On The Convergence Properties of a Simple Self-Adaptive Evolutionary Algorithm

John DeLaurentis
Sandia National Laboratories
P. O. Box 5800, MS 1110
Albuquerque, NM 87185-1110
jdlauren@sandia.gov

Lauren Ferguson
Texas Technical University
alaferg@yahoo.com

William E. Hart
Sandia National Laboratories
P. O. Box 5800, MS 1110
Albuquerque, NM 87185-1110
wehart@sandia.gov

Abstract

We consider the convergence properties of self-adaptive evolutionary algorithms (EAs). The self-adaptive search component of these EAs *implicitly* adapts the step lengths in response to their efficacy for generating improving points. We analyze the convergence of a $(1, \lambda)$-EA with simpler mutation updates than are commonly used in Evolutionary Strategies or Evolutionary Programming methods. Although self-adaptive EAs have been analyzed by several authors, our analysis provides the first exact proof of convergence for an implicitly self-adaptive EA. Our experimental and theoretical analysis demonstrates that this EA robustly converges to the optimum of a symmetric, unimodal problem.

1 INTRODUCTION

The distinguishing feature of self-adaptive evolutionary algorithms (EAs) is that the control parameters are evolved by the evolutionary algorithm. This is particularly important when using EAs to optimize over continuous design spaces, since an effective search requires a search over different neighborhoods of the domain as well as refined search at different length-scales within interesting neighborhoods [6]. Thus self-adaptation is a central feature of EAs like evolutionary stategies (ES) and evolutionary programming (EP), which are applied to continuous design spaces. Eiben et al. [6] distinguishes between *explicit self-adaptation*, in which the success of previous iterations is explicitly employed to adapt the step length, and *implicit self-adaptation*, in which the the step lengths are evolved along with the search parameters. Eiben et al. [6] distinguishes these forms of self-adaptation by denoting explicit methods as *adaptive* EAs and denoting implicit methods as *self-adaptive* EAs. However, this distinction between adaptive and self-adaptive methods does not appear to have been widely adopted.

Explicitly self-adaptive EAs have been analyzed by a number of authors, and a variety of analyses have proven convergence theories for different explicitly self-adaptive formulations [1, 7, 10, 9, 11, 12, 13, 15]. By contrast, Beyer [3, 4] has developed the only theoretical investigation of implicitly self-adaptive EAs; he considers the convergence of the implicitly self-adaptive (μ, λ)-ES. Beyer notes that EAs can be described by an inhomogeneous Markovian process, and that the stochastic evolution of the system can be expressed by Chapman-Kolmogorov equations. However, he further notes that a direct treatement of these equations is generally quite difficult, and thus his analysis treats the (μ, λ)-ES as a dynamical system from which simpler dynamical systems are derived and validated.

In this paper we reconsider the convergence properties of the implicitly self-adaptive $(1, \lambda)$-ES. We simplify this EA's dynamics by considering a mutation operator that employs a discrete random variable. In this EA, there are a finite (and small) number of possible individuals that can be generated in each iteration. Consequently, the expected behavior of the EA can be characterized from one iteration to the next. Our analysis provides a convergence theory for the $(1, \lambda)$-ES for one-dimensional symmetric, unimodal objective functions. Although this is clearly an artificial class of objective functions, we hope that the techniques used in this analysis will be applicable to broader problem domains.

A complete description of our analysis is beyond the scope of this paper, so we refer the reader to DeLaurentis et al. [5] for further details. In addition to our convergence analysis, we have also identified param-

eters for which this convergence theory is practically relevant, and we provide some simple comparisons of these EAs with an $(1, \lambda)$-ES using the standard log-normal mutation operator. Finally, we describe how this convergence theory can be exploited to show how non-elitist self-adaptive $(1, \lambda)$-EAs can fail to robustly converge to globally optimal solutions. This result follows from the convergence theory that we have proven, and thus we believe that this is a broader property of implicitly self-adaptive EAs.

2 BACKGROUND

Perhaps the most common approach to the analysis of ESs is to the consider the *progress rate*, φ. Typical progress rates reflect the expected change of the ES from one iteration to the next with respect to a progress metric, which reflects the distance of a point from a given local optimum. Beyer [3] considers the progress rate of a standard self-adaptive $(1, \lambda)$-ES with log-normal mutation and concludes:

> Furthermore, applying the scaling rule $\tau = c_{1,\lambda}\sqrt{N}$ *ensures the linear convergence* of the ES algorithm. (His italics)

Beyer models the $(1, \lambda)$-ES with an approximate noisy map, and his analysis considers both first-order and second-order dynamics of this ES.

