

Analysis of a Simple Evolutionary Algorithm for Minimization in Euclidean Spaces

Jens Jägersküpfer*

FB Informatik, LS 2, Univ. Dortmund, 44221 Dortmund, Germany
jj@Ls2.cs.uni-dortmund.de

Abstract. Although evolutionary algorithms (EAs) are widely used in practical optimization, their theoretical analysis is still in its infancy. Up to now results on the (expected) runtime are limited to discrete search spaces, yet EAs are mostly applied to continuous optimization problems. So far results on the runtime of EAs for continuous search spaces rely on validation by experiments/simulations since merely a simplifying model of the respective stochastic process is investigated.

Here a first algorithmic analysis of the expected runtime of a simple, but fundamental EA for the search space \mathbb{R}^n is presented. Namely, the so-called (1+1) Evolution Strategy ((1+1) ES) is investigated on unimodal functions that are monotone with respect to the distance between search point and optimum. A lower bound on the expected runtime is proven under the only assumption that isotropic distributions are used to generate the random mutation vectors. Consequently, this bound holds for any mutation adaptation mechanism. Finally, we prove that the commonly used “Gauss mutations” in combination with the so-called 1/5-rule for the mutation adaptation do achieve asymptotically optimal expected runtime.

Keywords: Evolutionary Algorithms, Black-Box Optimization, Continuous Search Space, Expected Runtime, Mutation Adaptation

1 Introduction

The optimization, here the minimization, of functions $f: S \rightarrow \mathbb{R}$ for some given search space S is one of the fundamental algorithmic problems. Discrete search spaces, e. g. $\{0, 1\}^n$, lead to combinatorial optimization problems like TSP, knapsack, or maximum matching. Mathematical optimization deals with continuous search spaces, usually \mathbb{R}^n . Here, problems are commonly defined by classes of functions, like polynomials of degree d , k -times differentiable functions, etc. Many problem-specific algorithms have been designed for each of these two scenarios. Since such algorithms are analyzed (in general), they can be compared and there is a theory on algorithms.

* supported by the Deutsche Forschungsgemeinschaft (DFG) as part of the collaborative research center “Computational Intelligence” (SFB 531)

If not enough resources are on hand to design a problem-specific algorithm, however, robust algorithms like randomized search heuristics are often a good alternative. Especially, if the knowledge about the function f to be optimized is not sufficient, classical mathematical optimization algorithms like the steepest descent method or the conjugate gradient method cannot be applied. In the extreme, for instance if f is only given implicitly, knowledge about f can solely be gathered by consecutively evaluating f at selected points. This situation is commonly named “black-box optimization.” In this scenario, runtime is measured by the number of f -evaluations. Obviously, if we know nothing about f , a (reasonable) theoretical analysis of the runtime of some search heuristic like an evolutionary algorithm is impossible. Thus, to get insight into why such algorithms do often work quite well in practice, assumptions about the properties of f must be made, with respect to which the analysis is carried out.

This approach has been taken since the early 1990s for the discrete search space $\{0, 1\}^n$. Probably the first function that was analyzed is $\text{ONEMAX}(\mathbf{b}) := b_1 + \dots + b_n$, $\mathbf{b} = (b_1, \dots, b_n) \in \{0, 1\}^n$ (the name reflects that maximization was considered rather than minimization). The algorithm investigated was the so-called (1+1) Evolutionary Algorithm ((1+1) EA), which is in fact the discrete counterpart of the (1+1) ES investigated here. Both algorithms use a population consisting of only one search point, called an individual in the field of Evolutionary Computation. Thus, recombination is precluded, and mutation is the only “evolutionary force.” Within each beat of the evolution loop, the mutation of the current individual temporarily generates a second individual, and selection determines which one of both founds the next generation. An $O(n \log n)$ bound on the expected runtime of the (1+1) EA for ONEMAX (if mutation consists in flipping each bit of the individual independently with probability $1/n$) is proved in [9]. Retrospectively, this bound is easy to obtain; yet more sophisticated papers on the (1+1) EA have been published: In [2] linear functions are analyzed, in [14] quadratic polynomials, and in [13] monotone polynomials. Furthermore, [4] investigates the (1+1) EA for the maximum matching problem. Even the effect of recombination has been analyzed for the search space $\{0, 1\}^n$ [7, 8], and the number of papers on algorithmic analyses is increasing.

The situation for continuous search spaces is different: The vast majority of results on EAs are empirical, i. e., based on experiments and simulations. In the few papers that focus on theoretical analyses, however, either (global) convergence is investigated or local changes from one generation to the next. In the former case, one must recall that EAs for continuous search spaces merely approximate an optimum rather than optimize the respective function. Convergence deals with the question of whether the algorithm reaches the ε -neighborhood of some (global) optimum in a finite number of steps or not (e. g. [11]). However, the order of the number of steps necessary remains open—in particular with respect to the dimension of the search space. On the other hand, results dealing with local changes in one step, for instance convergence rates, (generally) do not enable statements on the long-time behavior of EAs. Normally, the effect of mutation/recombination depends on the location of the respective individual(s) in the search space. Consequently, the changes from one generation to

the next one generally do not resemble the changes from the next generation to the second next. This is the reason why EAs for continuous search spaces apply so-called adaptation mechanisms, particularly mutation adaptation. The idea behind such adaptation mechanisms is to enable EAs to optimize as many types of functions as possible. Another idea behind mutation adaptation is that the mutative changes must in some way scale with the approximation quality. The rule of thumb reads: the closer the search approaches an optimum, the smaller the mutative changes. Unfortunately, (mutation) adaptation complicates the stochastic process an EA induces—and the analysis of the expected runtime.

