Available online at www.sciencedirect.com

ScienceDirect

Theoretical Computer Science 379 (2007) 329–347

Theoretical Computer Science

www.elsevier.com/locate/tcs

Algorithmic analysis of a basic evolutionary algorithm for continuous optimization[☆]

Jens Jägersküpper[*]

Department of Computer Science 2, Dortmund University, 44221 Dortmund, Germany

Abstract

In practical optimization, applying evolutionary algorithms has nearly become a matter of course. Their theoretical analysis, however, is far behind practice. So far, theorems on the runtime are limited to discrete search spaces; results for continuous search spaces are limited to convergence theory or even rely on validation by experiments, which is unsatisfactory from a theoretical point of view.

The simplest, or most basic, evolutionary algorithms use a population consisting of only one individual and use random mutations as the only search operator. Here the so-called *(1+1) evolution strategy* for minimization in $\mathbb{R}^n$ is investigated when it uses isotropically distributed mutation vectors. In particular, so-called *Gaussian mutations* are analyzed when the so-called *1/5-rule* is used for their adaptation.

Obviously, a reasonable analysis must respect the function to be minimized, and furthermore, the runtime must be measured with respect to the approximation error. A first algorithmic analysis of how the runtime depends on n, the dimension of the search space, is presented. This analysis covers all unimodal functions that are monotone with respect to the distance from the optimum. It turns out that, in the scenario considered, Gaussian mutations in combination with the $1/5$-rule indeed ensure asymptotically optimal runtime; namely, $\Theta(n)$ steps/function evaluations are necessary and sufficient to halve the approximation error.

Keywords: Runtime analysis; Evolutionary algorithms; Continuous optimization

1. Introduction

Finding an optimum of a given function $f : S \to \mathbb{R}$ is one of the fundamental problems — in theory as well as in practice. Methods for solving continuous optimization problems, essentially $S = \mathbb{R}^n$, are usually classified into first-order, second-order, and zeroth-order methods, depending on whether they utilize the gradient (the first derivative) of the objective function, the gradient and the Hessian (the second derivative), or neither of both. A zeroth-order method is also called *derivative-free* or *direct search method.* Note that here "continuous" relates to the search space rather than to f. Newton's method is the example of a second-order method; first-order methods can be (sub)classified into

[☆] Expanded and extended version of a talk presented at the International Conference on Automata, Languages, and Programming (ICALP) 2003 in Eindhoven, NL.

[*] Tel.: +49 231 755 2435; fax: +49 231 755 2047.
E-mail address: Jens.Jaegerskuepper@udo.edu.

doi:10.1016/j.tcs.2007.02.042

quasi-Newton, steepest decent, and conjugate gradient methods. Classical zeroth-order methods try to approximate the gradient and to then plug this estimate into a first-order method. Finally, amongst the "modern" zeroth-order methods, evolutionary algorithms (EAs) come into play. EAs for continuous optimization, however, are usually subsumed under the term *evolution(ary) strategies (ESs)*.

When information about the gradient is not available, for instance if f relates to a property of some workpiece and is given by simulations or even by real-world experiments, first-order (and also second-order) methods just cannot by applied. As the approximation of the gradient usually involves $\Omega(n)$ f-evaluations, a single optimization step of a classical zeroth-order method is computationally intensive, especially if f is given implicitly by simulations. In practical optimization, especially in mechanical engineering, this is often the case, and particularly in this field EAs are becoming more and more popular. However, the enthusiasm in practical EAs has led to an unclear variety of very sophisticated and problem-specific EAs. Unfortunately, from a theoretical point of view, the development of such EAs is solely driven by practical success, whereas the aspect of a theoretical analysis is left aside. In other words, (concerning EAs) theory has not kept up with practice, and thus, we should not try to analyze the most sophisticated EA en vogue, but concentrate on very basic, or call them "simple", EAs to build a sound and solid basis for EA-theory.[1]

For discrete search spaces, essentially $\{0, 1\}^n$, such a theory has been started successfully in the mid-1990s ([10], cf. [17] and [5]). There first results for non-artificial, but well-known problems have been obtained recently (namely for the maximum matching problem by Giel and Wegener [6] and for sorting and the shortest path problem by Scharnow et al. [15]). As mentioned above, the situation for continuous evolutionary optimization is different. Here, the vast majority of the results are based on empiricism, i.e., experiments are performed and their outcomes are interpreted. Also convergence properties of EAs have been studied to a considerable extent (cf. [14] and [2]). Before we take a closer look at such convergence results, however, we return to the zeroth-order optimization scenario and take a look from the complexity-theoretical point of view.

When f is given to the optimization algorithm as an oracle for f-evaluations (zeroth-order oracle) and the cost of the optimization (the runtime) is defined as the number of queries to this oracle, we are in the so-called *black-box optimization* scenario. Nemirovsky and Yudin ([11] p. 333) state in their book *Problem Complexity and Method Efficiency in Optimization:* "From a practical point of view this situation would seem to be more typical. At the same time it is objectively more complicated and it has been studied in a far less extend than the one [with first-order oracles/methods] considered earlier". After 20 years there still seems to be some truth in their statement; yet to a smaller extent, though. For discrete black-box optimization, just recently a complexity theory has been started (cf. [4] and [18]). Here lower bounds on the number of f-evaluations (*black-box complexity)* are proved with respect to classes of functions when an optimization heuristic, for instance an EA, knows about the class of functions to which f belongs, but nothing about f itself. The benefits of such results are obvious: They can prove that an allegedly poor performance of an apparently simple EA on f is not due to its simpleness, but due to the inherent black-box complexity of f.

In the mathematical programming community, however, it is common to describe the performance of an optimization method by means of convergence theory. As an example, let us take a closer look at "linear convergence". Let x^* denote the optimum of a unimodal function and x_k the approximate solution after k optimization steps. If *strong* convergence is meant, i.e. convergence in the search space,[2] then we have

$$\frac{d(x^*, x_{k+1})}{d(x^*, x_k)} \to c < 1 \quad \text{as} \quad k \to \infty$$

where $d(\cdot, \cdot)$ denotes the distance between two points in the search space. From a computer scientist's point of view, the first rub with such a result is that we do not know when k is large enough to ensure $d(x^*, x_{k+1}) \leq c' \cdot d(x^*, x_k)$ for some constant $c' < 1$. The second rub is that there seems to be no connection to n, the dimension of the search space. Only if c is an absolute constant, there is true independence of n; yet in general, the convergence rate c depends on n. When we are interested in, say, the number of steps necessary to halve the approximation error (the distance from the optimum), the order of this number with respect to n precisely depends on how c depends on n. If for instance $c = 1 - 0.5/n$, we need $\Theta(n)$ steps; if $c = 1 - 0.5/n^2$, however, we need $\Theta(n^2)$ steps — when k is large enough,

[1] in the sense of mathematics, rather than physics

[2] as apposed to *weak* convergence which means convergence in the objective space

respectively. Thus, the order of convergence ("linear" in the example above) tells us something about the "final speed" of the optimization, but in general nothing about the n-dependence of the number of steps necessary to ensure a certain approximation error. [3]

Regarding the approximation error, for unconstrained optimization it is generally not clear how the runtime can be measured solely with respect to the absolute error of the approximation. In contrast to discrete and finite problems (like CLIQUE), the initial error is generally not bounded (for CLIQUE the trivial solution consisting of a single vertex is an approximation with bounded error). Hence, the question how many steps it takes to get into the ε-ball around x^* does not make sense without specifying the starting conditions. This cannot be solved inconsiderately by restricting ourselves to the search space, say, $[-w, w]^n \subset \mathbb{R}^n$ for some w (which might depend on n). Obviously, such an assumption would put constraints on the originally unconstrained optimization problem. Hence, we must consider the runtime with respect the relative improvement of the approximation.

The simple optimization problems we will consider result in a somehow homogeneous optimization which enables us to measure the performance of the algorithm by the number of steps necessary to halve the approximation error, i.e. the Euclidean distance from the optimum. Starting at distance $\delta > 2\varepsilon$ then gives an additional factor of $\log_2(\delta/\varepsilon)$ for the number of steps necessary for an ε-approximation, i.e. $d(x^*, x_k) \leq \varepsilon$. We will come back to the point why it makes sense (in our investigations) to consider the approximation error in the "strong" sense, i.e. with respect to the search space, rather than in the "weak" sense, i.e. with respect to the objective space.

1.1. The algorithm

As mentioned in the abstract, we will concentrate on the (1+1) evolution strategy ((1+1) ES), which dates back to the mid-1960s (cf. [12]). This simple EA uses solely mutation due to a single-individual population, where here "individual" is just a synonym for "search point".[4] Let $\boldsymbol{c} \in \mathbb{R}^n$ denote the current individual. Given a starting point, i.e. an initialization of $\boldsymbol{c}$, the (1+1) ES performs the following evolution loop:

(1) Choose a random mutation vector $\boldsymbol{m} \in \mathbb{R}^n$, where the distribution of $\boldsymbol{m}$ may depend on the course of the optimization process.
(2) Generate the mutant $\boldsymbol{x} \in \mathbb{R}^n$ by $\boldsymbol{x} := \boldsymbol{c} + \boldsymbol{m}$.
(3) If $f(\boldsymbol{x}) \leq f(\boldsymbol{c})$ then $\boldsymbol{x}$ becomes the current individual ($\boldsymbol{c} := \boldsymbol{x}$), otherwise $\boldsymbol{x}$ is discarded ($\boldsymbol{c}$ unchanged).
(4) If the stopping criterion is met then output $\boldsymbol{c}$ else goto 1.

Each execution of the loop is called a *step*, and if the mutant is discarded, the mutation/mutant is said to be *rejected*, otherwise *accepted*. Also extensions of this algorithm exist: The (1+λ) ES generates λ mutants per step and (one of) the best of them competes with $\boldsymbol{c}$. The (1,λ) ES also generates λ mutants, yet (one of) the best of them replaces $\boldsymbol{c}$ regardless of whether it is better than $\boldsymbol{c}$ or not. Finally, a population consisting of $\mu > 1$ individuals can be used, which leads to $(\mu+\lambda)$ ES or (μ,λ) ES (cf. [16]). Since a worse mutant (with respect to the function to be minimized) is always discarded, the (1+1) ES is a randomized hill climber, and the selection rule is called *elitist selection.*

Note that by choosing a different, namely an appropriately randomized selection rule, the (1+1) ES may resemble the Metropolis algorithm or simulated annealing (i.e., with a fixed resp. a monotone decreasing probability the mutant is accepted irrespective of whether it is better than $\boldsymbol{c}$). Some of the results we will present hold also for such selection rules; in fact, these results hold for any selection rule. This will be stated explicitly in each case.