Although this analysis provides significant insight into the dynamics of this ES, it does not provide an exact convergence theory of this ES. Beyer's analysis makes a variety of nontrivial approximations to simplify the stochastic process underlying this ES. For example, the approximations that Beyer makes partially decouples the interaction between the step-length parameter and position parameter, which may fundamentally affect the convergence of these random variables; variations in the step-length parameters are modeled with normally distributed noise, which does not depend upon the position parameters. Further, Beyer's analysis does not carefully characterize the type of stochastic convergence exhibited by the approximate noisy map. Although the analysis of first- and second-order dynamics suggests that this ES has linear convergence, Beyer's results do not state this result in terms of standard forms of stochastic convergence (e.g. almost sure, in probability, etc.) [8].

Thus it is clear that an exact a convergence theory has not been previously developed for any class of implicitly self-adaptive EAs. The convergence theory that we describe considers the sequence of best points found in each iteration of an implicitly self-adaptive $(1, \lambda)$-ES, and we show that these points converge *almost surely* (i.e. with probability one). If Y and Y_t are random variables, then we say that the sequence $\{Y_t\}_{t\geq 0}$ converges almost surely to Y if $P\{\lim_{t\to\infty} Y_t = Y\} = 1$. We write this as $Y_t \xrightarrow{a.s.} Y$. See Grimmett and Stirzaker [8] for a thorough discussion of stochastic convergence.

3 ANALYSIS OF A SIMPLE ES

3.1 OVERVIEW

In this section we describe a class of self-adaptive EAs that are guaranteed to converge on one-dimensional, unimodal objective functions. We consider a simplified $(1, \lambda)$-ES that generates λ new points in each iteration and selects the best point generated for the next iteration. Figure 1 describes Algorithm A, which updates the mutation scale with the standard update rule: $\sigma_t^i = \sigma_{t-1} \cdot D_t^i$. However, this EA is distinguished by the fact that it uses a discrete random variable for D_t^i, as well as the discrete random variable, B_t^i, to generate x_t^i.

Given x_0, σ_0
For $t = 1, \ldots$
  For $i = 1 : \lambda$
    $\sigma_t^i = \sigma_{t-1} \cdot D_t^i$
    $x_t^i = x_{t-1} + \sigma_t^i \cdot B_t^i$
  End
  $j = \arg\min_{i=1:\lambda} f(x_t^i)$
  $x_t = x_t^j$
  $\sigma_t = \sigma_t^j$
End

Figure 1: Algorithm A: A self-adaptive $(1, \lambda)$-ES for one-dimensional problems. The random variables D_t^i and B_t^i are discrete random variables described in the text.

Let d_t^i be the realization of the random variable D_t^i: $D_t^i \in \{\gamma, 1, 1/\gamma\}$, $1/2 < \gamma < 1$. Let $\nu_1 = P\{D_t^i = \gamma\}$, $\nu_2 = P\{D_t^i = 1\}$ and $\nu_3 = P\{D_t^i = 1/\gamma\}$ for all t; we assume that these probabilities are nonzero. Thus $\sigma_t^i = \sigma_{t-1} d_t^i$. A step length σ_t^i is used to generate the point $x_t^i = x_{t-1} + \sigma_t^i \cdot b_t^i$, where b_t^i is the realization of the random variable B_t^i: $B_t^i \in \{-1, +1\}$ with probabilities $\{\frac{1}{2}, \frac{1}{2}\}$ respectively. Each point x_t with its corresponding step length σ_t becomes the parent point for the next iteration.

Let X_t^λ and Σ_t^λ be random variables that describe the distribution of the values of x_t and σ_t respectively when a population of size λ is used by Algorithm A. Let $\mathcal{F}_t$ be the sequence of σ-algebras that describe the random events that underly X_t^λ and Σ_t^λ. The key observation that underlies our analysis is that the convergence of X_t^λ and Σ_t^λ are complementary. For example, Σ_t^λ tends to grow when X_t^λ is far from the optimum, but in this situation X_t^λ is moving toward the optimum as much as Σ_t^λ grows. Similarly, when X_t^λ is very close to the global optimum then X_t^λ may need to move away from the optimum. But in this case, the value of Σ_t^λ is likely to decrease.

This observation led us to consider the convergence of the random variable $Z_t^\lambda = \left|X_t^\lambda\right| + W_\gamma \Sigma_t^\lambda$, where

$$W_\gamma = \frac{3\gamma + 1}{4(1-\gamma)}.$$

Our analysis applies when Algorithm A is applied to objective functions that satisfy the following assumption:

Assumption 1 *The function $f : \mathbf{R} \to \mathbf{R}$ has the property that*

1. *There exists a unique global minimum $x^* = 0$,*
2. *f is symmetric about x^* (i.e. $f(x) = f(2x^* - x)$), and*
3. *f is monotonically increasing for $x \in (x^*, \infty)$.*

Note that we assume that $x^* = 0$ only for convenience sake, since if an EA converges on a function that satisfies this condition, then we can show convergence for any other function h with nonzero global optimizer by optimizing the function $f(x) = h(x + x^*)$. Assumption 1 requires that f be unimodal, but it is quite weak otherwise. Assumption 1 does *not* require that f be continuous, and the global optimum can be at an isolated point (e.g., see Figure 2). Finally, note that since f is monotonically increasing for $x \in (x^*, \infty)$ then because of symmetry f is monotonically decreasing for $x \in (-\infty, x^*)$.