The Scenario

As mentioned above, we will concentrate on the (1+1)ES which uses solely mutation because of a single-individual population. Let $\mathbf{c} \in \mathbb{R}^n$ denote this current individual. For a given initialization of \mathbf{c} , i. e., for a given starting point, the rough structure of the (1+1)ES is given by the following evolution loop:

1. Randomly choose the mutation vector $\mathbf{m} \in \mathbb{R}^n$.
2. Generate the mutant $\mathbf{x} \in \mathbb{R}^n$ by $\mathbf{x} := \mathbf{c} + \mathbf{m}$.
3. Using $f(\mathbf{c})$ and $f(\mathbf{x})$, the selection rule determines whether this mutant becomes the current individual ($\mathbf{c} := \mathbf{x}$) or is discarded (\mathbf{c} unchanged).
4. If the stopping criterion is met then output \mathbf{c} else goto 1.

A single execution of the loop is called a *step* of the (1+1)ES, and if “ $\mathbf{c} := \mathbf{x}$ ” is executed in a step, the mutation/mutant is said to be *accepted*, otherwise *rejected*. For a concrete instantiation of the (1+1)ES, the distribution of \mathbf{m} , the selection rule, and the stopping criterion must be specified. Although the stopping criterion is important in practice, we investigate the (1+1)ES as an infinite process. Let $T_f \in \mathbb{N}$ denote the number of steps the (1+1)ES needs to reach some fixed approximation quality when optimizing f . Then we are interested in $E[T_f]$ and in $P\{T_f \leq \tau\}$ for a given number of steps τ . By defining an appropriate randomized selection rule, simulated annealing can be realized for instance. However, we will investigate the commonly and originally used elitist selection where the mutant \mathbf{x} becomes/replaces the current individual \mathbf{c} if and only if $f(\mathbf{x}) \leq f(\mathbf{c})$. As this selection rule precludes worsenings, the (1+1)ES becomes a randomized hill-climber. If mutation adaptation is applied, obviously, the distribution of the mutation vector \mathbf{m} is not fixed, but varies during the optimization process. Here we concentrate on mutation vectors that are isotropically distributed.

Definition 1. For $\mathbf{m} \in \mathbb{R}^n$, let $|\mathbf{m}|$ denote its length, i. e., its L_2 -norm, and $\widehat{\mathbf{m}} := \mathbf{m}/|\mathbf{m}|$ the normalized vector. The distribution of the random vector \mathbf{m} is isotropic if $|\mathbf{m}|$ is independent of $\widehat{\mathbf{m}}$ and $\widehat{\mathbf{m}}$ is uniformly distributed upon the unit hyper-sphere $\{\mathbf{u} \in \mathbb{R}^n \mid |\mathbf{u}| = 1\}$.

Under these two assumptions (elitist selection and isotropically distributed mutation vectors) the lower bound on the runtime will be proved. That is, in each

step the mutation adaptation is free to choose an arbitrary isotropic distribution for \mathbf{m} . Consequently, the lower bound particularly holds for so-called “Gauss mutations” which are very common in practice (cf. Lemma 5 for the isotropy).

Definition 2. Let $\widetilde{\mathbf{m}} \in \mathbb{R}^n$ be $(N_1(0, 1), \dots, N_n(0, 1))$ -distributed (each component is independently standard normal distributed). A mutation is called Gauss mutation if the mutation vector’s distribution equals the one of $s \cdot \widetilde{\mathbf{m}}$, $0 < s \in \mathbb{R}$.

In particular, the upper bound on the runtime of the (1+1) ES will be proved with respect to Gauss mutations.

This scenario, (1+1) ES using elitist selection and Gauss mutations, has been introduced by Rechenberg, whose 1973 book *Evolutionstrategie* [10] is one starting point of evolutionary optimization. Rechenberg applied the (1+1) ES to optimize the shape of some workpiece. Furthermore, he presents some rough calculations on what length of the mutation vector maximizes the expected spacial gain in one step. These calculations are carried out with respect to two different kinds of functions. On the one hand, the so-called corridor model is considered, and on the other hand, the SPHERE function, where $\text{SPHERE}(\mathbf{x}) := x_1^2 + \dots + x_n^2$ for $\mathbf{x} = (x_1, \dots, x_n) \in \mathbb{R}^n$. The calculations for the one-step behavior of the (1+1) ES on SPHERE have been improved by Beyer and can be found in his 2001 book *The Theory of Evolution Strategies* [1]. As a conclusion, Rechenberg states that the length of a Gauss mutation vector should be adapted such that the *success probability of a step*, the probability that the mutation in this step is accepted, is about $1/5$. This led to the notion of the *1/5-rule* for mutation adaptation: The (expected) length of the Gauss mutation vectors are scaled by adapting the factor s in Definition 2 as follows. For a certain number of steps (originally $\Theta(n)$ many), the relative frequency of successful steps is observed without changing s . Subsequent to each observation phase, the relative share of successful steps in the respective phase is evaluated; if it is smaller than $1/5$, s is divided by some fixed constant greater than 1, and otherwise, s is multiplied by some fixed (possibly different) constant greater than 1. The upper bound on the runtime will be proved with respect to this $1/5$ -rule.

Finally, the class of functions we consider contains all unimodal $f: \mathbb{R}^n \rightarrow \mathbb{R}$, $n \in \mathbb{N}$, such that for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and the respective optimum/minimum $\mathbf{o}_f \in \mathbb{R}^n$: $|\mathbf{x} - \mathbf{o}_f| < |\mathbf{y} - \mathbf{o}_f| \Rightarrow f(\mathbf{x}) < f(\mathbf{y})$. In other words, if an individual is closer to the optimum than some other, also its function value is better/smaller. We assume w. l. o. g. that the optimum \mathbf{o}_f coincides with the origin, and thus, w. l. o. g. $|\mathbf{x}| < |\mathbf{y}| \Rightarrow f(\mathbf{x}) < f(\mathbf{y})$ for $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$. Obviously, the L_2 -norm itself and for instance SPHERE (as well as all their translations) bear this property.