For a concrete application of the (1+1) ES, the stopping criterion and the distribution of the mutation vectors must be specified. Obviously, the stopping criterion is crucial in practical optimization. However, we investigate the (1+1) ES as an infinite process since we are interested in the random variable S_f over $\mathbb{N}$ defined as the number of steps the (1+1) ES performs until a predefined (relative) approximation error is realized. In particular, we are interested in $\mathsf{E}[S_f]$ and in $\mathsf{P}\{S_f \leq k\}$ for a given number of steps k.

[3] Unless c is an absolute constant; then it takes a constant number of steps to halve the distance from x^* independently of n.

[4] In general, an individual may contain additional information (which might influence the optimization process) or even more than one search point.

1.2. Gaussian mutations and the 1/5-rule for their adaptation

Fortunately, for the type of results we are after, we need not define a reasonable stopping criterion. How the mutation vectors are generated must be specified, though. Originally (cf. [12]), the mutation vector $\boldsymbol{m} \in \mathbb{R}^n$ is generated by firstly generating a so-called *Gaussian mutation (vector)* $\widetilde{\boldsymbol{m}} \in \mathbb{R}^n$ each component of which is independently standard normal distributed; subsequently, this vector is scaled by the multiplication with the so-called *mutation strength,* a scalar $s \in \mathbb{R}_{>0}$, i.e. $\boldsymbol{m} = s \cdot \widetilde{\boldsymbol{m}}$. In practice, Gaussian mutations are the most common type of mutations (for the search space $\mathbb{R}^n$).

The question that naturally arises is how the scaling factor s is to be chosen. Obviously, the smaller the approximation error is, i.e., the closer $\boldsymbol{c}$ is to an optimum, the shorter $\boldsymbol{m}$ needs to be for a further improvement of the approximation to be possible. Unfortunately, the algorithm does not know about the current approximation error, but can utilize only the knowledge obtained by f-evaluations. Based on experiments and rough calculations for two function scenarios (namely SPHERE and a corridor function), Rechenberg proposed the *1/5-(success-)rule.* The idea behind this adaptation mechanism is that in a step of the (1+1) ES the mutant should be accepted with probability 1/5. Hereinafter, a mutation that results in $f(\boldsymbol{x}) \leq f(\boldsymbol{c})$ is called *successful,* and hence, when talking about a mutation, *success probability* denotes the probability that the mutant $\boldsymbol{x} = \boldsymbol{c} + \boldsymbol{m}$ is at least as fit as $\boldsymbol{c}$. Obviously, when elitist selection is used, the success probability of a step equals the probability that the mutation is accepted in this step. If every step was successful with probability 1/5, we would observe that on average one fifth of the mutations are successful. Thus, the 1/5-rule works as follows: The optimization process is observed for n steps without changing s; if more than one fifth of the steps in this observation phase have been successful, s is doubled, otherwise, s is halved. The number of steps for observation varies in the literature, but is almost always $\Theta(n)$. Also the choice of the constants for the adaptation of s, here 2 and 1/2, seems somewhat arbitrary. In fact, one result we will obtain is that – for the function scenario considered – the asymptotic runtime is "robust" with respect to the concrete instantiation of the 1/5-rule. Namely, any 1/5-rule that performs the s-adaptation every $\Theta(n)$ steps using any two constants for the scaling of s that are greater resp. smaller than 1 results in the same asymptotic runtime; even the 1/5 can be replaced by any positive constant smaller than 1/2 in our function scenario without affecting the order of the runtime.

1.3. The function scenario

As mentioned in the abstract, we concentrate on unimodal functions that are monotone with respect to the distance from the minimum. More formally, a function $f : \mathbb{R}^n \to \mathbb{R}$ belongs to this class if

(1) a minimizer $\boldsymbol{x}^* \in \mathbb{R}^n$ exists, i.e. $f(\boldsymbol{x}^*)$ is minimal, and
(2) $d(\boldsymbol{x}^*, \boldsymbol{x}) < d(\boldsymbol{x}^*, \boldsymbol{y}) \Rightarrow f(\boldsymbol{x}) < f(\boldsymbol{y})$ for any two points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$.

The crucial property of such a function with respect to the (1+1) ES is that any mutant that is closer to the minimum is accepted, whereas any mutant that is farther away is discarded. In other words, a reduction of the approximation error is always accepted, whereas an increase is always rejected. We do not know, however, whether a mutant having the same distance from the optimum as $\boldsymbol{c}$ is accepted; yet this does not matter as we will see. All in all, when starting with the same initial approximation error, the stochastic process induced by the (1+1) ES depends on the class-defining properties, but not on the function itself.

In particular, the so-called *sphere function* defined by $\text{SPHERE}(\boldsymbol{x}) := \sum_{i=1}^{n} x_i^2$ belongs to our class, where $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Presumably, SPHERE, which is merely the L_2-norm squared, is the most investigated and discussed function in the field of ES [12,13,2,3], and it is easy to see that a function $f = g \circ L_2$ belongs to our class if $g : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is monotone increasing and bounded from below. With respect to the search space, the optimization process is independent of g; the progression of the approximation with respect to the objective space, however, crucially depends on g; consider for instance $g(x) = x^2$, i.e. $f = \text{SPHERE}$, as opposed to $g(x) = 2^x$. Results with respect to g can easily be obtained from ones with respect to the search space, and hence, it makes sense to concentrate on the later. Some of the results to be presented, however, hold independently of this function scenario; this will be stated explicitly in each case.

In the next section we are going to investigate some general properties of isotropic mutations, and in Section 3 we will see what the consequences of these properties are with respect to our concrete scenario. In the two succeeding sections these results will be used to obtain a lower bound (Section 4) and an upper bound (Section 5) on the (expected) runtime of the (1+1) ES for our function scenario. Finally, we finish with some concluding remarks in Section 6.

2. A closer look at isotropic mutations in general

The reason why we are going to have a closer look at isotropically distributed mutation vectors is that isotropy is just the crucial property of a Gaussian mutation in our analysis. Before we start with a formal definition of isotropy, however, a few notions and notations are introduced.

Definition 1. A probability $p(n)$ is *exponentially small* in n if for a constant $\varepsilon > 0$, $p(n) = \exp(-\Omega(n^{\varepsilon}))$. An event $A(n)$ happens *with overwhelming probability* (w. o. p.) with respect to n if $1 - \mathsf{P}\{A(n)\}$ is exponentially small in n.

For a vector $\boldsymbol{x} \in \mathbb{R}^n$, $|\boldsymbol{x}|$ denotes the L_2-norm of the vector, i.e. its length in Euclidean space, and $x_i \in \mathbb{R}$ its ith component. Furthermore, for instance "n-volume" abbreviates "n-dimensional volume".

Definition 2. For a random vector $\boldsymbol{x} \in \mathbb{R}^n$ let $\widehat{\boldsymbol{x}} := \boldsymbol{x}/|\boldsymbol{x}|$ denote the normalized vector. The distribution of $\boldsymbol{x}$ is *isotropic* if

(1) $\widehat{\boldsymbol{x}}$ is uniformly distributed upon the unit hyper-sphere $\{\boldsymbol{u} \in \mathbb{R}^n \mid |\boldsymbol{u}| = 1\}$,
(2) $|\boldsymbol{x}|$ is independent of $\widehat{\boldsymbol{x}}$.

Obviously, if $\boldsymbol{x}$ is isotropically distributed, the distribution of $s \cdot \boldsymbol{x}$ for $s \in \mathbb{R}_{>0}$ is also isotropic. Thus, to prove that a scaled Gaussian mutation vector is isotropically distributed, we must merely show that the distribution of $\widetilde{\boldsymbol{m}}$ (cf. Section 1.2) is isotropic. As each component of $\widetilde{\boldsymbol{m}}$ is independently standard normal distributed, the density at the point $\boldsymbol{x} \in \mathbb{R}^n$ equals $\prod_{i=1}^{n} \exp(-{x_i}^2/2)/\sqrt{2\pi} = \exp(-|\boldsymbol{x}|^2/2)/\sqrt{2\pi}$. In other words, the distribution of $\widetilde{\boldsymbol{m}}$ is invariant with respect to orthonormal transformations, and it is easy to see that this implies the isotropy of the distribution:

Proposition 3. *A scaled Gaussian mutation vector is isotropically distributed.*

In less formal words, for an isotropically distributed vector, all directions are equiprobable (in fact "equidense"), and furthermore, the distribution of its length is independent of its direction. This enables us to assume that an isotropically distributed vector $\boldsymbol{m}$ is generated in two independent steps: First ℓ is chosen according to the distribution of $|\boldsymbol{m}|$ (which is well defined due to the isotropy); subsequently, by a *deferred decision*, $\boldsymbol{m}$ is chosen uniformly upon the hyper-sphere $S_\ell := \{\boldsymbol{x} \in \mathbb{R}^n \mid |\boldsymbol{x}| = \ell\}$. This decomposition will turn out very useful in our investigations of a mutation. After ℓ is chosen according to the distribution of $|\boldsymbol{m}|$ in a step of the (1+1) ES, the mutant $\boldsymbol{x} = \boldsymbol{c} + \boldsymbol{m}$ is only partially random since we know that it will be located in the hyper-sphere $S_{\boldsymbol{c},\ell} := \{\boldsymbol{x} \in \mathbb{R}^n \mid d(\boldsymbol{x}, \boldsymbol{c}) = \ell\}$. Namely, under the condition $\{|\boldsymbol{m}| = \ell\}$ the mutant $\boldsymbol{x}$ is uniformly distributed upon $S_{\boldsymbol{c},\ell}$.

Let $A_{\boldsymbol{c}} := \{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) \le f(\boldsymbol{c})\}$ (the lower level set of $\boldsymbol{c}$ with respect to f). Then the success probability of a step equals the mass of $A_{\boldsymbol{c}}$ with respect to the probability measure induced by the distribution of the mutant. For instance, we could also chose $A_{\boldsymbol{c}}$ to denote the set containing all points that are at least $\Delta > 0$ closer to the optimum than $\boldsymbol{c}$. Then the mass of $A_{\boldsymbol{c}}$ would give us the probability that the approximation error of the mutant is by at least Δ smaller than the one of $\boldsymbol{c}$. In the following, $A_{\boldsymbol{c}}$ may denote an arbitrary set so that the knowledge of $\mathsf{P}\{\boldsymbol{c} + \boldsymbol{m} \in A_{\boldsymbol{c}}\}$ helps us with the analysis of the (1+1) ES.

