UNIVERSITY OF DORTMUND

REIHE COMPUTATIONAL INTELLIGENCE

COLLABORATIVE RESEARCH CENTER 531

Design and Management of Complex Technical Processes and Systems by means of Computational Intelligence Methods

---

Runtime Analysis of a $(\mu+1)$ ES for the Sphere Function

Jens Jägersküpper and Carsten Witt

No. CI-197/05

Technical Report   ISSN 1433-3325   May 2005

Runtime Analysis of a $(\mu{+}1)$ ES for the Sphere Function

Jens Jägersküpper[*]
FB Informatik, LS 2
Univ. Dortmund
44221 Dortmund, Germany
jens.jaegerskuepper@udo.edu

Carsten Witt[*]
FB Informatik, LS 2
Univ. Dortmund
44221 Dortmund, Germany
carsten.witt@cs.uni-dortmund.de

May 11, 2005

Abstract

Evolutionary algorithms (EAs) are general, randomized search heuristics applied successfully to optimization problems both in discrete and in continuous search spaces. In recent years, substantial progress has been made in theoretical runtime analysis of EAs, in particular for pseudo-Boolean fitness functions $f\colon \{0,1\}^n \to \mathbb{R}$. Compared to this, little is known about the runtime of simple and, in particular, more complex EAs for continuous functions $f\colon \mathbb{R}^n \to \mathbb{R}$.

In this paper, a first rigorous runtime analysis of a population-based EA in continuous search spaces is presented. A simple $(\mu{+}1)$ evolution strategy $((\mu{+}1)\,\mathrm{ES})$ that uses Gaussian mutations adapted by the 1/5-rule as its search operator is studied on the well-known SPHERE function and the influence of μ and n on its runtime is examined. By generalizing the proof technique of randomized family trees, developed before w. r. t. discrete search spaces, asymptotically upper and lower bounds on the time for the population to make a predefined progress are derived. Furthermore, the utility of the 1/5-rule in population-based evolution strategies is shown. Finally, the behavior of the $(\mu{+}1)\,\mathrm{ES}$ on multimodal functions is discussed.

1 Introduction

Since the very first successful applications of evolutionary algorithms (EAs), one has desired to explain the principles of EAs from a theoretical point of view (cf., e. g., the Schema Theorem by Holland [5] and Rechenberg's [13] analysis of

the $(1+1)$ ES). In the last decade, there has been a growing interest in theoretical runtime analysis of specific EAs on specific fitness functions and classes of functions, which is also called computational-time complexity of EAs. In this framework, it is examined by means of mathematical proofs how many evaluations of the fitness function are performed until the EA finds a global optimum. The hope is to identify practically relevant classes of functions where the EA behaves efficiently, i.e., where on average it takes only a small (polynomial) number of evaluations.

Runtime analysis started with very simple EAs such as the $(1+1)$ EA on example functions like ONEMAX (e.g., [2]). Nowadays, one is able to analyze the runtime of the $(1+1)$ EA on practically relevant problems such as the maximum matching problem [4], the minimum spanning tree problem [12], and simple scheduling problems [17]. It has turned out that from our complexity-theoretical perspective, the $(1+1)$ EA is surprisingly efficient on such problems.

Despite the successful analyses of the $(1+1)$ EA, runtime analyses should also explain the utility of the ingredients of more complex EAs, e.g., populations and variation operators. Studies of the utility of populations were performed, amongst others, for a $(1+\lambda)$ EA [9] and a $(\mu+1)$ EA [15]. The impact of a crossover operator was also investigated [14]. All these studies of more complex EAs, however, have been confined to discrete search spaces, more precisely, to pseudo-Boolean functions $f\colon \{0,1\}^n \to \mathbb{R}$. With respect to continuous search spaces, i.e., the optimization of functions $f\colon \mathbb{R}^n \to \mathbb{R}$, most of the studies are experimental. Local performance measures, namely the so-called progress rate and the so-called quality gain, have also been studied theoretically, in particular by Beyer [1]. These two measures are usually used to obtain estimates of the best possible gain of/in a single step assuming optimal adaptation. Predictions of the runtime of an EA, however, then generally rely on experimental validation since, obviously, optimal adaptation cannot be achieved/realized. Theoretical analyses of the runtime taking the adaptation into account succeeded only recently for the simple $(1+1)$ ES [7, 8]. Just like in the first runtime analyses of the $(1+1)$ EA, very simple functions such as the well-known SPHERE function and straightforward generalizations thereof are considered in these works. A study of more complex EAs in continuous search spaces should also start with such simple functions.

In this paper, a first runtime analysis of a population-based EA in continuous search spaces is presented. We define a $(\mu+1)$ ES with 1/5-rule that represents a canonical generalization of the $(1+1)$ ES with 1/5-rule. We present bounds on its runtime for the SPHERE function, depending on the population size μ and the dimensionality of the search space n. To prove the lower bounds, the technique of analyzing randomized family trees introduced by Witt [15] is generalized to continuous search spaces.

In Section 2, we formally define the $(\mu+1)$ ES and the fitness landscape induced by the SPHERE function, and in Section 3, we recapitulate basic proof techniques needed for our analyses. In Section 4, we present a general upper bound on the progress of the $(\mu+1)$ ES in the search space, implying a lower bound on the expected runtime for the SPHERE function. In Section 5, we

derive an asymptotically matching upper bound on the runtime for this function, showing that the 1/5-rule also works in our population-based ES. Moreover, the proof demonstrates that the $(\mu+1)$ ES on SPHERE quickly loses diversity, i.e., that all individuals are at almost the same distance from the optimum after a small number of steps. However, our results reveal that the population of the $(\mu+1)$ ES nevertheless has a negative impact on the runtime for the simple SPHERE function. Hence, in Section 6, we discuss more difficult functions where the $(\mu+1)$ ES could outperform the $(1+1)$ ES. We finish with some conclusions.

2 Fitness Landscape and Algorithm

As mentioned before, we will consider the well-known SPHERE function defined by $\text{SPHERE}(\boldsymbol{x}) := \sum_{i=1}^{n} x_i^2 = |\boldsymbol{x}|^2 = \boldsymbol{x}^\top \boldsymbol{I} \boldsymbol{x}$, where $\boldsymbol{I}$ denotes the identity matrix and $|\boldsymbol{x}|$ the L^2-norm of the vector $\boldsymbol{x}$, i.e. its length in Euclidean space. It can easily be seen that the results are valid also for translations of this function, i.e., $f(\boldsymbol{x}) = (\boldsymbol{x} - \boldsymbol{x}^*)^\top \boldsymbol{I} (\boldsymbol{x} - \boldsymbol{x}^*)$ for some fixed minimizer $\boldsymbol{x}^* \in \mathbb{R}^n$. Since we concentrate on the approximation error in the search space (defined as the distance from the optimizer), the results are in fact valid for any unimodal function satisfying $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n \colon |\boldsymbol{x} - \boldsymbol{x}^*| < |\boldsymbol{y} - \boldsymbol{x}^*| \implies f(\boldsymbol{x}) < f(\boldsymbol{y})$, where $\boldsymbol{x}^* \in \mathbb{R}^n$ is the unique minimizer. Note that for SPHERE, a reduction of the approximation error in the search space by an α-fraction corresponds to a reduction of the SPHERE-value by a $(2\alpha - \alpha^2)$-fraction and that $\alpha < 2\alpha - \alpha^2 < \min\{2\alpha, \alpha + 1/4\}$ for $0 < \alpha < 1$; e.g., if the distance from the optimum is halved, the SPHERE-value is reduced to 25 %.