3.2 ANALYSIS

Our analysis begins by showing that Z_t^λ converges almost surely for sufficiently large λ. Given this, we show that the step lengths Σ_t^λ converge to zero, and finally we show that this implies that X_t^λ converges to x^*. A random process X_t is a super-martingale if $E[|X_t|] < \infty$ and $E[X_{t+1}|\mathcal{F}_t'] \leq X_t$, where $\mathcal{F}_t'$ is the family of σ-algebras that describe the events underlying X_t [8]. The following lemma provides the key result for our proof of Proposition 1.

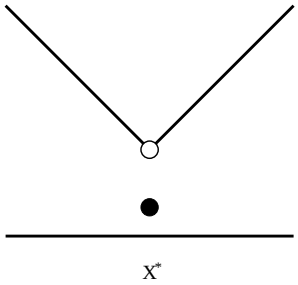


Figure 2: An example of a function satisfying Assumption 1 with an isolated global minimum.

Lemma 1 *Let $\frac{1}{2} < \gamma < 1$, and suppose that f satisfies Assumption 1. There exists $\lambda_0 > 0$ such that for all $\lambda \geq \lambda_0$, $E(Z_{t+1}^\lambda \mid \mathcal{F}_t) \leq Z_t^\lambda$.*

The following proposition shows that Z_t^λ is a super-martingale, which roughly shows that Z_t^λ decreases on average.

Proposition 1 *Let $\frac{1}{2} < \gamma < 1$, and suppose that the function f satisfies Assumption 1. Then there exists $\lambda_0 > 0$ such that for all $\lambda \geq \lambda_0$, Z_t^λ is a super-martingale with respect to the σ-algebras $\mathcal{F}_t$.*

Proof. Note that $E(Z_t^\lambda) < \infty$, since there are a finite number of states that can be reached by Algorithm A after t iterations, and Z_t^λ is finite for each of these states. From Lemma 1 we know that $E(Z_{t+1}^\lambda \mid \mathcal{F}_t) \leq Z_t^\lambda$. Together, these results show that Z_t^λ is a super-martingale with respect to $\mathcal{F}$. ∎

In the following results, we use the value λ_0 described by Proposition 1. The following corollary follows immediately from the fact that Z_t^λ is a nonnegative super-martingale.

Corollary 1 *Let $\frac{1}{2} < \gamma < 1$, and suppose that the function f satisfies Assumption 1. For all $\lambda \geq \lambda_0$, there exists a random variable Z_∞^λ such that $Z_t^\lambda \xrightarrow{a.s.} Z_\infty^\lambda$.*

Corollary 1 ensures that X_t^λ and Σ_t^λ almost surely generate a convergent sequence. This result confirms the observation noted above: X_t^λ and Σ_t^λ converge in a complementary fashion. However, this result is not sufficient to demonstrate that both of these random variables converge to zero. The following theorem uses Corollary 1 to show that Σ_t^λ converges to zero.

Theorem 1 *Let $\frac{1}{2} < \gamma < 1$, and suppose that the function f satisfies Assumption 1. For all $\lambda \geq \lambda_0$, $\Sigma_t^\lambda \xrightarrow{a.s.} 0$.*

Finally, we prove our main result: X_t^λ converges to zero almost surely. The following technical assumption is required for this analysis.

Assumption 2 *ν_1, ν_3 and λ are chosen so that*

$$1 - \left(1 - \frac{\nu_3}{2}\right)^\lambda + \left(\frac{\nu_3}{2}\right)^\lambda \geq \left(\frac{1+\nu_1}{2}\right)^\lambda - \left(\frac{1-\nu_1}{2}\right)^\lambda.$$

Theorem 2 *Let $\frac{1}{2} < \gamma < 1$, and suppose that the function f satisfies Assumption 1. Further, let ν_1, ν_3 and λ satisfy Assumption 2. There exists $\lambda_1 \geq \lambda_0$ such that for all $\lambda \geq \lambda_1$, $X_t^\lambda \xrightarrow{a.s.} 0$.*

4 PRACTICAL RELEVANCE

In this section, we consider the practical relevance of this convergence theory in cases where $\lambda \leq 5$. Note that Theorem 2 is not practically relevant when $\lambda \geq 6$. In this case, the stochastic sampling performed by Algorithm A is unnecessessary, since it is at least as efficient to simply enumerate the six possible new points that can be generated in each iteration. The following section considers the range of parameters for Algorithm A for which the convergence theory applies, particular when $\lambda = 5$. Subsequently, we consider some simple experiments that confirm that the performance of Algorithm A is roughly comparable to a standard self-adaptive ES using these parameters.