A direct algebraic computation of this *hitting probability*, i.e. the mass of $A_{\boldsymbol{c}}$ with respect to the probability measure induced by the distribution of $\boldsymbol{m}$, has turned out to be very difficult (cf. [2]) even for simple situations. This is usually the point where simplifications of the model are conceded, which leads to the undesirable need for experimental validations of whether the results obtained in this simplified model really predict the behavior of the true algorithm well.[5] Thus, we are going to take a "less is sometimes more" approach here: We will be comfortable with less accuracy, namely with asymptotic estimates of probabilities, and by this, we will gain more insight into the asymptotic n-dependence of the runtime of the (1+1) ES.

As we have seen above, if the mutation vector takes the length ℓ, the mutant is uniformly distributed upon $S_{\boldsymbol{c},\ell}$, in other words:

$$\mathsf{P}\{\boldsymbol{c} + \boldsymbol{m} \in A_{\boldsymbol{c}} \mid |\boldsymbol{m}| = \ell\} = \frac{(n-1)\text{-volume of } S_{\boldsymbol{c},\ell} \cap A_{\boldsymbol{c}}}{(n-1)\text{-volume of } S_{\boldsymbol{c},\ell}}.$$

[5] Unfortunately, in most works a formal definition of the measure for "well" is omitted and experiments are not validated statistically.

In general, even the algebraic computation of the volume of $S_{\boldsymbol{c},\ell} \cap A_{\boldsymbol{c}}$ might be tricky or a closed formula might not exist at all. However, if we can pick a hyper-plane H that separates $A_{\boldsymbol{c}}$ from $\boldsymbol{c}$, we obtain an upper bound on the (conditional) hitting probability (given that $\{|\boldsymbol{m}| = \ell\}$). Let $C_{\boldsymbol{c},\ell,H} \subset S_{\boldsymbol{c},\ell}$ denote the hyper-spherical cap whose boundary is given by $H \cap S_{\boldsymbol{c},\ell}$ and which lies in the half-space containing $A_{\boldsymbol{c}}$. If $d(H, \boldsymbol{c}) := \inf_{\boldsymbol{x}\in H} d(\boldsymbol{x}, \boldsymbol{c}) > \ell$, then the hyper-plane and the hyper-sphere do not intersect, and consequently, the set $C_{\boldsymbol{c},\ell,H}$ is empty and the mutation cannot hit $A_{\boldsymbol{c}}$. If $d(H, \boldsymbol{c}) = \ell$ then $C_{\boldsymbol{c},\ell,H}$ is a singleton, i.e. degenerated, and is hit only with zero probability. Finally, if $d(H, \boldsymbol{c}) < \ell$ then $C_{\boldsymbol{c},\ell,H}$ has a positive $(n-1)$-volume, and thus:

$$\mathsf{P}\{\boldsymbol{c} + \boldsymbol{m} \in A_{\boldsymbol{c}} \mid |\boldsymbol{m}| = \ell\} \le \frac{(n-1)\text{-volume of } C_{\boldsymbol{c},\ell,H}}{(n-1)\text{-volume of } S_{\boldsymbol{c},\ell}}.$$

On the other hand, if we can pick a hyper-plane G such that each point in the cap also lies in $A_{\boldsymbol{c}}$, i.e. $C_{\boldsymbol{c},\ell,G} \subseteq A_c$, then we obtain a lower bound on the conditional hitting probability:

$$\mathsf{P}\{\boldsymbol{c} + \boldsymbol{m} \in A_{\boldsymbol{c}} \mid |\boldsymbol{m}| = \ell\} \ge \frac{(n-1)\text{-volume of } C_{\boldsymbol{c},\ell,G}}{(n-1)\text{-volume of } S_{\boldsymbol{c},\ell}}.$$

Needless to say, these bounds might by arbitrarily weak depending on the shape of the hyper-surface $A_{\boldsymbol{c}} \cap S_{\boldsymbol{c},\ell}$. When applying these bounds, however, H and G should be chosen such that $d(H, \boldsymbol{c})$ is as large as possible and $d(G, \boldsymbol{c})$ as small as possible.

The considerations just presented are not limited to Gaussian mutations and our function scenario, they are only specific to generating a mutant by adding an isotropically distributed vector. In fact, for our function scenario it turns out that we can choose $H = G$, i.e., the lower and the upper bound meet. Therefore, we now concentrate on the ratio of the hyper-surface area of a hyper-spherical cap to the one of the hyper-sphere it is cut off by the intersection with some hyper-plane J. In particular, we are interested in how this ratio depends on the height of the cap and on n, the dimension of the search space.

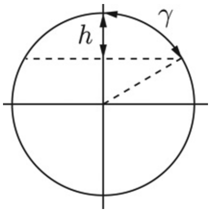
In the following, polar/spherical coordinates are used: Let r denote the distance from the origin $\boldsymbol{o}$, α the azimuthal angle with range $[0, 2\pi)$, and $\beta_3, \ldots, \beta_n$ the remaining angles with range $[0, \pi]$. Here, for a given $\boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{o}\}$, β_i is the angle between (the positive half of) the x_i-axis and the half-line starting at $\boldsymbol{o}$ and passing through $\boldsymbol{x}$. Let ρ denote an arbitrary permutation on $\{3, \ldots, n\}$. Fixing r in n-space, but none of the angles, defines an n-sphere $S_r^{(n)}$ with radius r; additionally fixing $\beta_{\rho(n)}$ results in an $(n-1)$-sphere $S_r^{(n-1)} \subseteq S_r^{(n)}$ having radius $r \sin \beta_{\rho(n)}$; fixing $\beta_{\rho(n-1)}$ in addition to r and $\beta_{\rho(n)}$ results in an $(n-2)$-sphere $S_r^{(n-2)} \subseteq S_r^{(n-1)} \subseteq S_r^{(n)}$ with radius $r \sin \beta_{\rho(n)} \sin \beta_{\rho(n-1)}$, and so on (cf. [9]). Thus, assuming $\rho = \mathrm{id}$ for notational convenience, the hyper-surface area of an n-sphere with radius r is given by

$$\int_{\beta_n=0}^{\pi} \int_{\beta_{n-1}=0}^{\pi} \cdots \int_{\beta_3=0}^{\pi} \int_{\alpha=0}^{2\pi} (r \sin \beta_n \cdots \sin \beta_3 \mathrm{d}\alpha)(r \sin \beta_n \cdots \sin \beta_4 \mathrm{d}\beta_3) \cdots \quad \cdots (r \sin \beta_n \mathrm{d}\beta_{n-1})(r \mathrm{d}\beta_n).$$

Regrouping the factors and solving the α-integral, namely $\int_0^{2\pi} \mathrm{d}\alpha = 2\pi$, yields

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

for the area of an n-sphere with radius r. Naturally, we could have looked up the formula for the hyper-surface area of an n-sphere in a formulary, but we also need a formula for the cap. The formula for the cap can easily be derived from the one above — yet only if one knows about the derivation of the later.



The area of an n-dimensional spherical cap is calculated by adjusting the upper limit on $\beta_{\rho(n)}$ appropriately. In the figure on the right, the interdependence between the upper limit (γ) on $\beta_{\rho(n)}$ and the height (h) of a spherical cap is shown (where the sheet this figure is drawn on corresponds to the x_1–$x_{\rho(n)}$-plane when $\alpha = 0$). Consequently, the area of a hyper-spherical cap with radius r and height $h = r \cdot (1 - \cos\gamma) \in [0, 2r]$ is given by

$$r^{n-1} \cdot 2\pi \cdot \left(\int_0^{\gamma} (\sin\beta)^{n-2} \mathrm{d}\beta \right) \cdot \left(\prod_{i=1}^{n-3} \int_0^{\pi} (\sin\beta)^{i} \mathrm{d}\beta \right).$$

All in all, in n-space, $n \geq 3$, the ratio of the hyper-surface area of a spherical cap with height $h \in [0, 2r]$ to the one of the sphere with radius r the cap is cut off reduces to

$$\frac{\int_0^{\gamma} (\sin\beta)^{n-2} \mathrm{d}\beta}{\int_0^{\pi} (\sin\beta)^{n-2} \mathrm{d}\beta} \quad \text{with} \quad \gamma = \arccos(1 - h/r).$$

Let $g := r - h$ denote the distance between the center of the sphere and the cutting hyper-plane, i.e. the hyper-plane containing the boundary of the cap. Then $\gamma = \arccos(g/r)$. In the bounds on the conditional hitting probability $\mathsf{P}\{\boldsymbol{m} \in A_{\boldsymbol{c}} \mid |\boldsymbol{m}| = \ell\}$, the radius r is given by ℓ, and g is given by the distance between $\boldsymbol{c}$ and H resp. G.

We use the abbreviation $\Psi_k(\alpha) := \int_0^{\alpha} (\sin\beta)^k \mathrm{d}\beta$. In Appendix A it is shown that for $n \geq 4$

$$\sqrt{\frac{2\pi}{n-1}} < \int_0^{\pi} (\sin x)^{n-2} \mathrm{d}x = \Psi_{n-2}(\pi) < \sqrt{\frac{2\pi}{n-2}},$$

and hence, $\Psi_{n-2}(\pi) = \Theta(1/\sqrt{n})$. Unfortunately, solving $\Psi_{n-2}(\gamma)$ algebraicly results in a formula that does not reveal the n-dependence of the integral's value. With respect to the 1/5-rule which we want to analyze, it is particularly interesting which situation results in a hitting probability of 1/5. Obviously, when $|\boldsymbol{m}| = \ell$, this is the case if $\Psi_{n-2}(\arccos(d(J, \boldsymbol{c})/\ell)) = \Psi_{n-2}(\pi)/5$ (recall that J denotes the cutting hyper-plane such that $\boldsymbol{c}$ is located in the one half-space and the cap in the other one); yet this implicit result does not help much.

To estimate the probability that an isotropic mutation (which takes the length ℓ) hits a certain cap, we will now transform the formula just derived into one which makes such an estimation simple and which will turn out useful also in the forthcoming analysis of the spatial gain towards the optimum. Namely, we will concentrate on the probability density of hitting the boundary of the cap. With the help of this density, we will obtain an alterative formula for the probability of hitting a cap.