Results

As mentioned above, the one-step behavior of (1+1) ES on SPHERE has been investigated by Rechenberg and in great detail by Beyer. Unfortunately, at certain points within these calculations the limit $n \rightarrow \infty$ is taken without controlling the error terms; this is problematic in an algorithmic analysis, which exactly focuses on how the runtime depends on n . Thus, in Section 2 the n -dependence of the

one-step behavior of the (1+1) ES is investigated. The impact of the 1/5-rule on the convergence of the (1+1) ES is investigated in [12] and [5] for instance; yet the order of the number of steps is not tackled. Applying methods and concepts known from the field of randomized algorithms, the main results mentioned in the abstract are shown in Section 3. Finally, we close with some concluding remarks. Note that more detailed proofs can be found in [6].

Notions and Notations

As mentioned in Definition 1, $|\mathbf{x}|$ denotes the L_2 -norm of the vector $\mathbf{x} \in \mathbb{R}^n$, i. e., its length in Euclidean space, and $x_i \in \mathbb{R}$ its i th component. Furthermore, for instance, “ n -sphere” abbreviates “ n -dimensional sphere”

Definition 3. A probability $p(n)$ is exponentially small in n if for a positive constant ε , $p(n) = \exp(-\Omega(n^\varepsilon))$. An event $A(n)$ happens with overwhelming probability (w. o. p.) with respect to n if $\mathbb{P}\{\neg A(n)\}$ is exponentially small in n .

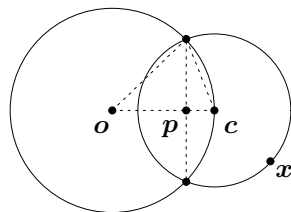
2 One-Step Behavior

As we are interested in how fast the “evolving” individual of the (1+1) ES approaches the optimum in the search space, the spatial gain towards the optimum in one step is the intermediate objective. Since the 1/5-rule for mutation adaptation is investigated, it is particularly interesting what length of the mutation vector results in the mutant being accepted with probability 1/5. Due to the independence of the random length of an isotropic mutation vector and its random direction (cf. Definition 1), we may assume that the length $\ell > 0$ of the mutation vector \mathbf{m} is chosen according to $|\mathbf{m}|$ ’s distribution first; then the mutant is uniformly distributed upon the n -sphere with radius ℓ centered at the current search point \mathbf{c} .

The situation is depicted by the figure on the right. The left sphere $F := \{\mathbf{c}' \in \mathbb{R}^n \mid |\mathbf{c}'| = |\mathbf{c}|\}$ will be called the *fitness sphere* since the properties of f imply that all points inside (resp. outside) the fitness sphere are better (resp. worse) than the current search point \mathbf{c} . The potential mutants define the *mutation sphere* $M := \{\mathbf{x} \in \mathbb{R}^n \mid |\mathbf{x} - \mathbf{c}| = \ell\}$.

Let $I := F \cap M \subset \mathbb{R}^n$ denote the intersection of the two spheres. Obviously, if $\ell > 2|\mathbf{c}|$, I is empty, and if $\ell = 2|\mathbf{c}|$, I is a singleton, such that we concentrate on $\ell < 2|\mathbf{c}|$. It is easy to see that I forms an $(n-1)$ -sphere, and that the hyperplane $P \supset I$ is orthogonal to the line passing through \mathbf{c} and \mathbf{o} . (Let $\mathbf{p} \in P$ denote the point where this line passes through P .) Hence, the mutation sphere’s part lying inside the fitness sphere forms a hyper-spherical cap $C \subset M - I$, the missing boundary of which is I . Basic geometry shows that the distance between \mathbf{c} and P equals $g := |\mathbf{c}| - |\mathbf{p}| = \ell^2 / (2|\mathbf{c}|)$ if $\ell \leq \sqrt{2}|\mathbf{c}|$.

Since the mutant \mathbf{x} is uniformly distributed upon the mutation sphere M , for any (Lebesgue measurable) $S \subseteq M$, $\mathbb{P}\{\mathbf{x} \in S \mid |\mathbf{m}| = \ell\}$ equals the ratio



of the $(n-1)$ -volume of S to the one of M , inducing a probability measure. Consequently, I is of zero measure, and since \mathbf{x} is better than \mathbf{c} if $\mathbf{x} \in C$, and worse if $\mathbf{x} \in M - (C \cup I)$, the probability that the mutant is accepted equals the ratio of the hypersurface area of C to the one of M . Now, the interesting question is how this ratio depends on $|\mathbf{c}|$, ℓ , and, of course, n , the number of dimensions.