As mentioned in the introduction, we will consider the 1/5-rule for the adaptation of the mutation strength, which was introduced by Rechenberg in the mid-1960s for the $(1+1)$ ES. The idea behind the 1/5-rule is that an isotropic mutation should result in an improvement with a probability of roughly 1/5. Therefore, the optimization is observed for $\Theta(n)$ steps. After each observation phase, the scaling factor σ for the adaptation of the length of the mutation vector (hereinafter called mutation strength) is decreased if less than 1/5 of the mutations in the respective observation phase have been successful, and otherwise, it is increased. Namely, σ is multiplied by a positive constant smaller resp. greater than 1. To keep the proofs as simple as possible, here the observation phase will last n steps and σ will be halved resp. doubled.

Commonly, in the $(\mu \overset{+}{,} \lambda)$ ES framework each individual consists of a search point and an associated mutation strength. As we are going to investigate the 1/5-rule, we additionally associate two counters, g and b, with each individual. "g" stands for the number of good/improving mutations and "b" for the number of bad mutations, where a mutation is called bad if the corresponding mutant is worse than its parent. In other words, an individual $\mathcal{X} = (\boldsymbol{x}, \sigma, g, b)$ is in $\mathbb{R}^n \times \mathbb{R}_{>0} \times \mathbb{N}_0 \times \mathbb{N}_0$.

The $(\mu+1)$ ES for minimization of $f \colon \mathbb{R}^n \to \mathbb{R}$ we consider works as follows: For a given initialization of the population of μ individuals (where all g- and b-counters are zero) an evolution loop is performed:

1. Choose one of the individuals in the (current) population uniformly at random. Let this be $\mathcal{X} = (\boldsymbol{x}, \sigma, g, b)$.

2. Create a new search point $\boldsymbol{y} := \boldsymbol{x} + \boldsymbol{m}$ with a mutation (vector) $\boldsymbol{m} := \sigma \cdot \widetilde{\boldsymbol{m}}$, where each component of the Gaussian mutation vector $\widetilde{\boldsymbol{m}} \in \mathbb{R}^n$ is independently standard-normal distributed.

3. IF $f(\boldsymbol{y}) \leq f(\boldsymbol{x})$ THEN $g := g + 1$ ELSE $b := b + 1$.

4. IF $b + g = n$ THEN BEGIN

   IF $g < n \cdot 1/5$ THEN $\sigma := \sigma/2$ ELSE $\sigma := \sigma \cdot 2$.

   $g := 0$; $b := 0$; END.

5. Create the mutant $\mathcal{Y} := (\boldsymbol{y}, \sigma, g, b)$.

6. Discard one of the $\mu + 1$ individuals by uniformly choosing one of the worst individuals (maximal f-value).

7. GOTO 1.

It is well known that a Gaussian mutation vector $\widetilde{\boldsymbol{m}}$ is isotropically distributed, and scaling by σ does not affect this property.

In practice, obviously, the GOTO is conditioned on a stopping criterion. Fortunately, for the results we are aiming at, we need not define a reasonable stopping criterion. Rather we will consider a run of the $(\mu+1)$ ES as an infinite stochastic process. We are interested in how fast (one of) the best individual(s) in the population evolves.

For the SPHERE scenario which is considered here this means that we are interested in how fast (number of function evaluations w.r.t. n, the dimensionality of the search space) the distance of the population from the optimum/origin $\boldsymbol{o}$ is reduced, where the distance of the population P from $\boldsymbol{o}$ is given as $d := \min\{|\boldsymbol{x}| \mid (\boldsymbol{x}, \cdot, \cdot, \cdot) \in P\}$. Note that for SPHERE (and all other functions for which every plateau of constant fitness has zero n-volume), the function value of the mutant generated in line 2 differs from every function value in the population with probability 1. Consequently, after (at most) μ improving mutations, there will never be two individuals of equal fitness with probability 1. As another consequence, we obtain that on the SPHERE function, the $(\mu+1)$ ES with $\mu = 1$ equals the $(1+1)$ ES as considered by Jägersküpper [7]. Hence, in this paper, the $(1+1)$ ES is defined as the $(\mu+1)$ ES with $\mu = 1$.

3 Preliminaries

In this section, some of the results for the $(1+1)$ ES on SPHERE obtained in [7] and the lower bound technique developed for the analysis of the $(\mu+1)$ EA in [15] will shortly be resumed. First, we concentrate on the effect of an isotropic mutation in the SPHERE scenario.

Let $\boldsymbol{x} \in \mathbb{R}^n - \{\boldsymbol{o}\}$ denote a search point and $\boldsymbol{m}$ an isotropic mutation. Note that $\mathrm{SPHERE}(\boldsymbol{x}) = |\boldsymbol{x}|^2$ (recall that $|\boldsymbol{x}|$ is the L^2-norm (Euclidean length) of $\boldsymbol{x}$). The analysis of the $(1+1)\,\mathrm{ES}$ for SPHERE has shown that

$$
\begin{array}{ll}
\mathsf{P}\{|\boldsymbol{x} + \boldsymbol{m}| \leq |\boldsymbol{x}| \mid |\boldsymbol{m}| = \ell\} \geq \varepsilon & \\
\text{for a constant } \varepsilon \in (0, \tfrac{1}{2}) & \Longleftrightarrow \quad \ell = O(|\boldsymbol{x}|/\sqrt{n}), \\
\text{and } n \text{ large enough} &
\end{array}
$$

i.e., under the condition that the mutation vector's length is ℓ, the mutant obtained by an isotropic mutation of $\boldsymbol{x}$ is closer to a predefined point, here the origin, with probability $\Omega(1)$ iff the length of the mutation vector is at most an $O(1/\sqrt{n})$-fraction of the distance between $\boldsymbol{x}$ and this point. On the other hand,

$$
\begin{array}{ll}
\mathsf{P}\{|\boldsymbol{x} + \boldsymbol{m}| \leq |\boldsymbol{x}| \mid |\boldsymbol{m}| = \ell\} \leq \varepsilon & \\
\text{for a constant } \varepsilon \in (0, \tfrac{1}{2}) & \Longleftrightarrow \quad \ell = \Omega(|\boldsymbol{x}|/\sqrt{n}), \\
\text{and } n \text{ large enough} &
\end{array}
$$

in other words, the mutant obtained by an isotropic mutation of $\boldsymbol{x}$ is closer to a predefined point, here again the origin, with a constant probability strictly smaller than $1/2$ iff the length of the mutation vector is at least an $\Omega(1/\sqrt{n})$-fraction of the distance between $\boldsymbol{x}$ and this point. (In the two equivalences above, the ε on the left side respectively correlates with the (multiplicative) constant concealed by the asymptotic notation on the right side.)

