac{x + a(2+a)}{4a}\right)^{\lambda-1} dx \ , \tag{11}$$

so that for $k = 1$ one obtains

$$E[V(\lambda)] = 2\,a \cdot \frac{\lambda - 1}{\lambda + 1} - a^2 \ . \tag{12}$$

Differentiation of (12) with respect to a leads to

$$a^* = \frac{\lambda - 1}{\lambda + 1} \tag{13}$$

resulting in a maximal relative improvement

$$E[V(\lambda)] = \left(\frac{\lambda - 1}{\lambda + 1}\right)^2 \in (0, 1) \ \text{ for } \lambda \geq 2 \ . \tag{14}$$

The optimal step length is $l^* = a^* \cdot R = a^* \cdot \|\nabla f(x_t)\|/2$, which requires the availability of the gradients $\nabla f(x)$. Approximations of the gradients, however, should be sufficient — at the expense of the convergence rate. For $\lambda = 1$ the $(1, 1)$–ES is a random walk and (13) reveals that it is optimal to stay at the current position by setting $l = 0$. Since the expectation (14) is finite and within the required range $(0, 1)$ we have shown that (D_t) is a nonnegative supermartingale converging a.s. to a random variable D_∞ as

$t \to \infty$. In order to apply Theorem 2 we must consider the squared process (D_t^2). From (5) we obtain the relation

$$D_{t+1}^2 = D_t^2 \left(1 - V_t(\lambda)\right)^2 = D_t^2 \left[1 - \left(2V_t(\lambda) - V_t^2(\lambda)\right)\right] .$$

The squared process (D_t^2) is a nonnegative supermartingale if $\mathrm{E}[2V(\lambda) - V^2(\lambda)]$ is finite and in the range $(0, 1]$. Setting $k = 2$ in (11) and using the optimal step length ratio (13) we obtain

$$\mathrm{E}[V^2(\lambda)] = \left(\frac{\lambda - 1}{\lambda + 1}\right)^2 \cdot \frac{\lambda^3 + 13\lambda + 2}{(\lambda + 1)^2(\lambda + 2)} .$$

Since $\mathrm{E}[2\,V(\lambda) - V^2(\lambda)] = 2 \cdot \mathrm{E}[V(\lambda)] - \mathrm{E}[V^2(\lambda)] =$

$$\left(\frac{\lambda - 1}{\lambda + 1}\right)^2 \cdot \left[2 - \frac{\lambda^3 + 13\lambda + 2}{(\lambda + 1)^2(\lambda + 2)}\right] \in (0, 1)$$

for $\lambda \geq 2$, the condition of Theorem 2 is satisfied and we may conclude that (D_t) also converges in mean to the random variable D_∞ as $t \to \infty$.

Finally, we apply Theorem 3 to show that $\mathrm{E}[D_\infty] = 0$ so that the $(1, \lambda)$–ES converges a.s. and in mean to the optimum: From (14) we obtain $c_t = \mathrm{E}[V_t] = const.$ for all $t \geq 0$, so that the sum (2) diverges for $\lambda \geq 2$. Moreover, we may conclude from (14) that the expected distance to the optimal objective function value converges geometrically fast.

V. Discussion

We have derived sufficient conditions for global convergence of non–elitist strategies. They are tailored for problems with feasible regions $M \subseteq \mathbb{R}^n$ and where the objective function is strictly convex at least in a neighborhood $\mathcal{N}_\epsilon(x^*) = \{x \in M : \|x - x^*\| < \epsilon\}$ around the globally optimal point $x^* \in M$ for some $\epsilon > 0$. It should be noted that former analyses of this kind summarized in [17] were incomplete (but remain true), because the uniform integrability condition was not checked. But for elitist $(\mu + \lambda)$–ES uniform integrability is guaranteed by the construction of the algorithm provided that the expectations are finite: The best point is replaced only if a better point is found, so that the worst solution is D_0. This is not the case for non–elitist strategies. Therefore, we developed Theorem 2 to ensure uniform integrability. This condition is only sufficient and might not be applicable in all cases.

Theorem 3 guarantees global convergence a.s. *and* in mean. This is a relatively strong property which might not be attainable for some algorithms. Moreover, this Theorem must be modified to cover those cases where the globally optimal points are sets of nonzero measure.

The advantage of this approach is twofold: First, if Theorem 3 can be applied to prove global convergence one also obtains the expected convergence rate. Second, it appears to be a possible route to analyze the self–adaptability properties of ES [3], i.e., on–line learning of the norm of the gradient. The latter remains for future research.

References

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

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

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

[4] D.B. Fogel. *Evolving Artificial Intelligence.* PhD thesis, University of California, San Diego, 1992.

[5] K.A. De Jong. *An analysis of the behavior of a class of genetic adaptive systems.* PhD thesis, University of Michigan, 1975.

[6] J. Born. *Evolutionsstrategien zur numerischen Lösung von Adaptationsaufgaben.* Dissertation A, Humboldt-Universität, Berlin, 1978.

[7] F.J. Solis and R.J.-B. Wets. Minimization by random search techniques. *Math. Operations Research*, 6:19–30, 1981.

[8] A.E. Eiben, E.H.L. Aarts, and K.M. Van Hee. Global convergence of genetic algorithms: A markov chain analysis. In H.-P. Schwefel and R. Männer, editors, *Parallel Problem Solving from Nature*, pages 4–12. Springer, Berlin and Heidelberg, 1991.

[9] J.H. Holland. *Adaptation in natural and artificial systems.* The University of Michigan Press, Ann Arbor, 1975.

[10] G. Rudolph. Convergence properties of canonical genetic algorithms. *IEEE Transaction on Neural Networks (special issue on Evolutionary Computation)*, 1994 (in press).

[11] J. Suzuki. A markov chain analysis on a genetic algorithm. In S. Forrest, editor, *Proceedings of the fifth International Conference on Genetic Algorithms*, pages 146–153. Morgan Kaufman, San Mateo (CA), 1993.

[12] E. Lukacs. *Stochastic Convergence.* Academic Press, New York, 2nd edition, 1975.

[13] D. Williams. *Probability with Martingales.* Cambridge University Press, Cambridge, 1991.

[14] J. Neveu. *Discrete–Parameter Martingales.* North Holland, Amsterdam and Oxford, 1975.

[15] M.A. Schumer and K. Steiglitz. Adaptive step size random search. *IEEE Transactions on Automatic Control*, 13:270–276, 1968.

[16] S.L. Resnick. *Extreme values, regular variation, and point processes.* Springer, New York, 1987.

[17] T. Bäck, G. Rudolph, and H.-P. Schwefel. Evolutionary programming and evolution strategies: Similarities and differences. In D.B. Fogel and W. Atmar, editors, *Proceedings of the 2nd Annual Conference on Evolutionary Programming*, pages 11–22. Evolutionary Programming Society, La Jolla (CA), 1993.