The distance between $\boldsymbol{c}$ and J, namely g, can be interpreted as the minimum spatial gain (measured perpendicular to J) which a mutation $\boldsymbol{m}$ with $|\boldsymbol{m}| = \ell$ yields when it hits this cap. Therefore, let G denote the random variable given by the spatial gain of an isotropic mutation $\boldsymbol{m}$ measured parallel to a fixed direction under the condition $|\boldsymbol{m}| = \ell$, i.e., the distance between $\boldsymbol{c}$ and the hyper-plane that contains its mutant $\boldsymbol{c}+\boldsymbol{m}$ and lies perpendicular to the direction. Then

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

In other words, for $x \in [-\ell, \ell]$

$$F_n(x) := 1 - \frac{\Psi_{n-2}(\arccos(x/\ell))}{\Psi_{n-2}(\pi)}$$

is the probability distribution of G over $[-\ell, \ell]$ in n-space. As Ψ_k is continuous, the probability density of G at $g \in [-\ell, \ell]$ is given by $F_n'(g) = \frac{\mathrm{d}F_n(x)}{\mathrm{d}x}(g)$,

$$\begin{aligned} F_n'(x) &= \frac{-1}{\Psi_{n-2}(\pi)} \cdot \frac{\mathrm{d}\,\Psi_{n-2}(\arccos(x/\ell))}{\mathrm{d}x} \\ &= \frac{-1}{\Psi_{n-2}(\pi)} \cdot \frac{\mathrm{d}}{\mathrm{d}x} \int_0^{\arccos(x/\ell)} (\sin\beta)^{n-2} \mathrm{d}\beta. \end{aligned}$$

Since $\sin^k$ is continuous, let Sin_k denote its anti-derivative, i.e. $\mathrm{Sin}_k' = \sin^k$, such that $\mathrm{Sin}_k(0) = 0$. Then

$$\begin{aligned} \frac{\mathrm{d}}{\mathrm{d}y} \int_0^{\arccos(y)} (\sin\beta)^k \mathrm{d}\beta &= \frac{\mathrm{d}\mathrm{Sin}_k(\arccos y)}{\mathrm{d}y} \\ &= \mathrm{Sin}_k'(\arccos y) \cdot \arccos' y \\ &= (\sin\arccos y)^k \cdot \arccos' y, \end{aligned}$$

and since $\sin\arccos y = \sqrt{1-y^2}$ and $\arccos' y = -1/\sqrt{1-y^2}$, we have for $k \geq 2$

$$\frac{\mathrm{d}}{\mathrm{d}y}\mathrm{Sin}_k(\arccos y) = \left(1-y^2\right)^{k/2} \cdot \frac{-1}{\sqrt{1-y^2}} = -1 \cdot \left(1-y^2\right)^{(k-1)/2}.$$

By applying the chain rule, we finally obtain the probability density of G at $g \in [-\ell, \ell]$ in n-space, $n \geq 4$:

$$F_n'(g) = \frac{-1}{\Psi_{n-2}(\pi)} \cdot \frac{-1}{\ell} \cdot (1-(g/\ell)^2)^{(n-3)/2} = \frac{1}{\Psi_{n-2}(\pi)\cdot\ell}\left(1-(g/\ell)^2\right)^{(n-3)/2}.$$

This density function can now be used to derive the announced alternative formula for the conditional probability that a certain cap is hit by an isotropic mutation:

$$\begin{aligned}
\mathsf{P}\{\boldsymbol{m} \in C_{\boldsymbol{c},\ell,J} \mid |\boldsymbol{m}| = \ell\} &= \frac{\Psi_{n-2}(\arccos(d(J,\boldsymbol{c})/\ell))}{\Psi_{n-2}(\pi)} \\
&= \mathsf{P}\{G \geq d(J,\boldsymbol{c})\} \\
&= \int_{d(J,\boldsymbol{c})}^{\ell} F_n'(g)\mathrm{d}g \\
&= \int_{d(J,\boldsymbol{c})}^{\ell} \frac{1}{\Psi_{n-2}(\pi)\cdot\ell}(1-(g/\ell)^2)^{(n-3)/2}\mathrm{d}g \\
(x \text{ substitutes } g/\ell) \qquad &= \frac{1}{\Psi_{n-2}(\pi)}\int_{d(J,\boldsymbol{c})/\ell}^{1}(1-x^2)^{(n-3)/2}\mathrm{d}x.
\end{aligned}$$

Note that, by definition, the height of $C_{\boldsymbol{c},\ell,J}$ is smaller than ℓ since $\boldsymbol{c}$ lies in the one half-space (with respect to J) and the cap in the other one.

Obviously, for a fixed condition $|\boldsymbol{m}| = \ell > 0$, if $d(J,\boldsymbol{c}) \to 0$, then the cap becomes a hyper-hemisphere and the hitting probability approaches 1/2. With respect to the 1/5-rule, which we want to investigate, it is of particular interest when a cap is hit with probability 1/5. By the alternative formula just derived, we obtain the following results.

Lemma 4. *Let $\boldsymbol{m} \in \mathbb{R}^n$ be isotropically distributed. Then*

(1) *for any constant κ with $0 < \kappa < 1/2$, for instance $\kappa = 1/5$,*

$$\mathsf{P}\{\boldsymbol{m} \in C_{\boldsymbol{c},\ell,J} \mid |\boldsymbol{m}| = \ell\} = \kappa \quad \Rightarrow \quad d(J,\boldsymbol{c}) = \Theta(\ell/\sqrt{n});$$

(2) *there exists a constant $\delta > 0$ such that*

$$d(J,\boldsymbol{c}) = \Theta(\ell/\sqrt{n}) \quad \Rightarrow \quad \delta \leq \mathsf{P}\{\boldsymbol{m} \in C_{\boldsymbol{c},\ell,J} \mid |\boldsymbol{m}| = \ell\} \leq 1/2 - \delta$$

for n sufficiently large[6]*;*

(3) *for any constant $\varepsilon > 0$*

$$d(J,\boldsymbol{c}) = \Omega(\ell \cdot n^{\varepsilon-1/2}) \quad \Rightarrow \quad \mathsf{P}\{\boldsymbol{m} \in C_{\boldsymbol{c},\ell,J} \mid |\boldsymbol{m}| = \ell\} \text{ is exponentially small in } n.$$

Proof. Note that $(1-1/n)^n < \exp(-1) < (1-1/n)^{n-1}$ for $n > 1$. Using the formula just derived, to prove the first two claims it is sufficient to show that for $d(J,\boldsymbol{c}) = \Theta(\ell/\sqrt{n})$ both

$$\int_0^{d(J,\boldsymbol{c})/\ell}(1-x^2)^{(n-3)/2}\mathrm{d}x \qquad \text{as well as} \qquad \int_{d(J,\boldsymbol{c})/\ell}^{1}(1-x^2)^{(n-3)/2}\mathrm{d}x$$

are $\Omega(\Psi_{n-2}(\pi))$, i.e. $\Omega(1/\sqrt{n})$, respectively. Therefore, let β be such that $\beta/\sqrt{n} = d(J,\boldsymbol{c})/\ell = \Theta(1/\sqrt{n})$, i.e., $\beta = \Theta(1)$. Since $(1-x^2)^{(n-3)/2}$ is non-increasing on $[0,1]$, for $n \geq 4$ and $2\beta/\sqrt{n} \leq 1$, i.e., for $n \geq \max\{4, 4\beta^2\}$

$$\begin{aligned}
\int_0^{\beta/\sqrt{n}}(1-x^2)^{(n-3)/2}\mathrm{d}x &\geq (\beta/\sqrt{n})\cdot(1-(\beta/\sqrt{n})^2)^{(n-3)/2} \\
&= \Theta(1/\sqrt{n})\cdot(1-\Theta(1/n))^{\Theta(n)} \\
&= \Theta(1/\sqrt{n})\cdot\exp(-\Theta(1))
\end{aligned}$$

[6] That is, a constant n_0 exists such that the statement holds for $n \geq n_0$; here n_0 depends only on the concealed constants in $\Theta(\ell/\sqrt{n})$.

and

$$\begin{aligned}\int_{\beta/\sqrt{n}}^{1}(1-x^2)^{(n-3)/2}\mathrm{d}x &\ge \int_{\beta/\sqrt{n}}^{2\beta/\sqrt{n}}(1-x^2)^{(n-3)/2}\mathrm{d}x\\ &\ge (\beta/\sqrt{n})\cdot(1-(2\beta/\sqrt{n})^2)^{(n-3)/2}\\ &= \Theta(1/\sqrt{n})\cdot(1-\Theta(1/n))^{\Theta(n)}\\ &= \Theta(1/\sqrt{n})\cdot\exp(-\Theta(1)).\end{aligned}$$

To prove the third claim it suffices to show that $\int_{d(J,\boldsymbol{c})/\ell}^{1}(1-x^2)^{(n-3)/2}\mathrm{d}x$ is exponentially small if $d(J,\boldsymbol{c}) = \Omega(\ell\cdot n^{\varepsilon-1/2})$ for a positive constant ε. Since $d(J,\boldsymbol{c})\ge\ell$ results in the degeneration of the cap, w. l. o. g. $\varepsilon\le 1/2$. Thus, $d(J,\boldsymbol{c})/\ell\ge\beta\cdot n^{\varepsilon-1/2}$ for a positive constant $\beta\le 1$ such that $0<\beta\cdot n^{\varepsilon-1/2}\le 1$, and

$$\begin{aligned}\int_{\beta\cdot n^{\varepsilon-1/2}}^{1}(1-x^2)^{(n-3)/2}\mathrm{d}x &\le (1-\beta^2 n^{2\varepsilon}/n)^{(n-3)/2}\cdot(1-\beta\cdot n^{\varepsilon-1/2})\\ &= \exp(-\Omega(n^{2\varepsilon}))\cdot\Theta(1)\end{aligned}$$

completes the proof. □

These results on the probability of hitting a cap of a certain height when $|\boldsymbol{m}|=\ell$ will now be used to investigate the spatial gain which a step of the (1+1) ES yields. Recall that hitting a cap of height $h\in[0,\ell]$ (defined by a hyperplane J) is just the same as obtaining a spatial gain of $g=\ell-h\ge 0$ parallel to a predefined direction (for instance perpendicular to J).

The spatial gain towards a fixed point $\boldsymbol{x}^*$ is given by $d(\boldsymbol{x}^*,\boldsymbol{c})-d(\boldsymbol{x}^*,\boldsymbol{c}+\boldsymbol{m})$. Now, note this simple, but crucial geometric fact: If the spatial gain towards $\boldsymbol{x}^*$ is non-negative then this gain is upper bounded by the respective gain measured parallel to the line passing through $\boldsymbol{c}$ and $\boldsymbol{x}^*$. Thus, the third claim of the preceding lemma directly yields the following result.

Corollary 5. *The spatial gain of an isotropic mutation $\boldsymbol{m}\in\mathbb{R}^n$ towards a predefined point in the search space, for instance the/an optimum, is for any constant $\varepsilon>0$ w. o. p. $O(|\boldsymbol{m}|\cdot n^{\varepsilon-1/2})$.*

When we optimistically assume that a mutation that reduces the approximation error was always accepted and that a mutation that increases the approximation error was always rejected, we obtain the following result.