In the following, we concentrate on properties of $\boldsymbol{m}$ that hold with high probability. For an event E, we say that E happens even *with overwhelming probability (w. o. p.)* iff $\mathsf{P}\{E\} = 1 - 2^{-\Omega(n^\delta)}$ for some constant $\delta > 0$. The expected length of a Gaussian mutation $\boldsymbol{m}$ equals $\sigma \cdot \mathsf{E}[|\widetilde{\boldsymbol{m}}|] = \sigma \cdot \sqrt{n} \cdot (1 - (4n)^{-1} + O(n^{-2}))$ since $|\widetilde{\boldsymbol{m}}|$ is χ-distributed with n degrees of freedom. Let $\bar{\ell} := \mathsf{E}[|\boldsymbol{m}|]$. Moreover, $\mathsf{P}\{\big||\boldsymbol{m}| - \bar{\ell}\big| \geq \beta \cdot \bar{\ell}\} \leq \beta^{-2}/(2n - 1)$ for $\beta > 0$, in other words, there is only small deviation in the length of a mutation; e. g., with probability $1 - O(1/n)$ the mutation vector's actual length differs from its expected length by no more than $\pm 1\%$. As a consequence, in an observation phase of the $(\mu+1)\,\mathrm{ES}$ we expect $n - O(1)$ of the n mutations to differ only by $\pm 1\%$ from $\bar{\ell}$, respectively. By Chernoff bounds [11], we obtain that w. o. p. only an ε-fraction of the mutations in a phase deviate by more than $\pm 1\%$, respectively, where the positive constant ε can be chosen arbitrarily small.

Concerning the mutation adaptation by the $1/5$-rule for SPHERE, we know that there exists a constant $p_h \in (0, 1/5)$ such that if the success probability of the mutation in the first step of an observation phase is smaller than p_h, then w. o. p. less than $1/5$ of the steps in this phase are successful so that the scaling factor is halved. Analogously, a constant $p_d \in (1/5, 1/2)$ exists such that if the first step of a phase is successful with probability at least p_d, then w. o. p. more than $1/5$ of the steps in this phase are successful so that s is doubled. This can be used to show that the $1/5$-rule in fact "works", i. e., that each step of the phase is successful with a probability in $[a, b] \subset (0, 1/2)$ for two constants a, b. In other words, while the $(1+1)\,\mathrm{ES}$ approaches the optimum, w. o. p. for any current search point $\boldsymbol{x}$, the mutation strength fulfills $\sigma = \Theta(|\boldsymbol{x}|/n)$, and,

therefore, the expected length of the mutation vector is $\bar{\ell} = \Theta(|\boldsymbol{x}|/\sqrt{n})$. In this case, we call σ resp. $\bar{\ell}$ *optimal up to a constant factor (optimal u. t. c. f.)* w. r. t. $\boldsymbol{x}$.

Let $\Delta = |\boldsymbol{x}| - |\boldsymbol{x} + \boldsymbol{m}|$ denote the spatial gain towards the origin, the optimum of SPHERE. When the $(1+1)$ ES minimizes SPHERE (using elitist selection), a mutation is accepted iff $\Delta \geq 0$. Consequently, negative gains are zeroed out, and thus, the expected spatial gain of a step of the $(1+1)$ ES equals $\mathsf{E}\big[\Delta \cdot \mathbb{1}_{\{\Delta \geq 0\}}\big]$. Jägersküpper [7] shows that $\mathsf{E}\big[\Delta \cdot \mathbb{1}_{\{\Delta \geq 0\}} \mid \sigma = \Theta(|\boldsymbol{x}|/n)\big] = \Theta(|\boldsymbol{x}|/n)$, i. e., if the scaling factor results in $\bar{\ell} = \Theta(|\boldsymbol{x}|/\sqrt{n})$ and, thus, in a mutation to be successful with a constant probability in $(0, 1/2)$—for instance $1/5$—then the distance from the optimum is expected to decrease by a $\Theta(1/n)$-fraction in this step. Furthermore, it is shown that for any constant α

$$\mathsf{P}\{\Delta \geq \alpha \cdot |\boldsymbol{x}|/n \mid \sigma = \Theta(|\boldsymbol{x}|/n)\} = \Omega(1), \qquad (1)$$

i. e., a step of the $(1+1)$ ES decreases the distance from the optimum by an α/n-fraction with probability $\Omega(1)$ in such situations (the greater we choose α, the smaller the probability is, of course). Jägersküpper [7] proves that this result, together with the above analysis of the 1/5-rule, implies the following theorem. Recall that $|\boldsymbol{x}|$ for search points $\boldsymbol{x} \in \mathbb{R}^n$ equals the distance from the optimizer.

Theorem 1 *Consider the (1+1) ES on* SPHERE. *For the initial individual* $\mathcal{X} = (\boldsymbol{x}, \sigma, 0, 0)$, *let* $\sigma = \Theta(|\boldsymbol{x}|/n)$ *hold. Then for* $k = poly(n)$, *the time until the distance of the current search point from the optimum has been reduced to a* 2^{-k}*-fraction is* $O(kn)$ *w. o. p.*

The analyses leading to Theorem 1 exploit that the distance from the optimum is non-increasing in the run of the $(1+1)$ ES. In the population of the $(\mu+1)$ ES, however, it can happen that a mutation is accepted even if the mutant is farther away from the optimum than its parent. Hence, the analysis of the $(1+1)$ ES on SPHERE cannot be simply transferred to the $(\mu+1)$ ES, yet a new, more sophisticated argumentation is required.

In the following, the lower-bound technique obtained for the $(\mu+1)$ EA will be presented. The technique is based on so-called *family trees*, which are undirected trees labeled with individuals. Let $\mathcal{X}$ be some individual from the initial population, i. e., at time 0, of the $(\mu+1)$ ES. The family tree $T_0(\mathcal{X})$ of $\mathcal{X}$ at time 0 contains only the root node labeled with $\mathcal{X}$. The family tree $T_t(\mathcal{X})$ at time $t > 0$ contains the tree $T_{t-1}(\mathcal{X})$ as a subtree and at most one additional node. If one of the individuals corresponding to nodes from $T_{t-1}(\mathcal{X})$ is chosen for mutation at time $t-1$, the corresponding node is connected to a new node, which is labeled with the new individual created by the mutation; otherwise, $T_t(\mathcal{X}) := T_{t-1}(\mathcal{X})$. Hence, $T_t(\mathcal{X})$ contains all descendants of $\mathcal{X}$ at time t. Moreover, if $\mathcal{Y}$ is contained in $T_t(\mathcal{X})$, the path leading from the root node to $\mathcal{Y}$ models the history of mutations that lead from $\mathcal{X}$ to $\mathcal{Y}$ via a set of intermediate individuals.

Intuitively, short paths in family trees do not contain much evolutionary progress. If all family trees at some given time t have a low depth, i. e., contain

6

only short paths, and the initial population is far away from the optimum, so will be the population at time t. This suggests a way of lower bounding the runtime of the $(\mu+1)$ ES. The main observation by Witt [15] states that it is possible to bound the depth of any family tree $T_t(\mathcal{X})$ irrespective of the fitness function and of the search space. This is made precise in the following lemma.