4.1 FEASIBLE PARAMETERS

We consider values of ν_i and γ for which the convergence theory for Algorithm A (in Proposition 1 and Theorems 1 and 2) applies for values of $\lambda < 6$. The following lemma makes it clear that the convergence theory does *not* apply for all possible values of ν_i.

Lemma 2 *Suppose that $\lambda \leq 5$. Then there exists $\epsilon > 0$ such that if $\nu_1 < \epsilon$ then Z_t^λ is not a super-martingale.*

Proof. To ensure that Z_t^λ is a super-martingale, we must have

$$E(Z_{t+1}^\lambda \mid \mathcal{F}_t) \leq Z_t^\lambda \tag{1}$$

for all possible values of Z_t^λ. Consider the case where Algorithm A is within $\sigma_{t-1}/(2\gamma)$ of the optimum. In this case, the six possible values of $|X_{t+1}^\lambda|$ that can be realized are worse than $|X_t^\lambda|$. Further, of the six terms in the expectation, only the terms representing contraction steps can reduce the value of Z_{t+1}^λ. Thus if Equation (1) is satisfied, the probability of the contraction steps, ν_1, cannot be arbitrarily small. But this is what we have assumed, so X_t^λ is not a super-martingale for sufficiently small values of ϵ. ∎

It is not clear that a simple analytic characterization can be developed to describe the different feasible choices for ν_i and γ. Consequently, we have numerically evaluated the set of choices for ν_i and γ for which our convergence theory is applicable. Our numerical model considers the ν_i and γ that satisfy (a) Assumption 2, (b) the constraint $\nu_1 + \nu_3 < 1$, and (c) the four main cases of the analysis in Proposition 1:

1. $|x_t + \sigma_t \gamma b| < |x_t|$,
2. $|x_t + \sigma_t b| < |x_t|$,
3. $|x_t + \sigma_t b/\gamma| < |x_t|$, and
4. $|x_t + \sigma_t \gamma b| \geq |x_t|$.

where $x_t + \sigma_t db$ is the best value of x_{t+1} possible for some $d \in \{\gamma, 1, 1/\gamma\}$ and $b \in \{-1, 1\}$. Note it is a simple matter to show that if case (4) does not hold then one of the other three cases holds. For each of these four cases we consider the expectation in Equation 1, which imposes a constraint on the values of ν_i and γ. In fact, this equation imposes several constraints on ν_i and γ, one for each possible rank-ordering of the six possible values of x_{t+1}. The details of how these constraints are derived is discussed in a technical report [5].

We enumerated feasible values of ν_i and γ with respect to these constraints. For these calculations, we assumed that no ties could occur in the rank-order of possible values of x_{t+1} (for example, this is always true if x_0 is irrational). Further, we fixed γ and sampled ν_1 and ν_3 on a fine mesh. The results of these calculations are shown in Figure 3. For values of $\gamma > 1/\sqrt{2}$ we did not find any feasible parameters in our calculations, which may simply reflect the weakness of our approximations in this case. However, if we assume that $\gamma < 1/\sqrt{2}$ then our results clearly indicate that there is a wide range of feasible parameters for Algorithm A. Further, there is significant overlap between these figures, which suggests that there may be values of ν_1 and ν_3 for which the convergence theory is valid for $0.5 < \gamma < 1/\sqrt{2}$.

4.2 EXPERIMENTAL COMPARISONS

We compared Algorithm A with a standard self-adaptive $(1, \lambda)$-ES on several simple unimodal problems. These experiments provide further confirmation