Corollary 6. *Let the (1+1) ES optimize an arbitrary function using an arbitrary selection rule. The spatial gain towards a predefined point in the search space in a step, in which the mutation vector $\boldsymbol{m}\in\mathbb{R}^n$ is isotropically distributed, is for any constant $\varepsilon>0$ w. o. p. $O(|\boldsymbol{m}|\cdot n^{\varepsilon-1/2})$.*

Also the efforts to derive the density function of hitting the boundary of a cap pay off now. With the help of this density we obtain a bound on the expected spatial gain towards a fixed point.

Lemma 7. *Irrespective of the function that is optimized and of the selection rule, the expected spatial gain towards a predefined point in the search space in a step of the (1+1) ES, in which $\boldsymbol{m}$ is isotropically distributed, is upper bounded by $|\boldsymbol{m}|/\sqrt{2\pi(n-1)}$, i.e., it is $O(|\boldsymbol{m}|/\sqrt{n})$.*

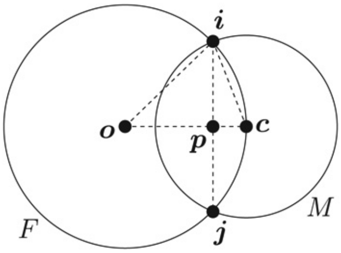
Proof. Optimistically assume that a negative gain is always rejected and that a positive gain is always accepted. Using the respective "parallel spatial gain" as an upper bound, when $\boldsymbol{m}$ takes the length ℓ the expected spatial gain is upper bounded by

$$\begin{aligned}\int_0^{\ell} g\cdot F'_{n-2}(g)\mathrm{d}g &= \frac{\ell}{\Psi_{n-2}(\pi)}\int_0^1 y\cdot\left(1-y^2\right)^{(n-3)/2}\mathrm{d}y\\ &= \frac{\ell}{\Psi_{n-2}(\pi)}\cdot\left[\frac{1}{n-1}\left(1-y^2\right)^{(n-1)/2}\right]_0^1\\ &< \frac{\ell\cdot\sqrt{n-1}}{\sqrt{2\pi}}\cdot\frac{1}{n-1}\end{aligned}$$

since $\Psi_{n-2}(\pi)>\sqrt{2\pi/(n-1)}$ for $n\ge 4$ (cf. Appendix A). □

Note (again) that the results obtained in this section are specific to isotropic mutations, but not to the function that is optimized nor to the selection rule that is used. In the next section we return to the function scenario described in the introduction and obtain more concrete results.

3. Isotropic mutations and the 1/5-rule in the concrete scenario



As mentioned above, we are now going to turn our attention to the function scenario described in the introduction. The situation is shown in the figure on the right. The left sphere $F := \{c' \in \mathbb{R}^n \mid d(c', o) = d(c, o)\}$ will be called *fitness sphere* since the properties of f imply that all points inside this fitness sphere are better than the current search point c, and that all points outside are worse. Consequently, due to the elitist selection, in our scenario a mutant that is closer to the optimum than c, i.e. a positive spatial gain towards the optimum, is accepted, whereas a mutant that is farther away, i.e. a negative spatial gain towards the optimum, is rejected. The potential mutants when the mutation vector m takes the length ℓ form the *mutation sphere* $M := \{x \in \mathbb{R}^n \mid d(x, c) = \ell\}$. Obviously, the intersection $I = F \cap M \subset \mathbb{R}^n$ of these two spheres is empty if $\ell > 2 \cdot d(c, o)$, and it is a singleton if $\ell = 2 \cdot d(c, o)$. The interesting case which we will focus on in the following is when $\ell < 2 \cdot d(c, o)$. Then, if the mutant is accepted in a step, this step yields a non-negative spatial gain towards the optimum, and the question is: How large is this gain and for which ℓ is it maximum? As we want to investigate the 1/5-rule for the adaptation of $|m|$, it is particularly interesting what the gain is when a mutant is accepted with probability 0.2.

Since for all $i, j \in I$, $d(i, c) = d(j, c)$ as well as $d(i, o) = d(j, o)$, I is a subset of a hyper-plane which is perpendicular to the line passing through the optimum o and the current search point c. Let P denote this hyper-plane and $p \in P$ the point where the line penetrates the hyper-plane. Let $C \dot{\cup} I \dot{\cup} D = M$, i.e., C and D are hyper-spherical caps the missing boundary of each of which is I; let C denote the cap inside the fitness sphere. According to the function scenario, if a mutant having distance ℓ from the current search point hits C then it is accepted, and if it hits D, it is rejected. If this mutant hits I, however, our assumptions on the properties of f tell us nothing about whether it is accepted or not, yet this is no problem: This event has zero probability since I, an $(n-1)$-sphere, has no $(n-1)$-volume. Following the argumentation presented in the preceding section, for the function scenario considered we have:

$$\mathsf{P}\{\text{mutation is accepted} \mid |m| = \ell\} = \frac{(n-1)\text{-volume of } C_{c,\ell,P}}{(n-1)\text{-volume of } S_{c,\ell}}.$$

The ratio on right has already been investigated in the preceding section; to use the results obtained there, we must know the height of the cap C which is given by $\ell - d(P, c)$. Note that the distance between the cutting hyper-plane P and the current search point, namely $d(P, c)$, equals $d(p, c)$. Due to Pythagoras, when $d(c, p) \le d(c, o)$ then for any $i \in I$

$$d(o, i)^2 - d(o, p)^2 = d(i, p)^2 = d(c, i)^2 - d(c, p)^2.$$

Since $d(o, i) = d(o, c)$, $d(o, p) = d(o, c) - d(c, p)$, and $d(c, i) = \ell$,

$$d(o, c)^2 - (d(o, c) - d(c, p))^2 = \ell^2 - d(c, p)^2,$$

and solving for $d(c, p)$ yields $d(c, p) = \ell^2/(2 \cdot d(o, c))$ for the distance between c and P. A similar argumentation yields that this equality holds also for $d(c, o) < d(c, p) \le 2 \cdot d(o, c)$, i.e., if the optimum o and the current search point c are not separated by P, but lie in the same half-space.

Now that we know how the distance of the cutting hyper-plane and the current search point depends on ℓ (the length of the mutation vector) and on $d(o, c)$ (the current approximation error) we can apply Lemma 4.

Lemma 8. *In the scenario considered, when a suboptimal search point* $\boldsymbol{c}$ *is mutated using an isotropically distributed vector* $\boldsymbol{m}$ *then under the condition* $|\boldsymbol{m}| = \ell$

(1) *if the mutation is successful with a constant probability* κ, $0 < \kappa < 1/2$, *then* $\ell = \Theta(d(\boldsymbol{o}, \boldsymbol{c})/\sqrt{n})$;
(2) *if* $\ell = \Theta(d(\boldsymbol{o}, \boldsymbol{c})/\sqrt{n})$, *then a positive constant* δ *exist such that for sufficiently large* n *the mutation is successful with a probability in* $[\delta, 1/2 - \delta]$;
(3) *if* $\ell = \Omega(d(\boldsymbol{o}, \boldsymbol{c}) \cdot n^{\varepsilon - 1/2})$ *for a positive constant* ε, *then* $\boldsymbol{m} \notin C$ *w.o.p., i.e., the success probability is exponentially small.*

Proof. According to Lemma 4, $d(P, \boldsymbol{c}) = \Theta(\ell/\sqrt{n})$ and according to the preceding argumentation specific to our scenario $d(P, \boldsymbol{c}) = \ell^2/(2 \cdot d(\boldsymbol{o}, \boldsymbol{c}))$. Thus, we have $\ell^2/(2 \cdot d(\boldsymbol{o}, \boldsymbol{c})) = \Theta(\ell/\sqrt{n})$, i.e., $\ell = \Theta(d(\boldsymbol{o}, \boldsymbol{c})/\sqrt{n})$.

For the proof of the third claim we have $d(P, \boldsymbol{c}) = \ell^2/(2 \cdot d(\boldsymbol{o}, \boldsymbol{c})) = \ell \cdot \Omega(n^{\varepsilon}/\sqrt{n})$ so that the third part of Lemma 4 completes the proof. □

The preceding lemma says that in our scenario the length of the mutation vector is by a factor of $\Theta(1/\sqrt{n})$ smaller than the current approximation error if the 1/5-rule is able to ensure a success probability close to 1/5 in the respective step. Thus, we investigate the (expected) spatial gain towards the optimum in this situation next.

Therefore, assume $|\boldsymbol{m}| = \ell = \Theta(d(\boldsymbol{o}, \boldsymbol{c})/\sqrt{n})$. Then

$$d(P, \boldsymbol{c}) = \ell^2/(2 \cdot d(\boldsymbol{o}, \boldsymbol{c})) = \ell \cdot \Theta(1/\sqrt{n}) = \Theta(d(\boldsymbol{o}, \boldsymbol{c})/n),$$

i.e., the distance between the current search point $\boldsymbol{c}$ and the hyper-plane P that separates the cap C (which contains only acceptable mutants) from the cap D (which contains only mutants that are rejected) is an $\Theta(1/n)$-fraction of the current approximation error. Just as in the proof of the preceding lemma, Lemma 4 tells us that C is hit by the mutation with probability $\Omega(1)$. Now consider the hyper-plane P' that is parallel to P and by $d(P, \boldsymbol{c})$ closer to the optimum than P, i.e., $d(P', \boldsymbol{c}) = 2 \cdot d(P, \boldsymbol{c})$. Let $C' \subset C \subset M$ denote the cap defined by P'. Then any point $\boldsymbol{x} \in C'$ is at least $d(P', P) = d(P, \boldsymbol{c}) = \Theta(d(\boldsymbol{o}, \boldsymbol{c})/n)$ closer to the optimum than a point $\boldsymbol{i} \in I = F \cap M \subset P$. Since $\boldsymbol{i}$ has the same distance from the optimum as $\boldsymbol{c}$, a point in C' is in fact by $\Omega(d(\boldsymbol{o}, \boldsymbol{c})/n)$ closer to the optimum than the current search point $\boldsymbol{c}$. Furthermore, the cap C' is also/still hit with probability $\Omega(1)$ because $d(P', \boldsymbol{c})$ is $\Theta(\ell/\sqrt{n})$. All in all, we have proved the following result.

Lemma 9. *In the function scenario considered, when a suboptimal point* $\boldsymbol{c}$ *is mutated by an isotropic mutation vector that has length* $\ell = \Theta(d(\boldsymbol{o}, \boldsymbol{c})/\sqrt{n})$, *then the spatial gain of this mutation towards the optimum is* $\Omega(d(\boldsymbol{o}, \boldsymbol{c})/n)$ *with probability* $\Omega(1)$.