Lemma 1 *Let $\mathcal{X}$ be some individual from the initial population of the $(\mu+1)$ ES and let $D(t)$ denote the depth of the family tree $T_t(\mathcal{X})$. Then $\mathrm{P}\{D(t) \geq 3t/\mu\} \leq (e/3)^{t/\mu} = 2^{-\Omega(t/\mu)}$.*

The proof of Lemma 1 can be found in a technical report by Witt [16], which is a complete presentation of the results in [15]. The reader can easily verify that all arguments referring to family trees of the $(\mu+1)$ EA remain valid for the $(\mu+1)$ ES considered here.

According to Lemma 1, it is very unlikely that a family tree at time t is deeper than $3t/\mu$. The general lower bound technique proceeds in the following way. Suppose that E is some event epitomizing a success of the EA, e. g., the event of the $(\mu+1)$ ES halving the distance to the optimum. To estimate the probability of E not occurring within t steps, we first bound the probability of a family tree becoming deeper than $3t/\mu$. Afterwards, we consider arbitrary paths of length at most $3t/\mu$ in family trees. We call such paths short. If t is chosen appropriately, we would like to show that with a high probability, E does not occur on any short path in a family tree by time t. The main advantage is that paths may be studied in isolation to show the claim. However, it is still a challenge to show that E does not occur on a given short path. Obviously, we must study the total progress made by the mutation operator of the $(\mu+1)$ ES within $3t/\mu$ steps. This is a new contribution, described in the lower bound of Theorem 2 in Section 4.

4   Lower Bound on the Runtime

In this section, we will show a general lower bound on the time for the $(\mu+1)$ ES to attain a predefined spatial gain. With respect to the SPHERE function, the following theorem immediately implies the same bound on the time to decrease the function value by a constant fraction.

Theorem 2 *Let the $(\mu+1)$ ES, where $2 \leq \mu = poly(n)$, optimize an arbitrary function f with a unique optimizer $\boldsymbol{x}^* \in \mathbb{R}^n$. Then the number of steps until the distance to $\boldsymbol{x}^*$ has been halved is $\Omega(\mu n)$ with probability $1 - 2^{-\Omega(n)}$ and also in expectation.*

Proof: We use the lower bound technique described in Section 3. In the following, we speak of a success when an individual is created that has halved the distance to $\boldsymbol{x}^*$ with respect to the initial distance. Let $s := \lfloor c\mu n \rfloor$ for some positive constant c to be chosen appropriately later. Our aim is to show that the

success probability in the phase consisting of the first s steps of the $(\mu+1)\,\mathrm{ES}$ is $2^{-\Omega(n)}$ if c is chosen small enough.

According to Lemma 1, in the phase, any fixed family tree reaches depth at most $3cn$ with a probability of $1 - 2^{-\Omega(n)}$. Since μ is a polynomial, this also holds for all μ family trees together. We study the probability of observing in a fixed tree a success at a node of depth at most $3cn$. If we can show that this probability is bounded above by $2^{-\Omega(n)}$ for small enough c, the theorem will follow since the number of trees is a polynomial.

Let $\mathcal{X}$ be some arbitrary initial individual and let p be an arbitrary path of length $\ell \leq 3cn$ and starting at the root node in the family tree $T_s(\mathcal{X})$. Now, given some mutation strength σ, consider a sequence of search points $\boldsymbol{x}_0, \ldots, \boldsymbol{x}_\ell$ such that $\boldsymbol{x}_i$, $1 \leq i \leq \ell$, is the result of a mutation of $\boldsymbol{x}_{i-1}$, performed independently of other steps and at mutation strength σ. Let such a sequence of search points be called an *independent-mutation sequence.* The idea is to relate an independent-mutation sequence to the sequence of individuals labeling the nodes along p. However, the sequence of mutations performed along p is not independent since we impose conditions on the path, namely by considering the event that the path emerges in $T_s(\mathcal{X})$. The event that a path emerges is influenced by the f-values of the individuals on the path since the $(\mu+1)\,\mathrm{ES}$ deletes individuals with worst f-value. Moreover, only for short paths of length less than n we can guarantee that no adaptation of the mutation strength takes place along the path.

By treating the (search points of the) individuals along p as an independent-mutation sequence, we introduce an error. This error can be handled in the following way. We modify the $(\mu+1)\,\mathrm{ES}$ such that

- Step 6 is omitted; hence, the population grows in each step;

- if the size of the current population before Step 1 equals k, Steps 1–5 are executed $\lceil 2k/\mu \rceil$ times in parallel in one iteration of the evolution loop.

Let this new, artificial algorithm be called $(\mu+1)\,\mathrm{ES}^*$. Since it creates $\lceil 2k/\mu \rceil$ new individuals in one iteration of the evolution loop, the size $r(t)$ of its population at time t is exponentially growing with respect to t. It is easy to see that $\mu(1 + 2/\mu)^t \leq r(t) \leq \mu(1 + 3/\mu)^t$. In the following, we assume the $(\mu+1)\,\mathrm{ES}^*$ to start with the same population as the $(\mu+1)\,\mathrm{ES}$.

This $(\mu+1)\,\mathrm{ES}^*$ has two crucial properties. Firstly, consider paths in family trees of individuals of the $(\mu+1)\,\mathrm{ES}^*$ starting at the root and having length less than n. Such paths have the property of containing independent-mutation sequences of individuals. This holds since the fitness function influences neither the selection for reproduction nor the selection for replacement and no adaptation of the mutation strength takes place along the path. Secondly, if q denotes the probability that a path labeled with some fixed sequence of search points exists in $T_t(\mathcal{X})$, then such a path exists in the family tree $T_t^*(\mathcal{X})$ of the $(\mu+1)\,\mathrm{ES}^*$ at time t with probability at least q. The reason is that in each iteration of the evolution loop of the $(\mu+1)\,\mathrm{ES}^*$, each individual from the current population is chosen with probability at least $1/\mu$. This holds since one iteration consists of

at least $2k/\mu$ uniform choices from the population of size k and the probability of choosing an individual is bounded below by

$$1 - \left(1 - \frac{1}{k}\right)^{2k/\mu} \geq 1 - e^{-2/\mu} \geq \frac{1}{\mu}$$

since $1 - e^{-x} \geq x/2$ holds for $x \leq 1$ and by assumption $\mu \geq 2$. Therefore, the probability $q^*(s)$ of a success at depth at most $3cn$ in $T_s^*(\mathcal{X})$ upper bounds the probability of such a success in $T_s(\mathcal{X})$.

In Lemma 2, we will show that any independent-mutation sequence of individuals that starts with $\mathcal{X}$ (and may have even arbitrary length) leads to a success with a probability of at most $2^{-c'n}$ for some constant $c' > 0$. Since $T_s^*(\mathcal{X})$ contains at most $\mu(1 + 3/\mu)^s$ paths starting at the root, we obtain

$$q^*(s) \ \leq \ \mu \cdot \left(1 + \frac{3}{\mu}\right)^{c\mu n} \cdot 2^{-c'n} \ \leq \ \mathrm{poly}(n) \cdot e^{3cn} \cdot 2^{-c'n},$$

which is $2^{-\Omega(n)}$ if c is chosen small enough. This proves the theorem. $\qquad\square$

We are left with the claim on the success probability for the so-called independent-mutation sequences as mentioned in the proof of Theorem 2. This claim is stated as the forthcoming Lemma 2, where without loss of generality the unique optimizer $\boldsymbol{x}^*$ is assumed to equal the origin. The proof of the lemma relies on the following observation.