As the height of the mutation sphere's cap that is cut off by the fitness sphere equals $h := \ell - g = \ell - \ell^2/(2|\mathbf{c}|)$, the *relative height* of C , the ratio h/ℓ , equals $1 - \ell/(2|\mathbf{c}|)$. It can be shown (cf. [3, Appendix B] for instance) that

$$\frac{\text{hypersurface area of } C}{\text{hypersurface area of } M} = \frac{\Psi_{n-2}(\arccos(1 - h/\ell))}{\Psi_{n-2}(\pi)}$$

in n -space, $n \geq 3$, where $\Psi_k(\gamma) := \int_0^\gamma (\sin \beta)^k d\beta$. Note that $1 - h/\ell = (\ell - h)/\ell = g/\ell$. This formula may be directly used to estimate a step's success probability, yet it can also be utilized more generally: The ratio $\Psi_{n-2}(\arccos(g/\ell))/\Psi_{n-2}(\pi)$ not only equals the probability that the mutation hits C , but also the one of "the spatial gain of an isotropic mutation vector \mathbf{m} parallel to some fixed direction (for instance $\overline{\mathbf{c}\mathbf{o}}$) is greater than g ," under the condition $|\mathbf{m}| = \ell$. Therefore, let G denote the random variable given by the spatial gain of an isotropic mutation \mathbf{m} parallel to a fixed direction under the condition $|\mathbf{m}| = \ell$. Then

$$\mathbb{P}\{G \leq g\} = 1 - \mathbb{P}\{G > g\} = 1 - \frac{\Psi_{n-2}(\arccos(g/\ell))}{\Psi_{n-2}(\pi)},$$

and hence, $F_n(x) := 1 - \Psi_{n-2}(\arccos(x/\ell))/\Psi_{n-2}(\pi)$ for $x \in [-\ell, \ell]$ is G 's probability distribution over $[-\ell, \ell]$ in n -space. Since Ψ_k is continuous, the probability density of G at $g \in [-\ell, \ell]$ equals $\frac{dF_n(x)}{dx}(g) = F'_n(g)$,

$$\begin{aligned} F'_n(x) &= \Psi_{n-2}(\pi)^{-1} \cdot (-1) \cdot \frac{d}{dx} \Psi_{n-2}(\arccos(x/\ell)) \\ &= \Psi_{n-2}(\pi)^{-1} \cdot (-1) \cdot \frac{d}{dx} \int_0^{\arccos(x/\ell)} (\sin \beta)^{n-2} d\beta \\ &= \Psi_{n-2}(\pi)^{-1} \cdot \ell^{-1} \cdot (1 - (g/\ell)^2)^{(n-3)/2} \end{aligned}$$

for $n \geq 4$. To make things clear, this is the density of the spatial gain of an isotropically distributed mutation vector \mathbf{m} parallel to an arbitrarily fixed direction—independently of the function optimized—if $|\mathbf{m}|$ takes the value ℓ , not the one towards the optimum after selection.

With the help of this density function, we obtain an alternative formula for the success probability of a step, in which \mathbf{c} is mutated using an isotropically distributed mutation vector \mathbf{m} with $|\mathbf{m}| = \ell$ (y substitutes g/ℓ):

$$\begin{aligned} \mathbb{P}\{\mathbf{x} \text{ is accepted} \mid |\mathbf{m}| = \ell\} &= \mathbb{P}\{\mathbf{x} \in C \mid |\mathbf{m}| = \ell\} = \mathbb{P}\left\{G \geq \frac{\ell^2}{2|\mathbf{c}|}\right\} \\ &= \int_{\ell^2/(2|\mathbf{c}|)}^{\ell} F'_n(g) dg = \frac{\ell}{\Psi_{n-2}(\pi) \cdot \ell} \int_{\ell/(2|\mathbf{c}|)}^1 (1 - y^2)^{(n-3)/2} dy \end{aligned}$$

With respect to the 1/5-rule, which will be investigated for the upper bound on the expected runtime, we can now answer what length of the mutation vector results in a step of the (1+1)ES having success probability 1/5. Note that, obviously, this probability approaches 1/2 as $\ell/|\mathbf{c}| \rightarrow 0$.

Lemma 1. *In the scenario considered, the mutant $\mathbf{c} + \mathbf{m} \in \mathbb{R}^n$ is accepted with a constant probability greater than 0 and smaller than 1/2 if and only if $|\mathbf{m}|$ takes a value $\ell = \Theta(|\mathbf{c}|/\sqrt{n})$ in the respective step.*

Proof. The distance between \mathbf{c} and P , the hyperplane containing the intersection of mutation sphere and fitness sphere, equals $\ell^2/(2|\mathbf{c}|) = \ell \cdot (\lambda/\sqrt{n})$ with $\lambda = \Theta(1)$, i. e., the relative height of the cap C equals $1 - \lambda/\sqrt{n}$. Using the formula derived above, we must show that $\int_{\lambda/\sqrt{n}}^1 (1 - y^2)^{(n-3)/2} dy$ as well as $\int_0^{\lambda/\sqrt{n}} (1 - y^2)^{(n-3)/2} dy$ are in $\Omega(\Psi_{n-2}(\pi))$, respectively. See [6]. \square

In other words, if the 1/5-rule was able to ensure a success probability of exactly 1/5 in a step, the length of the mutation vector would be $\Theta(|\mathbf{c}|/\sqrt{n})$ in this step. Thus, the expected spatial gain towards the optimum in this situation is of particular interest and is estimated in the following.

Lemma 2. *If (in the scenario considered) $|\mathbf{m}| = \Theta(|\mathbf{c}|/\sqrt{n})$ in a step then the spatial gain towards the optimum is $\Omega(|\mathbf{m}|/\sqrt{n}) = \Omega(|\mathbf{c}|/n)$ with probability $\Omega(1)$ in this step, and thus, also the expected decrease in distance to the optimum in this step is $\Omega(|\mathbf{c}|/n)$.*