When we additionally take in account that elitist selection is used, this result can easily be extended with respect to the expectation of the spatial gain in a step of the (1+1) ES.

Lemma 10. *Let the* (1+1) *ES use isotropic mutations end elitist selection to minimize a function in the considered scenario. If in a step the mutation vector takes a length* $\ell = \Theta(d(\boldsymbol{o}, \boldsymbol{c})/\sqrt{n})$, *then the expected spatial gain towards the optimum in this step is* $\Theta(d(\boldsymbol{o}, \boldsymbol{c})/n)$, *i.e., the approximation error is expected to decrease by a* $\Theta(1/n)$*-fraction.*

Proof. Due to elitist selection, in our function scenario a negative spatial gain towards the optimum, i.e. $\boldsymbol{c}+\boldsymbol{m} \in D$, is always rejected. Thus, the expected spatial gain is lower bounded by $\Omega(1) \cdot \Omega(d(\boldsymbol{o}, \boldsymbol{c})/n)$ according to the preceding lemma. That it is also $O(d(\boldsymbol{o}, \boldsymbol{c})/n)$ directly follows from the general upper bound given in Lemma 7. □

In the results just shown, we have assumed that the length of the mutation vector is $\Theta(d(\boldsymbol{o}, \boldsymbol{c})/\sqrt{n})$ since this is the idea behind the 1/5-rule as we have seen. When the length of the mutation vector is considerably smaller, then the general upper bound on the expected spatial gain (Lemma 7) also yields a smaller bound on the expected spatial gain in our concrete scenario. For instance, if $|\boldsymbol{m}| = \Theta(d(\boldsymbol{o}, \boldsymbol{c})/n)$, the expected spatial gain towards the optimum is $O(d(\boldsymbol{o}, \boldsymbol{c})/n^{1.5})$. Naturally, the next question one might ask is whether a mutation vector with a length that is considerably larger than $\Theta(d(\boldsymbol{o}, \boldsymbol{c})/\sqrt{n})$ leads to an expected gain towards the optimum that is larger than $\Theta(d(\boldsymbol{o}, \boldsymbol{c})/n)$. The general upper bound does not give the answer.

Since for $|\boldsymbol{m}| > 2 \cdot d(\boldsymbol{o}, \boldsymbol{c})$ the spatial gain towards the optimum is negative (and therefore rejected when elitist selection is used in our scenario), we assume $|\boldsymbol{m}| \le 2 \cdot d(\boldsymbol{o}, \boldsymbol{c})$ in the following. For the general upper bound, in fact, we computed the expectation of the spatial gain measured parallel to some fixed direction under the restriction that this gain is positive (we zeroed out negative gains). In our concrete function scenario, as we have seen, the spatial gain measured parallel to the line passing through $\boldsymbol{o}$ and $\boldsymbol{c}$ must be at least $d(P, \boldsymbol{c}) = \ell^2/(2 \cdot d(\boldsymbol{o}, \boldsymbol{c}))$ for the spatial gain towards the optimum to be non-negative. If a mutation does yield a positive gain towards the optimum, however, then the respective parallel spatial gain is again an upper bound on the gain towards the optimum. Let G denote the spatial gain towards the optimum, yet this time with respect to our concrete function scenario. Then by the same arguments which have been used in the proof of Lemma 7, we obtain for $n \geq 4$

$$\begin{aligned}\int_{\ell^2/(2\cdot d(\boldsymbol{o},\boldsymbol{c}))}^{\ell} g \cdot F'_{n-2}(g)\,\mathrm{d}g &= \frac{\ell}{\Psi_{n-2}(\pi)} \cdot \left[\frac{1}{n-1}\left(1-y^2\right)^{(n-1)/2}\right]_{\ell/(2\cdot d(\boldsymbol{o},\boldsymbol{c}))}^{1} \\ &< \frac{\ell}{\sqrt{2\pi}\sqrt{n-1}} \cdot \left(1 - \left(\frac{\ell}{2\cdot d(\boldsymbol{o},\boldsymbol{c})}\right)^2\right)^{(n-1)/2}\end{aligned}$$

as an upper bound on $\mathsf{E}[G]$, the expected spatial gain towards the optimum in the concrete scenario. Obviously, in the rightmost expression the quotient on the left increases with ℓ, whereas the power on the right decreases with ℓ (recall that $\ell \leq 2 \cdot d(\boldsymbol{o}, \boldsymbol{c})$). Now we merely need to maximize this upper bound with respect to ℓ to obtain an upper bound on the expected gain towards the optimum that is independent of $|\boldsymbol{m}|$, or in other words, that holds even when $|\boldsymbol{m}|$ is chosen optimally (with respect to the expected gain).

Let $x := \ell/(2 \cdot d(\boldsymbol{o}, \boldsymbol{c}))$ and $n \geq 4$. Then

$$\mathsf{E}[G] < \frac{\ell}{\sqrt{2\pi}\sqrt{n-1}}\left(1-\left(\frac{\ell}{2\cdot d(\boldsymbol{o},\boldsymbol{c})}\right)^2\right)^{(n-1)/2} = \frac{2\cdot d(\boldsymbol{o},\boldsymbol{c})}{\sqrt{2\pi}\sqrt{n-1}} \cdot x \cdot (1-x^2)^{(n-1)/2},$$

and since $0 \leq \ell \leq 2 \cdot d(\boldsymbol{o}, \boldsymbol{c})$, we have $x \in [0, 1]$. Since

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}x}\, x \cdot (1-x^2)^{(n-1)/2} &= (1-x^2)^{(n-1)/2} - x^2 \cdot (n-1) \cdot (1-x^2)^{(n-1)/2-1} \\ &= (1-x^2)^{(n-1)/2-1} \cdot \Big((1-x^2) - x^2 \cdot (n-1)\Big) \\ &= (1-x^2)^{(n-1)/2-1} \cdot (1 - x^2 \cdot n),\end{aligned}$$

$x \cdot (1-x^2)^{(n-1)/2}$ takes its maximum at $1/\sqrt{n}$ for $x \in [0, 1]$, and hence,

$$\mathsf{E}[G] < \frac{2\cdot d(\boldsymbol{o},\boldsymbol{c})}{\sqrt{2\pi}\sqrt{n-1}} \cdot \frac{1}{\sqrt{n}} \cdot \left(1-\frac{1}{n}\right)^{(n-1)/2} < \frac{\sqrt{2}\cdot d(\boldsymbol{o},\boldsymbol{c})}{\sqrt{\pi}\cdot(n-1)} \cdot (3/4)^{3/2}.$$

Note that in the preceding calculations we again optimistically assumed that a positive gain towards the optimum is always accepted, whereas a negative one is always rejected. Consequently, the upper bound obtained is subject to our function scenario, yet it holds irrespective of the selection rule used. Hence, finally using $\sqrt{2/\pi} \cdot (3/4)^{3/2} < 0.52$, we have proved the following result.

Lemma 11. *Let the (1+1) ES minimize a function in the considered function scenario using an arbitrary selection rule. When in a step a suboptimal point $\boldsymbol{c} \in \mathbb{R}^n$ is mutated by adding an isotropic mutation $\boldsymbol{m} \in \mathbb{R}^n$, then – even if $|\boldsymbol{m}|$ were chosen optimally – for $n \geq 4$ the expected spatial gain towards the optimum $\boldsymbol{o} \in \mathbb{R}^n$ in this step is smaller than $0.52 \cdot d(\boldsymbol{o}, \boldsymbol{c})/(n-1)$, i.e., it is $O(d(\boldsymbol{o}, \boldsymbol{c})/n)$.*

This result shows that if the 1/5-rule is able to adapt the length of the mutation vector such that the mutant is accepted with a probability close to 0.2, and elitist selection is used, then the expected spatial gain towards the optimum, i.e. the expected reduction of the approximation error, has indeed optimal order (with respect to our scenario; cf. Lemma 10).

4. Analyzing the runtime: Lower bound

Since the upper bound on the expected spatial gain towards the optimum in a single step, which we have just proved, holds even when the length of the mutation vectors is adapted optimally, we can easily proof a lower bound on the expected runtime of the (1+1) ES when it minimizes a function in our function scenario, for instance the well-known SPHERE function. Therfore, we simply assume that in each step the length of the mutation vector is chosen such that the expected spatial gain in this step is maximal. It is easy to see that this "greedy" assumption in fact maximizes the expected total gain of a sequence of steps. For our lower bound, however, we need not prove this explicitly; rather we are going to apply the following plausible fact about the sum of a sequence of random variables.

Lemma 12. *Let $X_1, X_2, \ldots$ denote random variables with bounded range and S the random variable defined by $S = \min\{s \mid X_1 + \cdots + X_s \geq g\}$ for a given $g > 0$. If $\mathsf{E}[S]$ exists and $\mathsf{E}[X_i \mid S \geq i] \leq u$ for $i \in \mathbb{N}$, then $\mathsf{E}[S] \geq g/u$.*

The proof of this lemma, which essentially follows the one of Wald's equation, can be found in Appendix B since it could not be found in the literature (although it is surely there).

The application of this lemma is obvious: S denotes the number of steps that the (1+1) ES needs to ensure a certain approximation error; namely, g denotes the reduction of the approximation error we are interested in. For instance, when the optimization is started at $\boldsymbol{a} \in \mathbb{R}^n \setminus \{\boldsymbol{o}\}$, then choosing $g := d(\boldsymbol{o}, \boldsymbol{a})/2$ yields a bound on the expected number of steps to halve the approximation error. Finally, X_i denotes the spatial gain towards the optimum in the ith step of the (1+1) ES, and the upper bound on $\mathsf{E}[X_i \mid S \geq i]$, namely u, has just been proved in the previous section. Note that in our scenario the condition "$S \geq i$" corresponds to the event that the aspired reduction of the approximation error has not already been realized within the steps $1, \ldots, i-1$.

Theorem 13. *Let the (1+1) ES minimize a function $f : \mathbb{R}^n \to \mathbb{R}$ in the considered function scenario using isotropic mutations and elitist selection. Furthermore, let S denote the number of steps, i.e. the number of f-evaluations, to halve the approximation error. Then irrespective of how the mutations are adapted*

(1) $\mathsf{E}[S] > 0.96 \cdot (n-1)$, *i.e.,* $\mathsf{E}[S] = \Omega(n)$,
(2) *for any constant $\varepsilon > 0$, w. o. p. $S = \Omega(n^{1-\varepsilon})$.*

Proof. If $\mathsf{E}[S]$ is not defined (due to improper mutation adaptation), one may informally reason that "$\mathsf{E}[S] = \infty = \Omega(n)$" since S is positive.