Fact 1 Let $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_k$ be independently (not necessarily identically) isotropically distributed vectors. Then the vector $\boldsymbol{m} := \boldsymbol{m}_1 + \cdots + \boldsymbol{m}_k$ is isotropically distributed.

Lemma 2 *Let $\boldsymbol{m} \in \mathbb{R}^n$ be isotropically distributed and $\boldsymbol{s} \in \mathbb{R}^n$ be fixed. Then $\mathsf{P}\{|\boldsymbol{s} + \boldsymbol{m}| \leq 0.841\,|\boldsymbol{s}|\} = 2^{-\Omega(n)}$.*

Proof: Let $d := |\boldsymbol{s}|$ and $z := 0.841d$. Furthermore, let $r := |\boldsymbol{m}|$ be the length of $\boldsymbol{m}$. Henceforth, we assume r to take some fixed value maximizing $\mathsf{P}\{|\boldsymbol{s} + \boldsymbol{m}| \leq z \,|\, |\boldsymbol{m}| = r\}$. According to the law of total probability, this leads to an upper bound on $\mathsf{P}\{|\boldsymbol{s} + \boldsymbol{m}| \leq z\}$.

Let $S_r \subseteq \mathbb{R}^n$ be the hypersphere with radius r centered at $\boldsymbol{s}$ and let $S_z \subseteq \mathbb{R}^n$ be the hypersphere with radius z centered at the origin $\boldsymbol{o}$. Moreover, let $S_{\leq z}$ denote the ball with hypersurface S_z. Let $C := S_r \cap S_{\leq z}$ denote the hyperspherical cap cut off from S_r by S_z. Obviously, $|\boldsymbol{s} + \boldsymbol{m}| \leq z$ iff $\boldsymbol{s} + \boldsymbol{m} \in C$. Let $A(C)$ and $A(S_r)$ denote the hypersurface areas of C resp. S_r. Since $\boldsymbol{m}$ is isotropically distributed, $\mathsf{P}\{|\boldsymbol{s} + \boldsymbol{m}| \leq z\} = A(C)/A(S_r)$, and hence, we need to estimate $A(C)$. Therefore, let $R := S_r \cap S_z$ denote the boundary of the cap C.

Let L denote the line segment connecting $\boldsymbol{o}$ to $\boldsymbol{s}$ and let $\boldsymbol{c} := L \cap C$ (see Fig. 1). Then $\boldsymbol{c} \in C$ is the center of the cap C. Note that all points in R, the boundary of the cap, have equal distance from L. Let h denote this distance and let $\alpha := \arcsin(h/r)$. Then $r\alpha$ is the distance of $\boldsymbol{c}$ from R, the boundary of
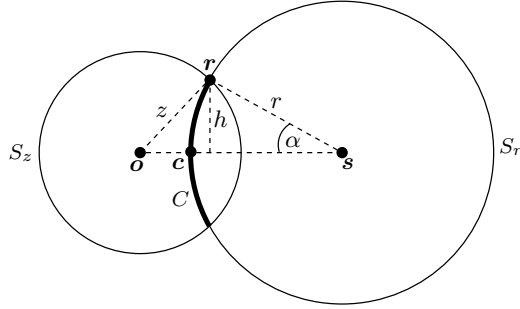
Figure 1: Illustration of spherical cap

the cap, within the $(n-1)$-space C, namely the spherical distance w. r. t. the hypersphere $S_r \supset C$.

If we can show that $A(C)$ is bounded above by the volume of an $(n-1)$-dimensional ball of radius αr, we can apply standard formulas for hypersurface areas and volumes [3] and obtain (where "Γ" denotes the well-known Gamma function):

$$
\begin{aligned}
\frac{A(C)}{A(S_r)} &\leq \frac{\pi^{\frac{n-1}{2}} \cdot (\alpha \cdot r)^{n-1}}{\Gamma\left(\frac{n-1}{2}+1\right)} \Big/ \frac{n \cdot \pi^{\frac{n}{2}} \cdot r^{n-1}}{\Gamma\left(\frac{n}{2}+1\right)} \\
&= \frac{\pi^{\frac{n-1}{2}} \cdot (\alpha \cdot r)^{n-1} \cdot \Gamma\left(\frac{n}{2}+1\right)}{n \cdot \sqrt{\pi} \cdot \pi^{\frac{n-1}{2}} \cdot r^{n-1} \cdot \Gamma\left(\frac{n-1}{2}+1\right)} \\
&= \frac{\alpha^{n-1} \cdot \Gamma(n/2+1)}{n \cdot \sqrt{\pi} \cdot \Gamma(n/2+1/2)} \\
&= \frac{\alpha^{n-1}}{n\sqrt{\pi}} \Theta(\sqrt{n}) \\
&= \frac{\alpha^{n-1}}{\Theta(\sqrt{n})}.
\end{aligned}
$$

Hence, $\alpha \leq 1 - \varepsilon$ for some constant $\varepsilon > 0$ will imply that $\mathsf{P}\{|\boldsymbol{s}+\boldsymbol{m}| \leq z\} = 2^{-\Omega(n)}$. To show that α is indeed bounded away from 1, we have to estimate h/r. A geometric argument (in the next paragraph) yields $h/r \leq z/d$. Since $z/d = 0.841$ and $h, r > 0$, $\arcsin(h/r) \leq \arcsin(0.841) < 0.9992$. This will imply the lemma.

To show $h/r \leq z/d$, let $\boldsymbol{r} \in R$ be some point in the boundary of the cap and observe that the area of the triangle defined by $\boldsymbol{o}, \boldsymbol{s}, \boldsymbol{r}$ is bounded above by half the area of a rectangle with edges of length z and r. Since the area of the triangle equals $hd/2$, we obtain $hd/2 \leq zr/2$, implying the claimed inequality since $d, r > 0$.

We still have to show that the hypersurface area $A(C)$ can be bounded by the volume of an $(n-1)$-dimensional ball of radius αr. Intuitively, we are confronted with the error that is introduced by mapping the area of a part of a

10

sphere (e. g., the area of a continent) onto a plane (e. g., a map of the continent). Of course, the larger the area compared to the sphere, the greater the error is. Formally, for $\alpha \leq \pi/2$, the hypersurface area of C can be expressed as

$$A(C) = r^{n-1} \cdot 2\pi \cdot \int_0^\alpha (\sin\beta)^{n-2} \mathrm{d}\beta \cdot \prod_{i=1}^{n-3} \int_0^\pi (\sin\beta)^i \mathrm{d}\beta$$

(see [6] for a derivation). Since $\sin\beta \leq \beta$ holds for $\beta \geq 0$, we obtain, by estimating the first integral, that

$$A(C) \leq \frac{2\pi}{n-1} \cdot (\alpha r)^{n-1} \prod_{i=1}^{n-3} \int_0^\pi (\sin\beta)^i \mathrm{d}\beta.$$

The last expression is the anti-derivative of the hypersurface area of an $(n-1)$-dimensional ball of radius αr, i. e., the volume of this ball. $\qquad \square$

Theorem 2 shows in conjunction with Theorem 1 that w. o. p., the $(\mu{+}1)\,\mathrm{ES}$ on SPHERE is asymptotically no more efficient than the $(1{+}1)\,\mathrm{ES}$. Although this result might have been suspected, Theorem 2 has been the first formal proof analyzing the runtime of the $(\mu{+}1)\,\mathrm{ES}$. Moreover, there seem to exist more complex, multimodal functions where the population of the $(\mu{+}1)\,\mathrm{ES}$ leads to an advantage in runtime compared to the $(1{+}1)\,\mathrm{ES}$. This will be discussed in Section 6.