Proof. As in Lemma 1, the assumptions imply that C has height $\ell \cdot (1 - \lambda/\sqrt{n})$ for $\lambda = \Theta(1)$. One result of that Lemma is that the mutation hits C with probability $\Omega(1)$. Let $A \subset C$ denote the cap with height $\ell \cdot (1 - 2\lambda/\sqrt{n})$ such that its pole coincides with the one of C . Then each point in A is at least $\ell \cdot (1 - \lambda/\sqrt{n}) - \ell \cdot (1 - 2\lambda/\sqrt{n}) = \ell \cdot \lambda/\sqrt{n}$ distance units closer to the optimum than a point belonging to the boundary of C . Since the boundary of C equals the intersection of mutation sphere and fitness sphere, the distance to the optimum is decreased by at least $\ell \cdot \lambda/\sqrt{n} = \Theta(|\mathbf{c}|/n)$ distance units if the mutation hits A . This still happens with probability $\Omega(1)$ because the relative height of A equals $1 - \Theta(1/\sqrt{n})$ like the one of C . Since the properties of f in combination with the selection rule preclude a negative spatial gain, the expected decrease in distance to the optimum is $\Omega(|\mathbf{c}|/n)$. \square

Consequently, if the 1/5-rule is capable of adjusting the mutation vector's length such that the success probability is close to 1/5, the distance to the optimum is expected to decrease by an $\Omega(1/n)$ -fraction. Note that, e. g., an 1/8-rule or an 1/3-rule would lead to the same asymptotic expected gain. Naturally, one might ask if an expected spatial gain $\omega(|\mathbf{c}|/n)$ is possible. We prove that in our scenario the expected spatial gain towards the optimum is $O(|\mathbf{c}|/n)$ for any adaptation of the length of an isotropic mutation vector. Hence, the 1/5-rule indeed tries to adjust the mutation vector's length to have optimal order $\Theta(|\mathbf{c}|/\sqrt{n})$ such that the expected spatial gain towards the optimum has maximum order $\Theta(|\mathbf{c}|/n)$.

Obviously, the spatial gain of a step equals 0 if the mutation is rejected, and is upper bounded by the mutation's spatial gain parallel to $\overline{\mathbf{c}\sigma}$, otherwise. A mutation is accepted (resp. rejected) if the spatial gain parallel to $\overline{\mathbf{c}\sigma}$ is greater (resp. smaller) than $\ell^2/(2|\mathbf{c}|)$. Using the probability density function obtained above, the expected spatial gain of a step, call it $\mathbb{E}[\text{gain}]$, is bounded above by

$$\begin{aligned} \int_{\ell^2/(2|\mathbf{c}|)}^{\ell} gF'_{n-2}(g) dg &= \frac{\ell}{\Psi_{n-2}(\pi)} \int_{\ell/(2|\mathbf{c}|)}^1 y \cdot (1-y^2)^{(n-3)/2} dy \\ &= \frac{\ell}{\Psi_{n-2}(\pi) \cdot (n-1)} \cdot \left(1 - \left(\frac{\ell}{2|\mathbf{c}|}\right)^2\right)^{(n-1)/2} \\ &< \frac{\ell}{\sqrt{2\pi}\sqrt{n-1}} \cdot \left(1 - \left(\frac{\ell}{2|\mathbf{c}|}\right)^2\right)^{(n-1)/2} \end{aligned}$$

because for $n \geq 4$, $\int y \cdot (1-y^2)^{(n-3)/2} dy = (1-y^2)^{(n-1)/2}/(-(n-1))$ and $\Psi_{n-2}(\pi) > \sqrt{2\pi}/\sqrt{n-1}$ (cf. [6]). Consequently, $\mathbb{E}[\text{gain}] = O(|\mathbf{m}|/\sqrt{n})$ independently of the *scaled distance from the optimum* $|\mathbf{c}|/|\mathbf{m}|$ (remember that $|\mathbf{m}| > 2|\mathbf{c}|$ results in the mutant being rejected since it lies outside the fitness sphere). Furthermore, the inequality enables the proof that $\mathbb{E}[\text{gain}] = O(|\mathbf{c}|/n)$ for any adaptation of the mutation vector's length.

Lemma 3. *In the scenario considered, the expected spatial gain towards the optimum in a step is $O(|\mathbf{c}|/n)$ —for any isotropic mutation.*

Proof. To prove this claim, we must show that $\mathbb{E}[\text{gain}]/|\mathbf{c}| = O(1/n)$ even if the mutation vector's length ℓ is chosen such that the expected spatial gain is maximized. Let $d := |\mathbf{c}|/\ell$ denote the scaled distance from the optimum. Applying the upper bound on the expected spatial gain from above yields

$$\mathbb{E}[\text{gain}]/|\mathbf{c}| < (2\pi(n-1))^{-1/2} \cdot \underbrace{(1/d) \cdot (1-(2d)^{-2})^{(n-1)/2}}_{=: w_n(d)}.$$

Hence, an upper bound on $\mathbb{E}[\text{gain}]/|\mathbf{c}|$ can be derived by maximizing the function w_n . In fact, $w_n(d) = O(1/\sqrt{n})$ for $d > 0$ (cf. [6]), and thus,

$$\mathbb{E}[\text{gain}]/|\mathbf{c}| < w_n(d)/\sqrt{2\pi(n-1)} = O(1/\sqrt{n})/\sqrt{2\pi(n-1)} = O(1/n) \quad \square$$

3 Multi-Step Behavior and Expected Runtime

Obviously, the multi-step behavior of the (1+1)ES crucially depends on the mutation adaptation used. For a lower bound on the expected runtime, however, optimal mutation adaptation may be assumed. Surprisingly, we need not prove explicitly what mutation adaptation is optimal. Furthermore, it is not evident what “runtime” means since f is merely approximated rather than optimized.

Due to the symmetry and scalability properties of f , linearity of expectation enables further statements if one knows (for an arbitrary starting point) the

expected number of steps to halve the distance from the optimum using optimal mutation adaptation. Namely, the expected runtime to reduce the distance from the optimum to a $1/k$ -fraction is lower bounded by $\lceil \log_2 k \rceil$ times the lower bound on the expected runtime to halve it. We apply the following modification of Wald's equation to prove the lower bound on the expected number of steps the (1+1)ES needs to halve the distance from the optimum (cf. [6] for the proof of this lemma).