Hence, we assume that $\mathsf{E}[S]$ exists. Due to elitist selection and our function scenario, $X_i \geq 0$, and consequently, the initial approximation error, $d(\boldsymbol{o}, \boldsymbol{a})$, is an upper bound on the approximation error in each step. Hence, $X_i \leq d(\boldsymbol{o}, \boldsymbol{a})$. In other words, the X_i are bounded, respectively. Furthermore, due to Lemma 11, we have $\mathsf{E}[X_i \mid S \geq i] < 0.52 \cdot d(\boldsymbol{o}, \boldsymbol{a})/(n-1) =: u$, and with this, Lemma 12 finally yields

$$\mathsf{E}[S] \geq \frac{g}{u} = \frac{d(\boldsymbol{o}, \boldsymbol{a})/2}{0.52 \cdot d(\boldsymbol{o}, \boldsymbol{a})/(n-1)} > 0.96 \cdot (n-1).$$

For the proof of the second claim, assume that for some constant $\varepsilon > 0$, $O(n^{1-\varepsilon})$ steps suffice to halve the initial approximation error of $d(\boldsymbol{o}, \boldsymbol{a})$. Then at least in one step the spatial gain would have to be $\Omega(d(\boldsymbol{o}, \boldsymbol{a})/n^{1-\varepsilon}) = d(\boldsymbol{o}, \boldsymbol{a}) \cdot \Omega(n^{\varepsilon})/n$. Since a mutation yields a spatial gain of $|\boldsymbol{m}| \cdot \Omega(n^{\varepsilon/2})/\sqrt{n}$ only with exponentially small probability (cf. Corollary 6), in this step $|\boldsymbol{m}|$ would have to be larger than

$$\frac{d(\boldsymbol{o}, \boldsymbol{a}) \cdot \Omega(n^{\varepsilon})}{n} \Bigg/ \frac{\Omega(n^{\varepsilon/2})}{\sqrt{n}} = d(\boldsymbol{o}, \boldsymbol{a}) \cdot \Omega(n^{\varepsilon/2-1/2})$$

for the probability of the necessary spatial gain not to be exponentially small per se. The success probability of such a step, however, is exponentially small according to Lemma 8 (3). Hence, whatever the length of $\boldsymbol{m}$ may be, a step yields a spatial gain of $\Omega(d(\boldsymbol{o}, \boldsymbol{a})/n^{1-\varepsilon})$ only with exponentially small probability. Finally, the probability that in at least one of $O(n^{1-\varepsilon})$ steps such a gain is realized is still exponentially small. □

Our choice to investigate the runtime that is necessary to halve the approximation error was somehow arbitrary as we could have chosen a different g for the application of Lemma 12. However, after the approximation error is halved, the theorem can be applied anew, and hence, we obtain the main lower-bound result.

Theorem 14. *Let the* (1+1) *ES minimize a function $f : \mathbb{R}^n \to \mathbb{R}$ in the considered function scenario using isotropic mutations and elitist selection. Then the expected runtime, i.e. the expected number of f-evaluations, to reduce the initial approximation error to an α-fraction is $\Omega(\log(1/\alpha) \cdot n)$ for $0 < \alpha \leq 1/2$ (where α may decrease in n).*

As we assume that in each single step the mutation vector is adapted optimally, the lower bounds on the (expected) runtime we have proved in this section hold independently of how the mutation adaptation chooses the distribution of

$|\boldsymbol{m}|$ in a step. Consequently, they particularly hold when Gaussian mutations are used and the 1/5-rule is applied for their adaptation. The upper bound on the runtime, which we are going to prove in the next section, is subject to this scenario, which has already been described in the introduction.

5. Analyzing the runtime: Upper bound

In the calculations above we assumed that the isotropic mutation vector $\boldsymbol{m}$ would take the length ℓ. For the upper bounds on the (expected) spatial gain of a step, we additionally assumed that a positive gain is always accepted, whereas a negative one is always rejected, and that ℓ would be the optimal choice for the length of the mutation vector. In other words, we assumed that in each step the mutation adaptation magically chooses a distribution for $|\boldsymbol{m}|$ that is concentrated on an optimal value, respectively.

Obviously, for an upper bound on the runtime we must take into account that Gaussian mutations are used and that their adaptation is managed by the 1/5-rule. In addition to isotropy, the following properties of Gaussian mutations will turn out useful for an analysis.

Lemma 15. *For a scaled Gaussian mutation $\boldsymbol{m}$ over $\mathbb{R}^n$, the expectation $\ell_{\mathsf{E}} := \mathsf{E}[|\boldsymbol{m}|]$ exists, and moreover, $\mathsf{P}\{\big||\boldsymbol{m}| - \ell_{\mathsf{E}}\big| \geq \delta \cdot \ell_{\mathsf{E}}\} \leq \delta^{-2}/(2n-1)$.*

Let $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_n$ denote independent copies of $\boldsymbol{m}$, respectively. Then for any constant $\kappa < 1$, two positive constants a_κ and b_κ exist such that for the cardinality of $I := \{i \mid a_\kappa \cdot \ell_{\mathsf{E}} \leq |\boldsymbol{m}_i| \leq b_\kappa \cdot \ell_{\mathsf{E}}\}$, w. o. p. $\#I \geq \kappa \cdot n$.

Proof. Recall that $\boldsymbol{m} = s \cdot \widetilde{\boldsymbol{m}}$, where $s \in \mathbb{R}_{>0}$ and each component of $\widetilde{\boldsymbol{m}} \in \mathbb{R}^n$ is independently standard normal distributed. Then the random variable $|\widetilde{\boldsymbol{m}}|$ is χ-distributed (with n degrees of freedom; cf. [1]), and hence,

$$\mathsf{E}[|\widetilde{\boldsymbol{m}}|] \;=\; \sqrt{2} \cdot \frac{\Gamma(n/2+1/2)}{\Gamma(n/2)} \;=\; \Theta(\sqrt{n}).$$

Since $|\widetilde{\boldsymbol{m}}|^2$ is χ^2-distributed, $\mathsf{E}\big[|\widetilde{\boldsymbol{m}}|^2\big] = n$, and consequently,

$$\mathsf{Var}[|\widetilde{\boldsymbol{m}}|] \;=\; \mathsf{E}\Big[|\widetilde{\boldsymbol{m}}|^2\Big] - \mathsf{E}[|\widetilde{\boldsymbol{m}}|]^2 \;=\; n - 2 \cdot \left(\frac{\Gamma(n/2+1/2)}{\Gamma(n/2)}\right)^2.$$

Furthermore, $\mathsf{Var}[|\widetilde{\boldsymbol{m}}|]$ tends to $1/2$ from below as $n \to \infty$, i.e., $\mathsf{Var}[|\widetilde{\boldsymbol{m}}|] \leq 1/2$, and consequently, $\mathsf{E}[|\widetilde{\boldsymbol{m}}|]^2 \geq n - 1/2$.

If for a random variable Y, $\mathsf{E}\big[Y^2\big]$ exists and $\mathsf{E}[Y] > 0$, then Chebyshev's inequality yields that for any $\delta > 0$:

$$\mathsf{P}\big\{\big|Y - \mathsf{E}[Y]\big| \geq \delta \cdot \mathsf{E}[Y]\big\} \leq \frac{\mathsf{Var}[Y]}{(\delta \cdot \mathsf{E}[Y])^2}.$$

Since $\mathsf{E}[|\boldsymbol{m}|] = s \cdot \mathsf{E}[|\widetilde{\boldsymbol{m}}|]$ and $\mathsf{Var}[|\boldsymbol{m}|] = s^2 \cdot \mathsf{Var}[|\widetilde{\boldsymbol{m}}|]$, applying this bound to the random variable $|\boldsymbol{m}|$ yields

$$\mathsf{P}\big\{\big||\boldsymbol{m}| - \ell_{\mathsf{E}}\big| \geq \delta \cdot \ell_{\mathsf{E}}\big\} \leq \frac{s^2 \cdot 1/2}{\big(\delta \cdot s \cdot \mathsf{E}[|\widetilde{\boldsymbol{m}}|]\big)^2} \leq \frac{1}{\delta^2 \cdot (2n-1)}.$$

Finally, we consider n i. i. d. scaled Gaussian mutations. Since $|\boldsymbol{m}| = \Theta(\mathsf{E}[|\boldsymbol{m}|])$ with probability $1 - O(1/n)$ as we have just seen, $\mathsf{E}[\#I] = n - O(1)$ and applying Chernoff bounds completes the proof. □

Recall that the 1/5-rule (as defined in the introduction) reads: The scaling factor s is adapted after every nth step; it is halved if less than $n/5$ of the respective last n steps have been successful, and otherwise doubled. The asymptotic calculations we are going to present, however, are valid for any 1/5-rule keeping s unchanged for $\Theta(n)$ steps, respectively, and using any two positive constants (one greater and one smaller than 1) for the scaling of s.

A run of the (1+1) ES is virtually partitioned into phases each of which lasts n steps such that in each phase $\mathsf{E}[|\boldsymbol{m}|]$ is constant. Let s_i denote the scaling factor used throughout the ith phase and ℓ_i the corresponding $\mathsf{E}[|\boldsymbol{m}|]$. A phase after which s is doubled is symbolized by "×" and a phase after which s is halved by "÷". Furthermore, let d_i denote the distance from the optimum, i.e. the approximation error, at the beginning of the ith phase; hence, $d_i - d_{i+1}$ equals the spatial gain realized in the ith phase, i.e. the change of the approximation error in this phase.

The next result tells us that if the scaling factor, and with it $\mathsf{E}[|\boldsymbol{m}|]$, has the right order at the beginning of a phase, then this phase reduces the approximation error w. o. p. by a constant fraction. Furthermore, we will see that if the

1/5-rule decides to double the scaling factor after a phase, it is very unlikely that in this phase the scaling factor (and with it the expected length of the mutation vectors) has been too large. If the scaling factor is halved after a phase, however, it is very unlikely that the expected length of the mutation vector is too small in the following phase.

Lemma 16. *Let the (1+1) ES minimize a function in the considered scenario using Gaussian mutations adapted by the 1/5-rule and elitist selection. Then*

(1) *if $\ell_i = \Theta(d_i/\sqrt{n})$, then w. o. p. $d_{i+1} = d_i - \Omega(d_i)$, i.e., w. o. p. the approximation error is reduced by a constant fraction in the ith phase;*
(2) *if s is doubled after the ith phase, then $\ell_i = O(d_i/\sqrt{n})$ w. o. p.;*
(3) *if s is halved after the ith phase, then $\ell_{i+1} = \Omega(d_{i+1}/\sqrt{n})$ w. o. p.*

Proof. Assume that the total spatial gain of the ith phase is not $\Omega(d_i)$. Then the distance from the optimum is $\Theta(d_i)$ in each step of the phase (recall that the distance is non-increasing). Lemma 15 yields that w. o. p. $|\boldsymbol{m}| = \Theta(d_i/\sqrt{n})$ in $0.9n$ steps, and according to Lemma 9, in each such step the spatial gain is $\Omega(d_i/n)$ with probability $\Omega(1)$. Hence, the expected number of steps in each of which the distance is reduced by $\Omega(d_i/n)$ is $\Omega(n)$. By Chernoff bounds, the number of such steps is $\Omega(n)$ w. o. p. Since negative gains are rejected in the considered scenario, our initial assumption contradictorily implies that the total spatial gain of the ith phase is $\Omega(d_i)$ w. o. p.