5 Upper Bound on the Runtime and the Diversity

In this section, we are going to prove an upper bound on the time for the $(\mu{+}1)\,\mathrm{ES}$ on SPHERE to halve the distance from the optimum. This bound matches the lower bound from the previous section. By this upper bound, we rigorously show that the 1/5-rule works in a population-based EA, too. Furthermore, we will show that the *bandwidth* of the population, which we define as the ratio of the largest to the smallest distance from the optimum represented in the population, drops w. o. p. below 1.2 after a short period and remains smaller than 2 afterwards, which can be interpreted as a loss of diversity.

Obviously, efficient runtimes cannot be obtained if many individuals from the initial population have wrong mutation strength. In the following theorem, we therefore assume all initial mutation strengths to be optimal u. t. c. f., yet we do not put an assumption on the bandwidth of the initial population.

Theorem 3 *Consider the $(\mu{+}1)\,ES$ on* SPHERE*, where $2 \leq \mu = poly(n)$. For all initial individuals $\mathcal{X} = (\boldsymbol{x}, \sigma, 0, 0)$, let $\sigma = \Theta(|\boldsymbol{x}|/n)$ hold. Then after some number of steps that is $O(\mu n)$ w. o. p.*

1. the population's bandwidth drops below 1.1891*,*

11

2. the mutation strength of each individual in this population is optimal u. t. c. f.,

3. the population's bandwidth remains smaller than 2 for any number of subsequent steps that is $O(\mu n)$.

Furthermore,

4. the runtime to halve the distance from the optimum is $O(\mu n)$ w. o. p.

Proof: Let d_t be the distance from the optimum for the best individual of the population at time t, i. e., d_0 denotes the distance from the optimum of the initial population. We partition the search space into four regions. Region 0 contains all search points at distance at most $0.841d_0$ from the origin, Region 1 all search points having a distance in $(0.841d_0; d_0]$, Region 2 all search points with a distance in $(d_0; d_0/0.841]$, and Region 3 contains all remaining search points. Our goal is to show that after some phase of $O(\mu n)$ steps, w. o. p. all individuals are in Region 1 and none of them has an ancestor from Region 3. At this point, w. o. p. every individual of the current population is in Region 1 (and has only Region-1 and/or Region-2 ancestors) so that the bandwidth has become smaller than $1/0.841 < 1.1891$.

Consider the phase consisting of the first $s := c\mu n$ steps as in the proof of Theorem 2. Let S_t denote the subpopulation containing exactly those individuals of the population at time $t \geq 0$ that are in Region 3 or have an ancestor in Region 3. (Note that the subpopulation's size is non-increasing and that there is nothing to show if $\#S_0 = 0$.) We want to estimate the progress of this subpopulation by means of the general lower bound technique from Theorem 2. Consider the family tree rooted at some $\mathcal{X} = (\boldsymbol{x}, \cdot, \cdot, \cdot) \in S_0$. Then the proof of the lower bound tells us that w. o. p. no individual in this tree after s steps is in Region 0 or Region 1 since such an individual would have to have a distance from the optimum smaller than $0.841 |\boldsymbol{x}|$. Now let $\mathcal{X}$ denote an individual in $S_t - S_{t-1}$, i. e. a mutant of a non-Region-3 individual that hit Region 3 and was accepted, then the same argumentation as above applies to the individuals of the family tree rooted at $\mathcal{X}$ since the number of steps is less than s. All in all, we obtain that w. o. p. no individual in $S_\cup := S_0 \cup \cdots \cup S_s$ is in Region 0 $\cup$ Region 1. By the very same arguments, we can show that w. o. p. no initial individual (all of which are at distance at least d_0) produces a descendant from Region 0 within this phase of s steps. Next we will show that, on the other hand, after at most s steps w. o. p. the complete population has entered Region 1 (implying that w. o. p. $\#S_s = 0$). This will imply the first claim.

To show this, first recall that w. o. p. there will be no Region-0 individual after the s steps. Consider the subpopulation B_t exactly consisting of those individuals in the population at time t that have distance at most d_0 from the optimum, i. e., B_t is the part of the population that is located in Region 1 or even in Region 0. Consequently, B_0 contains all initially optimal individuals, and obviously, the subpopulation's size is non-decreasing so that $\#B_t = \mu$ implies that the complete population has made it into Region 1 $\cup$ Region 0 by time t.

Let t^* be the smallest t such that $\#B_t = \mu$. Then the bandwidth of $B_\cup :=$ $B_0 \cup \cdots \cup B_{t^*}$ is less than 1.1891 if there is no Region-0 individual in $B_\cup$—which is the case w. o. p. if $t^* \leq s$. Assume $t^* \leq s$. We are going to show that this assumption holds w. o. p. Note that, by definition of B_t and the preceding argumentation on S_t, w. o. p. $B_t \cap S_\cup = \emptyset$ for $t \leq t^*$ (as $t^* \leq s$). This means that no $\mathcal{X} = (\boldsymbol{x}, \sigma, \cdot, \cdot) \in B_t$ has a Region-3 ancestor, and since no adaptation takes place in the first s steps (cf. proof of Theorem 2), σ is optimal u. t. c. f. because $0.841 d_0 < |\boldsymbol{x}| \leq d_0$ and $\sigma = \Theta(d_0/n)$ (since $\mathcal{X}$ does not stem from a Region-3 individual). Moreover, this proves the theorem's second claim.

Therefore, w. o. p. for all $\mathcal{X} = (\boldsymbol{x}, \sigma, \cdot, \cdot) \in B_\cup$ we have $\sigma = \Theta(|\boldsymbol{x}|/n)$ respectively, i. e., σ (and with it the expected length of a mutation) is optimal u. t. c. f. Hence, we assume in the following that σ is in fact optimal u. t. c. f. whenever a $B_\cup$-individual is mutated. Then by the results from [7] restated in the Preliminaries, each mutation of a $B_\cup$-individual leads to a better individual with probability at least c' for some constant c'. Pessimistically assuming that we start with $\#B_0 = 1$, after at most μ/c' expected steps, the number of B-individuals increases from one to two. Afterwards, the probability of increasing the number of B-individuals is at least $2c'/\mu$. The length of the phase is bounded by the time that all individuals are B-individuals. The expectation of this time is bounded above by

$$\sum_{i=1}^{\mu} \frac{\mu}{ic'} \leq c''\mu \log \mu$$

for some constant c''. By Markov's inequality, the time is bounded above by $2c''\mu \log \mu$ with probability at least $1/2$. In case of a failure, we can repeat the argumentation with another phase of at most $2c''\mu \log \mu$ steps. Therefore, $t^* \leq 2kc''\mu \log \mu$ with probability at least $1 - 2^{-k}$. Since $\mu = \text{poly}(n)$, choosing $k := c'''n/\log n$ with some appropriate constant c''' yields that $t^* \leq c\mu n = s$ w. o. p., namely with probability $1 - 2^{-\Omega(n/\log n)}$.