Lemma 4. *Let X_1, X_2, \dots denote random variables with bounded range and T the random variable defined by $T = \min\{t \mid X_1 + \dots + X_t \geq g\}$ for a given $g > 0$. If $\mathbb{E}[T]$ exists and $\mathbb{E}[X_i \mid T \geq i] \leq u$ for $i \in \mathbb{N}$ then $\mathbb{E}[T] \geq g/u$.*

Theorem 1. *In the scenario considered, for any adaptation of isotropic mutations the expected number of steps to halve the distance to the optimum is $\Omega(n)$.*

Proof. For $i \geq 1$, let X_i denote the random variable that corresponds to the spatial gain towards the optimum in the i th step. Furthermore, let $\mathbf{a} \in \mathbb{R}^n - \{\mathbf{o}\}$ denote the starting point and T the (random) number of steps until $|\mathbf{c}| \leq |\mathbf{a}|/2$ for the first time. As mentioned previously, worsenings are precluded such that $X_i \geq 0$ and in particular $|\mathbf{c}| \leq |\mathbf{a}|$ in each step. Consequently, $X_i \leq |\mathbf{c}| \leq |\mathbf{a}|$, and according to Lemma 3, $\mathbb{E}[X_i \mid T \geq i] = O(|\mathbf{c}|/n) = O(|\mathbf{a}|/n)$. Choosing $g := |\mathbf{a}|/2$ in Lemma 4, $\mathbb{E}[T] \geq (|\mathbf{a}|/2)/O(|\mathbf{a}|/n) = \Omega(n)$ if $\mathbb{E}[T]$ exists. If $\mathbb{E}[T]$ is not defined (due to improper adaptation), one may informally argue that “ $\mathbb{E}[T] = \infty = \Omega(n)$ ” since T is positive. \square

This lower bound on the expected runtime holds independently of the mutation adaptation applied since theoretically optimal adaptation is (implicitly) assumed. For the upper bound, we concretize the lower-bound scenario by choosing Gauss mutations and the 1/5-rule for mutation adaptation. The following properties of Gauss-mutations are useful (and proved in [6]).

Lemma 5. *A Gauss-mutation $\mathbf{m} \in \mathbb{R}^n$ is isotropically distributed, and moreover, $\ell_{\mathbb{E}} := \mathbb{E}[|\mathbf{m}|]$ exists and $\mathbb{P}\{||\mathbf{m}| - \ell_{\mathbb{E}}| \geq \delta \cdot \ell_{\mathbb{E}}\} \leq \delta^{-2}/(2n - 1)$.*

Let $\mathbf{m}_1, \dots, \mathbf{m}_n$ denote independent copies of \mathbf{m} . For any constant $\lambda < 1$ two positive constants a_λ, b_λ exist such that $\#\{i \mid a_\lambda \ell_{\mathbb{E}} \leq |\mathbf{m}_i| \leq b_\lambda \ell_{\mathbb{E}}\} \geq \lambda n$ w. o. p.

Furthermore, we investigate this instantiation of the 1/5-rule: The scaling factor s (cf. Definition 2) is adapted after every n th step: if less than $n/5$ of the respective last n steps were successful, s is halved, otherwise doubled. The asymptotic calculations we present, however, are valid for any 1/5-rule keeping s unchanged for $\Theta(n)$ steps, respectively, and using any two constants, each greater than 1, for the scaling of s .

The run of the (1+1)ES is partitioned into phases each of which lasts n steps such that $\mathbb{E}[|\mathbf{m}|]$ is constant in each phase. Let s_i denote the scaling factor used throughout the i th phase and ℓ_i the corresponding $\mathbb{E}[|\mathbf{m}|]$. A phase after which s is doubled is symbolized by “ \times ”, and one after which s is halved by “ \div ”. Furthermore, let d_i denote the distance from the optimum at the beginning of the i th phase; hence, $d_i - d_{i+1}$ equals the spatial gain in/of the i th phase.

Lemma 6. *In the scenario considered for the 1/5-rule for Gauss mutations:*

1. if $\ell_i = \Theta(d_i/\sqrt{n})$ then $d_{i+1} = d_i - \Omega(d_i)$ w. o. p.,
2. if s is doubled after the i th phase then $\ell_i = O(d_i/\sqrt{n})$ w. o. p.,
3. if s is halved after the i th phase then $\ell_{i+1} = \Omega(d_{i+1}/\sqrt{n})$ w. o. p.

Proof. Assume that the total spatial gain of the i th phase is not $\Omega(d_i)$. Then the distance from the optimum is $\Theta(d_i)$ in each step of the phase (remember that the distance is non-increasing). Lemma 5 yields that w. o. p. $|\mathbf{m}| = \Theta(d_i/\sqrt{n})$ in $0.9n$ steps. According to Lemma 2, in each such step the spatial gain is $\Omega(d_i/n)$ with probability $\Omega(1)$. Hence, we expect $\Omega(n)$ steps each of which reduces the distance by $\Omega(d_i/n)$. By Chernoff bounds, the number of such steps is $\Omega(n)$ w. o. p. Consequently, our initial assumption contradictorily implies that the total spatial gain of the i th phase is $\Omega(d_i)$ w. o. p.