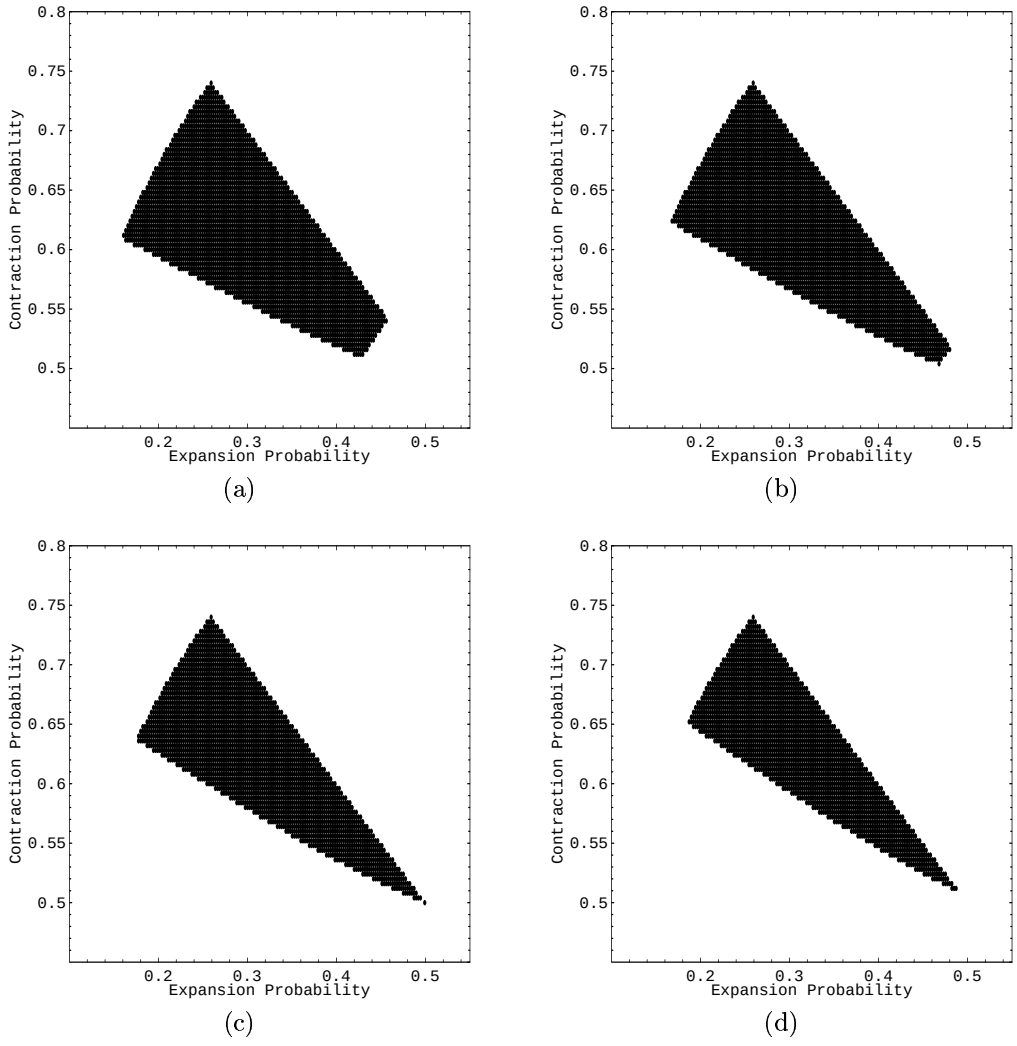


Figure 3: Domains of feasibility for ν_1 and ν_3 for (a) $\gamma = 0.55$, (b) $\gamma = 0.60$, (c) $\gamma = 0.65$ and (d) $\gamma = 0.70$. The axes are the expansion probability, ν_3, and the contraction probability, ν_1.

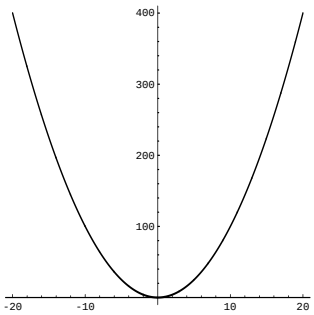
of the convergence properties of Algorithm A. Further, they demonstrate that Algorithm A can have comperable performance to standard self-adaptive EAs, despite the fact that Algorithm A uses a discrete random variables in the step-length adaptation and in the generation of mutation steps.

The following test functions were used in our experiments:

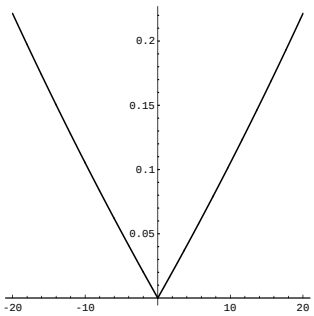
- $f_1(x) = x^2$
- $f_2(x) = e^{|x|/100} - 1$
- $f_3(x) = \sqrt{|x|}$
- $f_4(x) = 10 \lceil |x| \rceil + |x|$

All of these functions satisfy Assumption 1. The function f_1 is a standard test problem. Function f_2 becomes much steeper than f_1, which stresses some aspects of the convergence theory. Function f_3 is nonconvex, and function f_4 is discontinuous in regular intervals (and at the global optimum). These functions are illustrated in Figure 4.