For the proof of the theorem's third claim, naturally, we would like to iterate the argumentation from above. Unfortunately, there are two problems. The first one is that in the beginning of a phase of (w. o. p. at most) s steps we need all σs in the current population to be optimal u. t. c. f. Since this holds after the first such phase, this will also be the case w. o. p. after a constant number of phases, i. e. after a number of steps that is $O(\mu n)$. The second problem is that in the next (w. o. p. at most) s steps (following after the first t^* steps), we may observe (w. o. p. at most) one adaptation within each path growing in these s steps in the family trees. Hence, the argumentation in the proof of Theorem 2 must be split into two phases of independent-mutation sequences so that we obtain a factor of 0.841^2 instead. Let t^{**} be the first point of time when every individual in the population has distance at most d_{t^*} from the optimum. Then w. o. p. $t^{**} \leq t^* + s$, and if so, w. o. p. $d_{t^{**}} > 0.841^2 d_{t^*}$ so that the bandwidth of the population at time t^{**} is smaller than $0.841^{-2} < 1.414$. Since w. o. p. after time t^* all individuals have distance at most $0.841^{-2}d_{t^*}$ from the optimum, the total bandwidth of all populations between t^* and t^{**} is w. o. p. smaller than

13

$0.841^{-4} < 2$. Since w. o. p. there is always at most one adaptation within such paths growing in at most s steps, we can now iterate this argumentation a constant number of times.

Finally, we are going to prove the theorem's fourth claim, the upper bound on the runtime. Therefore, we pessimistically assume that $d_{t*} > d_0/2$, i. e., that the distance has not been halved already in the phase when the population's bandwidth collapses. Now, by the results on the bandwidth, we are able to show that the 1/5-rule also works for the $(\mu+1)$ ES. We know that by time t^* w. o. p. there is no individual $\mathcal{X} = (\boldsymbol{x}, \cdot, \cdot, \cdot)$ with $|\boldsymbol{x}| > d_0$ and that there will never be such an individual in the future.

We call the set of individuals $H_k(\mathcal{X})$ consisting of the k (most recent) ancestors of an individual $\mathcal{X}$ the *history line* of order k of individual $\mathcal{X}$. If the order is omitted, we refer to the complete history line. Note that a history line corresponds to a path in the family tree containing $\mathcal{X}$.

Consider a step $t = \text{poly}(n)$ in which the mutation strength of the individual $\mathcal{X} = (\boldsymbol{x}, \sigma', 0, 0)$ has just been adapted. Then any individual in $H_n(\mathcal{X})$ has been using the mutation strength $\sigma := 2\sigma'$ if the mutation strength has just been halved, or the mutation strength $\sigma := \sigma'/2$, otherwise. Note that the bandwith of $H_n(\mathcal{X})$ is smaller than 2 (unless the distance has just been halved and we are done). Consequently, since all n individuals in this history line use mutation strength σ, the success probabilities that correspond to these n individuals differ (at most) by a constant factor (cf. the results recapitulated in the Preliminaries); namely, if the greatest success probability were, say, 0.1 then the smallest one is lower bounded by $0.1/\kappa_1$, and if the smallest success probability were, say, $1/2 - 0.2 = 0.3$ then the greatest one is upper bounded by $1/2 - 0.2/\kappa_2$ for two constants $\kappa_1, \kappa_2 \geq 1$. Note that the expected number of successful mutations is upper bounded by $0.1n$ in the first case and lower bounded by $0.3n$ in the second case. By Chernoff bounds, the number of successful mutations is smaller resp. larger than $0.2n$ w. o. p. so that the mutation strength is halved resp. doubled. Thus, to prove that the 1/5-rule works, it suffices to show that halving the mutation strength in fact results in a greater success probability for $\mathcal{X}$ compared to any individual in $H_n(\mathcal{X})$ resp. that doubling results in a smaller success probability (the 1/5-rule does not necessarily fail if this is not the case; we consider this condition to keep things simple.) This happens in the first case if the largest distance from the optimum of an individual in $H_n(\mathcal{X})$ is less than $2|\boldsymbol{x}|$, and in the second case if the smallest distance of an individual in $H_n(\mathcal{X})$ is larger than $|\boldsymbol{x}|/2$. Since we know that $|\boldsymbol{x}| \in (d_0/2; d_0]$ (unless the population has just halved the distance from the optimum and we are done), this is indeed the case. Consequently, w. o. p. the 1/5-rule is able to keep the success probabilities in $[0.1/\kappa_1; 1/2 - 0.2/\kappa_2]$ (and with them mutation strengths that are optimal u. t. c. f., respecticely). All in all, we have shown that, if the success probabilities of all individuals at time t^* are in the interval $[0.1; 0.3]$, then they will be in $[0.1/\kappa_1; 1/2 - 0.2/\kappa_2]$ w. o. p. for all populations until the distance is halved. If the smallest success probability represented in the population at time t^* is smaller than 0.1, however, we would use this probability in the above

argumentation instead; correspondingly, if the largest one (which is strictly smaller than $1/2$) is larger than 0.3, we would use this one instead.

Finally, we show that in the situation just described, the (expected) progress of the population is as desired. Due to the properties of the success probabilities in our situation, within a polynomial number of steps w. o. p. all mutations will use a σ that is optimal u. t. c. f. until the distance is halved. We assume this to happen. Then, by the results from [7] mentioned in the Preliminaries, with probability $\Omega(1)$, a mutation choosing some individual $\mathcal{X} = (\boldsymbol{x}, \cdot, \cdot, \cdot)$ will result in a mutant $\mathcal{Y} = (\boldsymbol{y}, \cdot, \cdot, \cdot)$ such that we have $|\boldsymbol{y}| \leq |\boldsymbol{x}|\,(1 - \Omega(1/n))$. If the distance of $\mathcal{X}$ from the optimum is currently best, the best distance is improved by a factor $1 - \Omega(1/n)$ with probability $\Omega(1)$. Since a best individual is chosen with probability at least $1/\mu$, each step decreases the best distance by an $\Omega(1/n)$-fraction with probability $\Omega(1/\mu)$. By Chernoff bounds, in a phase of $\kappa\mu n$ steps, the best distance is decreased at least by a factor $(1 - \Omega(1/n))^{\Omega(\kappa n)}$ w. o. p. This factor is at most $1/2$ if the constant κ is chosen appropriately large. This proves the theorem's fourth claim. $\qquad\square$

6 Discussion

Theorem 2 and Theorem 3 together mean that $\mu = 1$ is (in an asymptotic sense) an optimal choice for the $(\mu{+}1)\,\mathrm{ES}$ on the Sphere function. It seems, however, that the utility of its population should become apparent on specific multimodal fitness functions. With respect to the $(\mu{+}1)\,\mathrm{EA}$ in discrete search spaces, Witt [15] presented a well-chosen example where an increase of μ by a small, polynomial factor decreased the expected runtime drastically, namely by an exponential factor. In order to show a corresponding result for the $(\mu{+}1)\,\mathrm{ES}$, we should also consider a well-chosen fitness function. Jansen and Wegener [10] have shown that a population can help to cross a small valley of bad fitness. We expect a similar benefit for the following function. For $\boldsymbol{x} \in \mathbb{R}^n$ and $0 \leq s < 1$, let

$$f_s(\boldsymbol{x}) := \begin{cases} |\boldsymbol{x}| + s & \text{if } |\boldsymbol{x}| < 1, \\ |\boldsymbol{x}| & \text{otherwise.} \end{cases}$$

If $s > 0$, $f_s(\boldsymbol{x})$ is a multimodal function. Search points $\boldsymbol{x}$ such that $|\boldsymbol{x}| = 1$ are locally optimal and the set of search points $\{\boldsymbol{x} \mid 1 - s < |\boldsymbol{x}| < 1\}$ forms "the gap" where the fitness is worse.