For the second claim, assume ℓ_i is not $O(d_i/\sqrt{n})$. Since the distance from the optimum is non-increasing, ℓ_i is not $O(|\mathbf{c}|/\sqrt{n})$ in each step of the i th phase. Lemma 5 yields that $|\mathbf{m}|$ is not $O(|\mathbf{c}|/\sqrt{n})$ in $0.9n$ steps w. o. p. According to Lemma 1, the success probability of each such step is $o(1)$. Hence, the expected number of unsuccessful steps is lower bounded by $0.9n - o(n)$. By Chernoff bounds, w. o. p. more than $0.8n$ steps are not successful. Thus, the assumption “ ℓ_i is not $O(d_i/\sqrt{n})$ ” contradictorily implies that s is halved w. o. p.

Assume ℓ_{i+1} is not $\Omega(d_{i+1}/\sqrt{n})$ for the third claim. Since $s_i = 2s_{i+1}$ also $\ell_i = 2\ell_{i+1}$. As the distance is non increasing, the assumption implies that ℓ_i is not $\Omega(|\mathbf{c}|/\sqrt{n})$ for each step of the \div -phase. Following the proof of the second claim with symmetric arguments, w. o. p. more than $0.8n$ steps are successful—contradictorily implying that the i th phase is a \times -phase w. o. p. \square

Now we can deal with sequences of phases in a run of the (1+1) ES.

Lemma 7. *If (in the scenario considered) the 1/5-rule for Gauss mutations causes a sequence $\div \times^k$ of phases, $k = \text{poly}(n)$, then w. o. p. the distance from the optimum is k times reduced by a constant fraction in the respective phases.*

Proof. Let the \div -phase be the i th one. By Lemma 6, $\ell_{i+1} = \Omega(d_{i+1}/\sqrt{n})$ w. o. p. Since the adaptation yields $\ell_{i+w} \geq \ell_{i+1}$, $1 \leq w \leq k$, and the distance is non-increasing, w. o. p. $\ell_{i+w} = \Omega(d_{i+w}/\sqrt{n})$ for $1 \leq w \leq k$. Lemma 6 also yields that w. o. p. $\ell_{i+w} = O(d_{i+w}/\sqrt{n})$ for $1 \leq w \leq k$. Consequently, w. o. p. $\ell_{i+w} = \Theta(d_{i+w}/\sqrt{n})$ for $1 \leq w \leq k$, and finally, again according to Lemma 6, in each of the k \times -phases the distance is reduced by a constant fraction w. o. p. \square

Lemma 8. *If (in the scenario considered) the 1/5-rule for Gauss mutations causes a sequence $\times \div^k$ of phases, $k = \text{poly}(n)$, then w. o. p. the distance from the optimum is k times reduced by a constant fraction in the respective phases.*

Proof. Let the \times -phase be the i th one. For $k = 1$, assume that the total spatial gain of the i th and the $(i+1)$ th phase is not $\Omega(d_i)$. According to Lemma 6, w. o. p. $\ell_i = O(d_i/\sqrt{n})$ and w. o. p. $\ell_{i+2} = \Omega(d_{i+2}/\sqrt{n})$. Hence, $\ell_i = \Theta(d_i/\sqrt{n})$ as well as $\ell_{i+1} = \Theta(d_{i+1}/\sqrt{n})$, and Lemma 6 contradictorily implies that in

each of the two phases the distance is reduced by a constant fraction w. o. p. Consequently, w. o. p. these two phases yield $d_{i+2} = d_i - \Omega(d_i)$.

For $k \geq 2$, the adaptation yields $s_{i+w} = s_i 2^{2-w} = 4 s_i / 2^w$ for $1 \leq w \leq k$, and according to Lemma 6, for $2 \leq w \leq k$ w. o. p. $\ell_{i+w} = \Omega(d_{i+w} / \sqrt{n})$. If $d_{i+w} \leq d_i / 2^w$ then by a simple accounting argument after the $(i+w)$ th phase $d_{i+w+1} \leq d_{i+w} \leq d_i / 2^w \leq d_i / \lambda^{w+1}$ for a constant $\lambda \geq \sqrt{2}$ and we are done. Thus, assume $d_{i+w} > d_i / 2^w$. As $\ell_{i+w} = 4 \ell_i / 2^w$, in this case “w. o. p. $\ell_i = O(d_i / \sqrt{n})$ ” implies that w. o. p. $\ell_{i+w} = O(d_{i+w} / \sqrt{n})$. Since also $\ell_{i+w} = \Omega(d_{i+w} / \sqrt{n})$, Lemma 6 yields that the $(i+w)$ th phase reduces the distance by a constant fraction w. o. p.

Altogether, the first two phases yield w. o. p. $d_{i+2} = d_i - \Omega(d_i)$, and for $2 \leq w \leq k$, either the distance from the optimum is reduced by a constant fraction in the $(i+w)$ th phase w. o. p., or after this phase $d_{i+w+1} \leq d_i / \lambda^{w+1}$ for a constant $\lambda \geq \sqrt{2}$ even if there was no spatial gain in the $(j+w)$ th phase. \square

Finally, the three preceding lemmas together with Theorem 1 yield the bound on the expected runtime, the expected number of steps the (1+1) ES needs for a predefined reduction of the distance from the optimum \mathbf{o} in the search space.

Theorem 2. *If (in the scenario considered) for the suboptimal initial search point $\mathbf{a} \in \mathbb{R}^n - \{\mathbf{o}\}$ and the initial scaling factor s_1 , $|\mathbf{a} - \mathbf{o}| / s_1 = \Theta(n)$ then the expected number of steps to obtain a search point \mathbf{c} such that $|\mathbf{c} - \mathbf{o}| \leq 2^{-t} |\mathbf{a} - \mathbf{o}|$ for $t \in \text{poly}(n)$ is $\Theta(t \cdot n)$.*

Proof. Assume w. l. o. g. that the optimum \mathbf{o} coincides with the origin. The lower bound $\Omega(t \cdot n)$ follows immediately from Theorem 1.