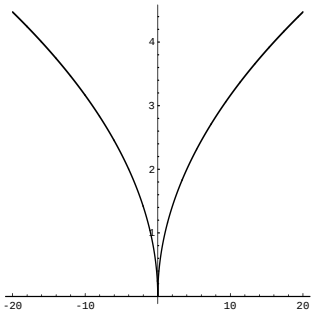
We compared Algorithm A with Algorithm B, a standard self-adaptive ES using log-normal self-adaptation of the step lengths and normally distributed mutation steps (e.g. see [2, 4]). Figure 5 describes Algorithm B; $N(0,1)$ refers to a normally distributed random variable with mean zero and variance one. Note that this algorithm is identical to Algorithm A, except for the updates to σ_t^i and x_t^i, which simply employ different random variables in place of D_t^i and B_t^i.



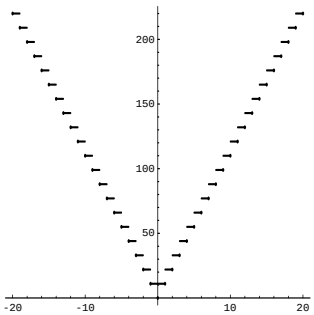
(a)



(b)



(c)



(d)

Figure 4: Graphs of functions (a) f_1, (b) f_2, (c) f_3 and (d) f_4.

Given x_0, σ_0
For $t = 1, \ldots$
  For $i = 1 : \lambda$
    $\sigma_t^i = \sigma_{t-1} \cdot \exp(N(0,1))$
    $x_t^i = x_{t-1} + \sigma_t^i \cdot N(0,1)$
  End
  $j = \arg\min_{i=1:\lambda} \ f(x_t^i)$
  $x_t = x_t^j$
  $\sigma_t = \sigma_t^j$
End

Figure 5: Algorithm B: A standard self-adaptive $(1, \lambda)$-ES for one-dimensional problems.

Both algorithms were run with 1000 different random seeds on each test function. Each optimizer has a single minimum at $x = 0$ with $f(0) = 0$. The optimizers were terminated after the function evaluation was less than 10^{-8}, except for f_4; f_4 has an isolated minimum at $x = 0$ and the function approaches 10 as x goes to zero. Thus we terminated these experiments when the optimizers found points whose function evaluation was less than $10 + 10^{-8}$. Further, we consider the behavior of these algorithms for two different initial conditions: (a) $x_0 = 10$ and $\sigma_0 = 100$ and (b) $x_0 = 1000$ and $\sigma_0 = 100$. Preliminary experiments suggested that Algorithm B worked well on these problems by effectively contracting the step length. Consequently, Algorithm A was run with $\nu_3 = 0.25$, $\nu_1 = 0.7$ and $\gamma = 0.55$.

Tables 1 and 2 compare the mean number of function evaluations before termination for these two initial conditions. Additionally, these tables show the results of a pairwise-comparison between these algorithms: for each random seed we compare the number of function evaluations needed for each algorithm and we report the percentage of trials for which Algorithm A is better.

Function	Mean Evals A	Mean Evals B	Pairwise Score
f_1	305.8	281.7	53.9
f_2	376.7	361.4	57.1
f_3	795.7	736.4	67.3
f_4	458.5	447.0	63.1

Table 1: Comparison of Algorithms A and B on the test functions when $x_0 = 1000$. The pairwise score is the percentage of trials that Algorithm A terminated after fewer function evaluations than Algorithm B.

Function	Mean Evals A	Mean Evals B	Pairwise Score
f_1	174.2	224.1	71.5
f_2	242.4	306.7	73.7
f_3	649.5	690.8	84.3
f_4	312.1	394.6	78.8

Table 2: Comparison of Algorithms A and B on the test functions when $x_0 = 10$. The pairwise score is the percentage of trials that Algorithm A terminated after fewer function evaluations than Algorithm B.

These results confirm the robustness of Algorithm A; this method found near-optimal points in every random trial of the experiments. Further, these experiments suggest that Algorithm A can perform as efficient a search as a standard self-adaptive EA. In fact, the experiments indicate that Algorithm A may me slightly more efficient, but these results are too preliminary to draw conclusions. Finally, we note that the total run-time of Algorithms A and B appears to be correlated with the slope of the function near the origin; when the slope is higher the run-time is longer. We conjecture that this reflects the fact that the step length needs to be contracted to a smaller scale to ensure convergence on narrower local minima.

5 GLOBAL CONVERGENCE

The analysis in Section 3 provides a concrete basis for expecting robust behavior from a self-adaptive $(1, \lambda)$-ES. This analysis also provides insight into the question of whether self-adaptive EAs are guaranteed (with probability one) to converge to globally optimal solutions in all cases. Rudolph [14] illustrates how self-adaptive, elitist EAs can fail to converge to globally optimal solutions with probability one. However, the EA that Rudolph considers uses an explicit self-adaptive control mechanism: whenever there is an improving mutation the step length is increased and it is decreased otherwise. By contrast, Algorithm A uses implicit self-adaptation, since the step length parameter is adapted through the evolutionary process itself.