Let the $(1{+}1)\,\mathrm{ES}$ start with some search point whose distance from the optimum is at least $1 + s$. If s is not too small, the $(1{+}1)\,\mathrm{ES}$ is likely to create some search point whose distance from the optimum is in the interval $[1, 1 + s]$. In this situation, any better search point is at distance at least s from the current one. The probability of a mutation overcoming this distance can be arbitrarily small, depending on s. Moreover, in the local optimum, the adaptation of the $(1{+}1)\,\mathrm{ES}$ would lead to smaller and smaller steps lengths, making an escape even more unlikely (finally resulting in premature convergence).

Suppose the $(\mu{+}1)\,\mathrm{ES}$ with large μ is in the situation of its closest-to-the-optimum individual being at some distance $d_\mathrm{c} \in [1, 1+s]$. Moreover, suppose that the farthest one is at distance $d_\mathrm{f} > 1 + s$. If the closest individual is chosen and its mutant is at distance less than 1, the mutant will be inserted into the next population. Hence, there is a chance of accepting individuals from the gap, which might lead to descendants at distance less than $1 - s$, i.e., outside the gap.

A formal proof of the utility of the $(\mu{+}1)\,\mathrm{ES}$ in this setting, however, would require several arguments. Here, one would have to show that the population maintains enough diversity for the above-mentioned step to happen. Moreover, it seems that s has to be chosen carefully since large gaps could only be overcome if the diversity of the population was very high. Before starting a theoretical analysis, careful experiments could give hints on suitable choices for s. Such experiments and, possibly, a formal proof are subjects for further research.

Conclusion

In this paper, we have presented a first rigorous runtime analysis of a population-based EA, namely the $(\mu{+}1)\,\mathrm{ES}$ with 1/5-rule, in continuous search spaces. With respect to the well-known SPHERE function, we have obtained asymptotically tight upper and lower bounds on the time for the $(\mu{+}1)\,\mathrm{ES}$ to halve the distance to the optimum. We have proven that the choice $\mu = 1$ leads to the (in the asymptotic framework) lowest runtime.

For the proof of the lower bound, we have generalized the technique of randomized family trees, developed before only w.r.t. to discrete search spaces. Interestingly, some of the related arguments on family trees could even be reused in the proof of the upper bound on the runtime. By this upper bound, we have shown that the 1/5-rule makes sense in our $(\mu{+}1)\,\mathrm{ES}$. In the proof that the 1/5-rule works, we have exploited that the $(\mu{+}1)\,\mathrm{ES}$ on SPHERE quickly loses diversity, i.e., that all individuals are at almost the same distance from the optimum after a small number of steps.

Due to the simplicity of the SPHERE function, we could not prove any advantage of the population of the $(\mu{+}1)\,\mathrm{ES}$. However, the SPHERE function serves as a starting point for the runtime analysis of general $(\mu\overset{+}{,}\lambda)\,\mathrm{ES}$ in continuous search spaces and for developing appropriate proof techniques. We have discussed multimodal example functions where the search might benefit from the population and regard rigorous runtime analyses of such functions and more general ES as subjects for further research.

References

[1] H.-G. Beyer. *The Theory of Evolution Strategies.* Springer, 2001.

[2] S. Droste, T. Jansen, and I. Wegener. On the analysis of the (1+1) evolutionary algorithm. *Theoretical Computer Science*, 276:51–81, 2002.

[3] T. Ericson and V. Zinoviev. *Codes on Euclidean spheres.* Elsevier, 2001.

[4] O. Giel and I. Wegener. Evolutionary algorithms and the maximum matching problem. In *Proceedings of the 20th International Symposium on Theoretical Computer Science (STACS '03)*, volume 2607 of *Lecture Notes in Computer Science*, pages 415–426. Springer, 2003.

[5] J. H. Holland. *Adaptation in Natural and Artificial Systems.* The University of Michigan Press, Ann Arbor, MI, USA, 1975.

[6] J. Jägersküpper. Analysis of a simple evolutionary algorithm for the minimization in Euclidian spaces. Technical Report CI 140/02, SFB 531, Universität Dortmund, 2002. `http://sfbci.uni-dortmund.de`.

[7] J. Jägersküpper. Analysis of a simple evolutionary algorithm for minimization in Euclidean spaces. In *Proceedings of the 30th International Colloquium on Automata, Languages, and Programming (ICALP '03)*, volume 2719 of *Lecture Notes in Computer Science*, pages 1068–1079. Springer, 2003.

[8] J. Jägersküpper. Rigorous runtime analysis of the (1+1) ES: 1/5-rule and ellipsoidal fitness landscapes. In *Proceedings of Foundations of Genetic Algorithms 8 (FOGA '05)*, volume 3469 of *Lecture Notes in Computer Science*. Springer, 2005. To appear.

[9] T. Jansen and K. De Jong. An analysis of the role of offspring population size in EAs. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '02)*, pages 238–246. Morgan Kaufmann, 2002.

[10] T. Jansen and I. Wegener. On the utility of populations. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '01)*, pages 1034–1041. Morgan Kaufmann, 2001.

[11] R. Motwani and P. Raghavan. *Randomized Algorithms.* Cambridge University Press, 1995.

[12] F. Neumann and I. Wegener. Randomized local search, evolutionary algorithms, and the minimum spanning tree problem. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '04)*, volume 3102 of *Lecture Notes in Computer Science*, pages 713–724. Springer, 2004.

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

[14] T. Storch and I. Wegener. Real royal road functions for constant population size. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '03)*, volume 2724 of *Lecture Notes in Computer Science*, pages 1406–1417. Springer, 2003.

[15] C. Witt. An analysis of the (μ+1) EA on simple pseudo-boolean functions. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO '04)*, volume 3102 of *Lecture Notes in Computer Science*, pages 761–773. Springer, 2004.

[16] C. Witt. An analysis of the (μ+1) EA on simple pseudo-boolean functions. Technical Report CI 170/04, SFB 531, Universität Dortmund, 2004. `http://sfbci.uni-dortmund.de`.

[17] C. Witt. Worst-case and average-case approximations by simple randomized search heuristics. In *Proceedings of the 22nd Annual Symposium on Theoretical Aspects of Computer Science (STACS '05)*, volume 3404 of *Lecture Notes in Computer Science*, pages 44–56. Springer, 2005.