If the sequence of phases starts with $\times \div$ or with $\div \times$, the two preceding lemmas yield that the number phases until $\mathbb{E}[|\mathbf{c}|] < 2^{-(t+1)} |\mathbf{a}|$ is $O(t)$. If the sequence starts with \times^k or with \div^k for $k \geq 2$, we must show that in these phases the distance is w. o. p. reduced k times by a constant fraction. The assumptions on the starting values ensure that in the first phase $\ell_1 = \mathbb{E}[|\widehat{\mathbf{m}}|] \cdot s_1 = \Theta(\sqrt{n}) \cdot s_1 = \Theta(d_1 / \sqrt{n})$ (cf. Definition 2 for $\widehat{\mathbf{m}}$ and [6] for $\mathbb{E}[|\widehat{\mathbf{m}}|] = \Theta(\sqrt{n})$). Therefore, the same argumentation as for $\div \times^k$ resp. $\times \div^k$ can be applied (without the preceding \div -phase resp. \times -phase).

Hence, the number of phases such that $\mathbb{E}[|\mathbf{c}|] < |\mathbf{a}| \cdot 2^{-t} / 2$ is bounded by $O(t)$. By Markov’s inequality, $\mathbb{P}\{|\mathbf{c}| \leq |\mathbf{a}| \cdot 2^{-t}\} \geq 1/2$ after these $O(t)$ phases. If this is not the case, after all $|\mathbf{c}| \leq |\mathbf{a}|$ such that again with probability at least $1/2$, $|\mathbf{c}| \leq |\mathbf{a}| \cdot 2^{-t}$ after another $O(t)$ phases. Repeating this argument, the expected number of phases is upper bounded by $\sum_{i \geq 1} 2^{-i} \cdot i \cdot O(t) = 2 \cdot O(t)$, and the expected number of steps is $O(t \cdot n)$. \square

For other starting conditions, the (expected) number of steps necessary to ensure the theorem’s assumptions must be estimated before the theorem can be applied—for instance by estimating the number of steps until the scaling factor is halved and doubled at least once, respectively. This is a rather simple task when using the results presented.

4 Conclusion

For the first time, the (expected) runtime of a simple, but fundamental evolutionary algorithm for optimization in \mathbb{R}^n is rigorously analyzed—not a simplifying model of it. In particular, this analysis shows that, in the scenario considered, the well-known 1/5-rule for mutation adaptation indeed results in asymptotically optimal expected runtime. As the analysis covers a wide range of realizations of the 1/5-rule, it additionally yields an interesting byproduct: Fine tuning the parameters of the 1/5-rule actually does not affect the order of the expected runtime; we could even replace 1/5 by 1/8 or by 1/3, for instance. This may be interpreted as an indicator for the robustness often ascribed to evolutionary algorithms; yet it is proved for the scenario considered only.

Acknowledgments. Thanks for productive discussions and for pointing out flaws especially go to Ingo Wegener, Carsten Witt, and Stefan Droste.

References

1. Beyer, H.-G. (2001). *The Theory of Evolution Strategies*. Springer, Berlin.
2. Droste, S., Jansen, T., Wegener, I. (2002). On the analysis of the (1+1) evolutionary algorithm. *Theoretical Computer Science*, 276, pp. 51–82.
3. Ericson, T., Zinoviev, V. (2001). *Codes on Euclidian Spheres*. Elsevier, Amsterdam.
4. Giel, O., Wegener, I. (2003). Evolutionary algorithms and the maximum matching problem. *Proceedings of the 20th International Symposium on Theoretical Computer Science (STACS 2003)*, LNCS 2607, pp. 415–426.
5. Greenwood, G. W., Zhu, Q. J. (2001). Convergence in evolutionary programs with self-adaptation. *Evolutionary Computation*, 9(2), pp. 147–157.
6. Jägersküpfer, J. (2002). Analysis of a simple evolutionary algorithm for the minimization in euclidian spaces. Tech. Rep. CI-140/02, Univ. Dortmund, SFB 531, <http://sfbci.uni-dortmund.de/home/English/Publications/Reference/>.
7. Jansen, T., Wegener, I. (2001). Real royal road functions—where crossover provably is essential. *Proceedings of the 3rd Genetic and Evolutionary Computation Conference (GECCO 2001)*, Morgan Kaufmann, San Francisco, pp. 375–382.
8. Jansen, T., Wegener, I. (2002). The analysis of evolutionary algorithms—A proof that crossover really can help. *Algorithmica*, 34, pp. 47–66.
9. Mühlenbein, H. (1992). How genetic algorithms really work: Mutation and hill-climbing. *Proceedings of the 2nd Parallel Problem Solving from Nature (PPSN II)*, North-Holland, Amsterdam, pp. 15–25.
10. Rechenberg, I. (1973). *Evolutionstrategie*. Frommann-Holzboog, Stuttgart, Germany.
11. Rudolph, G. (1997). *Convergence Properties of Evolutionary Algorithms*. Verlag Dr. Kovač, Hamburg.
12. Rudolph, G. (2001). Self-adaptive mutations may lead to premature convergence. *IEEE Transactions on Evolutionary Computation*, 5(4), pp. 410–414.
13. Wegener, I. (2001). Theoretical aspects of evolutionary algorithms. *Proceedings of the 28th International Colloquium on Automata, Languages and Programming (ICALP 2001)*, LNCS 2076, pp. 64–78.
14. Wegener, I., Witt, C. (2003). On the analysis of a simple evolutionary algorithm on quadratic pseudo-boolean functions. *Journal of Discrete Algorithms*, to appear.