In the following analysis, we show that there exists a function and initial conditions for which Algorithm A fails to converge to the globally optimal solution with probability one. Consider the following function:

$$g(x) = \begin{cases} |x| & , x < 10 \\ |x - 20| - 1 & , x \geq 10 \end{cases}.$$

This function is comprised of two local minima at $x = 0$ and $x = 20$. If $|x_0| < 10$ then this point is in the basin of attraction of the non-global local minima. Now if $\sigma_0 < \gamma(10 - |x_0|)$ then Algorithm A will begin searching within this basin of attraction (e.g. it cannot jump out in the first iteration). Locally, this function satisfies Assumption 1, and so we might expect that Algorithm A begins to search locally and thus misses the global optimum. The following theorem proves that this is true with some nonzero probability. Let $x^* = 20$ and recall that λ_1 is defined by Theorem 2.

Theorem 3 *Let $\{x_t\}$ be a sequence of points generated by Algorithm A on $g(x)$ starting with x_0 and σ_0 such that $|x_0| + W_\gamma \sigma_0 < 10$. If $\lambda \geq \lambda_1$ then $P(\lim_{t\to\infty} x_t = x^*) < 1$.*

Proof. Let $\overline{X}_t^\lambda$ and $\overline{\Sigma}_t^\lambda$ be the stochastic process, defined on some probability space $(\Omega, \mathcal{F}, P)$, that describes the behavior of Algorithm A on g. Consider the events in Ω for which Algorithm A converges to the global optimum: $A^* = \{\omega \in \Omega \mid \lim_{t\to\infty} \overline{X}_t^\lambda(\omega) = x^*\}$. We wish to show that $P(A^*) < 1$. Let $A_1 = \{\omega \in \Omega \mid \max_{t\geq 0} |\overline{X}_t^\lambda| \leq 10\}$, and note that $A_1^c \supseteq A^*$. Thus it suffices to show that $P(A_1) > 0$.

Now Z_t^λ is a non-negative super-martingale when Algorithm A is applied to the function $g'(x) = |x|$. Thus we can apply Kolgomorov's Theorem [8] to show that

$$P\left(\max_{t\geq 0} Z_t^\lambda > 10\right) < \frac{E(Z_0^\lambda)}{10} = \frac{|x_0| + W_\gamma \sigma_0}{10},$$

and from our assumptions we have $P(\max_{t\geq 0} Z_t^\lambda > 10) < 1 - \delta$ for some $\delta > 0$. Now consider the process $\overline{Z}_t^\lambda = |\overline{X}_t^\lambda| + W_\gamma \overline{\Sigma}_t^\lambda$ on the set of events $A_2 = \{\omega \in \Omega \mid \max_{t\geq 0} \overline{Z}_t^\lambda \leq 10\}$. For each event $\omega \in A_2$ the behavior of $Z_t^\lambda(\omega)$ is identical to $\overline{Z}_t^\lambda$. Thus we have $P(A_2) > \delta$. To conclude, note that $P(A_1) > P(A_2) > \delta > 0$. ∎

This result provides further theoretical evidence that we should not expect self-adaptive EAs to robustly perform global optimization for multimodal objective functions. This result complements Rudolph's result by considering a different form of self-adaptation. Additionally, our result applies to a *non-elitist* self-adaptive EA, and thus our result answers one of the open questions posed by Rudolph [14].

6 DISCUSSION

Rudolph [14] summarizes theoretical results concerning self-adaptive EAs and notes that the theoretical underpinnings for these methods are essentially unex-

plored. Our analysis exactly characterizes the stochastic process that underlies a class of self-adaptive $(1, \lambda)$-ESs, and we have developed a convergence theory that provides a robust guarantee that these methods converge. Our analysis is similar in spirit to Rudolph's analysis of non-elitist EAs [11]. The main difference in our work is the focus on EAs that *implicitly* self-adapt the mutation step length by coevolving these parameters within the EAs evolutionary process; our analysis appears to be the first result to exactly prove convergence for these self-adaptive EAs.

We expect that it will be difficult to prove similar results for general, nonconvex objective functions. However, we conjecture that you can relax the symmetry assumption that is made in our analysis. This assumption does not appear to be too critical for our proof that Z_t^λ is a super-martingale. However, this assumption may be more central to our proofs of Theorems 1 and 2, and this assumption is heavily exploited in our analysis of feasible values of ν_i and γ. Similarly, it should also be relatively straightforward to extend our analysis to the self-adaptive (μ, λ)-ES, which generates λ points and keeps the best μ for the next iterations. However, this may also weaken our analysis of the feasible values of ν_i and γ. In both of these cases, the more general convergence theory makes it more difficult to theoretically confirm that the convergence theory applies for small values of λ.

The use of discrete random variables in our self-adaptive EA is the key element that facilitates our analysis. This 'discrete' model of EAs on continuous design spaces is similar in spirit with our previous analyses of evolutionary pattern search methods [9, 10], where discretization of the mutation steps provides significant theoretical leverage. Although discrete self-adaptive EAs are not commonly used when optimizing continuous search domains, our empirical results suggest that these EAs may perform as well as standard self-adaptive EAs.

We believe that our result in Section 5 is quite interesting in several ways. First, this result illustrates the value of the underlying convergence theory that we have developed; our analysis is much simpler than Rudolph's analysis of explicitly self-adaptive EAs [14]. Second, this result confirms that despite their success as global optimizers, the self-adaptation in EAs limits their ability to perform global search. Thus as Rudolph notes, we should begin to focus on the global aspects of these methods. Finally, this result provides further justification for our analysis of evolutionary pattern search methods [9, 10], which focus on proving a local convegence theory. If we cannot expect that adaptation will provide global convergence, then we should ensure that it *does* provide local convergence on a wide range of objective functions.

We conclude by noting several open problems that are related to this work. First, it would be interesting to consider the progress rate for Algorithm A. We imagine that this analysis might follow directly from Beyer's analysis of the $(1, \lambda)$-ES. However, it may be possible exploit our use of discrete random variables to avoid some of the approximations that are made in that analysis

Next, the basic proof technique should be extensible to multidimensional problems. However, a significant complication is that in more than one dimension you can take a step that both increases the value of the objective function *and* the step length. Thus it appears that the basic convergence result for Z_t^λ may not be possible without further assumptions on the objective function.

Finally, a similar analysis should be possible for the self-adaptive $(1+\lambda)$-ES, which only replaces x_t in iteration t if one of the λ points generated is an improving point. This is an elitist EA, and consequently the discretized mutation steps would make it possible for this EA to get stuck near an optimum. Thus it is necessary to augment the EA to contract the step length with some fixed, nonzero probability in order to prevent it from converging to a point that is not locally optimal.

Acknowledgements

This work was supported by the DOE Office of Science, and it was performed at Sandia National Laboratories. Sandia is a multiprogram laboratory operated by Sandia corporation, a Lockheed Martin Company, for the United States Department of Energy under Contract DE-AC04-94AL85000.

References

[1] A. Agapie, *Theoretical analysis of mutation-adaptive evolutionary algorithms*, Evolutionary Computation, (2001), pp. 127–146.

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

[3] H.-G. Beyer, *Toward a Theory of Evolution Strategies: Self-Adaptation*, Evolutionary Computation, 3 (1995), pp. 311–347.

[4] ——, *The Theory of Evolution Strategies*, Springer-Verlag, 2001.

[5] J. D. DeLaurentis, L. Ferguson, and W. E. Hart, *On the convergence of an implicitly self-adaptive evolutionary algorithm: Symmetric, unimodal problems*, (2002). (submitted).

[6] A. E. Eiben, R. Hinterding, and Z. Michalewicz, *Parameter control in evolutionary algorithms*, IEEE Trans Evolutionary Computation, 3 (1999), pp. 124 – 141.

[7] G. W. Greenwood and Q. J. Zhu, *Convergence in evolutionary programs with self-adaptation*, Evolutionary Computation, (2001), pp. 147–158.

[8] G. R. Grimmett and D. R. Stirzaker, *Probability and Random Processes, Second Edition*, Oxford University Press, Oxford, 1992.

[9] W. E. Hart, *A convergence analysis of unconstrained and bound constrained evolutionary pattern search*, Evolutionary Computation, 9 (2001), pp. 1–23.

[10] ——, *Evolutionary pattern search algorithms for unconstrained and linearly constrained optimization*, IEEE Trans Evolutionary Computation, 5 (2001), pp. 388–397.

[11] G. Rudolph, *Convergence of non-elitist strategies*, in Proc of the First IEEE Conf on Evolutionary Computation, vol. 1, Piscataway, NJ, 1994, IEEE Press, pp. 63–66.

[12] ——, *Convergence Properties of Evolutionary Algorithms*, Kovač, 1997.

[13] ——, *Convergence rates of evolutionary algorithms for a class of convex objective functions*, Control and Cybernetics, 26 (1997), pp. 375–390.

[14] ——, *Self-adaptive mutations may lead to premature convergence*, IEEE Trans Evolutionary Computation, 5 (2001), pp. 410–414.

[15] G. Yin, G. Rudolph, and H.-P. Schwefel, *Analyzing* $(1, \lambda)$ *evolution strategy via stochastic approximation methods*, Evolutionary Computation, 3 (1996), pp. 473–489.