For the proof of the second claim, assume that ℓ_i is not $O(d_i/\sqrt{n})$. Since the distance from the optimum is non-increasing, ℓ_i is not $O(d(\boldsymbol{o}, \boldsymbol{c})/\sqrt{n})$ in each step of the ith phase. Lemma 15 yields that $|\boldsymbol{m}|$ is not $O(d(\boldsymbol{o}, \boldsymbol{c})/\sqrt{n})$ in $0.9n$ steps w. o. p. According to Lemma 8, 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 unsuccessful; in other words, w. o. p. less than 1/5 of the steps in the phase are successful. Thus, the assumption "ℓ_i is not $O(d_i/\sqrt{n})$" contradictorily implies that s is halved w. o. p., and not doubled.

Assume ℓ_{i+1} is not $\Omega(d_{i+1}/\sqrt{n})$ for the proof of the third claim. Since $s_i = 2s_{i+1}$, also $\ell_i = 2\ell_{i+1}$. As the distance from the optimum is non-increasing in our scenario, the assumption implies that ℓ_i is not $\Omega(d(\boldsymbol{o}, \boldsymbol{c})/\sqrt{n})$ in each step of the ith phase. Following the proof of the second claim with symmetric arguments, in the ith phase w. o. p. more than $0.8n$ steps are successful — contradictorily implying that s is doubled after the ith phase w. o. p., and not halved. □

The preceding lemma tells us that the 1/5-rule does what it is supposed to do, and we can deal with sequences of phases in a run of the (1+1) ES now, making an algorithmic analysis of the (1+1) ES possible.

Lemma 17. *Let the (1+1) ES minimize a function in the considered scenario using Gaussian mutations adapted by the 1/5-rule and elitist selection. If the 1/5-rule causes a sequence $\div\times^k$ of phases, $k = n^{O(1)}$, then w. o. p. the distance from the optimum is k times reduced by a constant fraction in these phases.*

Proof. Let the $\div$-phase be the ith one. By Lemma 16, $\ell_{i+1} = \Omega(d_{i+1}/\sqrt{n})$ w. o. p. Since the adaptation yields $\ell_{i+w} \geq \ell_{i+1}$ for $1 \leq w \leq k$ and the distance from the optimum is non-increasing, w. o. p. $\ell_{i+w} = \Omega(d_{i+w}/\sqrt{n})$ for $1 \leq w \leq k$. Lemma 16 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 16, in each of the k $\times$-phases the distance is reduced by a constant fraction w. o. p. □

The preceding proof has been easy since increasing the scaling factor surely decreases the success probability in subsequent steps (as the distance from the optimum cannot increase). Decreasing the scaling factor, however, need not necessarily result in a larger success probability. This depends on the decrease of the distance from the optimum, i.e., on how fast the approximation error is reduced. Thus, the symmetric situation is more complicated.

Lemma 18. *Let the (1+1) ES minimize a function $f\colon \mathbb{R}^n \to \mathbb{R}$ in the considered scenario using Gaussian mutations adapted by the 1/5-rule and elitist selection. If the 1/5-rule causes a sequence $\times\div^k$ of phases, $k = n^{O(1)}$, then w. o. p. the distance from the optimum is k times reduced by a constant fraction in these phases.*

Proof. Let the $\times$-phase be the ith one. For $k = 1$ assume that the total spatial gain of the ith and the $(i+1)$-th phase is not $\Omega(d_i)$. According to Lemma 16, 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 $\ell_{i+1} = \Theta(d_{i+1}/\sqrt{n})$ (recall that the distance from the optimum is non-increasing), and Lemma 16 contradictorily implies that in each of the two phases the distance is reduced by a constant fraction w. o. p. Consequently, w. o. p. the two phases result in $d_{i+2} = d_i - \Omega(d_i)$.

For $k \ge 2$, the adaptation yields $s_{i+w} = s_i\, 2^{2-w} = 4 \cdot s_i/2^w$ for $1 \le w \le k$, and according to Lemma 16, for $2 \le w \le k$ w. o. p. $\ell_{i+w} = \Omega(d_{i+w}/\sqrt{n})$. If $d_{i+w} \le d_i/2^w$, then a simple accounting argument yields that after the $(i+w)$th phase $d_{i+w+1} \le d_{i+w} \le d_i/2^w \le d_i/\kappa^{w+1}$ for a constant $\kappa \ge \sqrt{2}$ so that we were done. Therefore, assume $d_{i+w} > d_i/2^w$. Since $\ell_{i+w} = 4\,\ell_i/2^w$, 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 ℓ_{i+w} is also $\Omega(d_{i+w}/\sqrt{n})$, again Lemma 16 yields that the $(i+w)$th phase reduces the distance by a constant fraction w. o. p.

All in all, after the first two phases w. o. p. $d_{i+2} = d_i - \Omega(d_i)$, and for $2 \le w \le k$, 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 $d_{i+w+1} \le d_i/\kappa^{w+1} \le 2^{(w+1)/2}$ even if the $(j+w)$th phase did not yield any spatial gain at all. □

Finally, for our scenario the three preceding lemmas together enable a matching upper bound on the runtime, the number of steps the (1+1) ES needs to realize a predefined reduction of the approximation error, which is given by the distance from the optimum in the search space.

Theorem 19. *Let the (1+1) ES minimize a function $f : \mathbb{R}^n \to \mathbb{R}$ in the considered function scenario using Gaussian mutations adapted by the 1/5-rule and elitist selection. Given that the initialization ensures $d_1/s_1 = \Theta(n)$, the runtime/number of f-evaluations until the approximation error is no more than $\alpha \cdot d_1$ is w. o. p. $O(\log(1/\alpha) \cdot n)$, where $1/2 \ge \alpha = \exp(-n^{O(1)})$.*

Proof. First note that $\log(1/\alpha) = n^{O(1)}$ due to the lower bound on α.

When the sequence of phases starts with $\times\div$ or with $\div\times$, then the two preceding lemmas yield that the number of phases such that $d(\boldsymbol{o}, \boldsymbol{c}) \le \alpha \cdot d_1$ is w. o. p. $O(\log(1/\alpha))$ because w. o. p. each of the phases reduces the approximation error by a constant fraction (so that a polynomial number of phases suffice; by summing up a polynomial number of "error probabilities" each of which is exponentially small, we get around the problem of dependencies and still end up with a total error probability which is exponentially small).

If the sequence starts with $\times^k$ or with $\div^k$ for $k \ge 2$, however, it remains to show that in these phases the distance is w. o. p. reduced k times by a constant fraction. Here the assumption on the initialization of the scaling factor comes into play; namely, the assumption $d_1/s_1 = \Theta(n)$ ensures that in the first phase $\ell_1 = \mathsf{E}[|\boldsymbol{m}|] = \Theta(\sqrt{n}) \cdot s_1 = \Theta(d_1/\sqrt{n})$ (cf. Lemma 15) and the same argumentation as for $\div\times^k$ (resp. $\times\div^k$) applies (yet without the preceding $\div$-phase (resp. $\times$-phase)). Hence, also in such situations the number of phases until the current search point $\boldsymbol{c}$ is such that $d(\boldsymbol{o}, \boldsymbol{c}) \le \alpha \cdot d_1$ is $O(\log(1/\alpha))$ w. o. p. □

For other starting conditions, the number of steps, j, until the theorem's assumption is met, namely $d_{j-1}/s_{j-1} = \Theta(n)$, 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 utilizing the strong results presented in Lemma 16.

6. Conclusion

For the first time, the (expected) runtime of a simple, but fundamental evolutionary algorithm for continuous optimization in $\mathbb{R}^n$ is rigorously analyzed — rather than a simplifying model of it. In particular, this algorithmic analysis shows that for the considered function scenario Gaussian mutations adapted by the well-known 1/5-rule indeed results in asymptotically optimal runtime. Since the analysis covers a wide range of realizations of the 1/5-rule, it additionally yields an interesting by-product: Fine tuning the parameters of the 1/5-rule actually does not affect the order of the runtime. This result can be interpreted as an indicator for the robustness of the (1+1) ES; yet it is proved only for the investigated function scenario. Moreover, the results show that, in our scenario, even choosing a different type of isotropic mutation and/or a different adaptation rule cannot significantly decrease the order of the runtime.

Besides the concrete results on the (expected) runtime in the considered scenario, the estimates and, especially, the methods developed for their proofs, obtained for isotropically distributed mutation vectors can be reused in forthcoming analyses. They can be used to analyze the (1+1) ES for other function scenarios; yet they might also serve as a tool in the analyses of more complicated ES which use isotropic mutations, for instance, to mutate an individual obtained by recombination. That (in the considered function scenario) the upper bound and the lower bound on the runtime meet shows the strength of the estimates and of the methods presented.

For further reading

[8]

Acknowledgments

Thanks for productive discussions and for pointing out flaws especially go to Ingo Wegener, Carsten Witt, and Stefan Droste. The author was supported by the German Research Foundation (DFG) as part of the collaborative research center "Computational Intelligence" (SFB 531).

Appendix A. Bounding $\Psi_k(\pi) = \int_0^\pi (\sin x)^k \mathrm{d}x$

By the definition of the beta function (cf. [1]), namely

$$B(m+1, n+1) = 2 \cdot \int_0^{\pi/2} (\cos x)^{2m+1} (\sin x)^{2n+1} \, \mathrm{d}x \, ,$$

for $k \in \mathbb{N}$

$$\int_0^{\pi/2} (\sin x)^k \, \mathrm{d}x = \frac{1}{2} \cdot B\left(\frac{1}{2}, \frac{k+1}{2}\right).$$

As $B(m,n) = \Gamma(m) \cdot \Gamma(n) / \Gamma(m+n)$ and $\Gamma(1/2) = \sqrt{\pi}$,

$$\int_0^{\pi} (\sin x)^k \, \mathrm{d}x \;=\; 2 \cdot \int_0^{\pi/2} (\sin x)^k \, \mathrm{d}x \;=\; B\left(\frac{1}{2}, \frac{k+1}{2}\right) = \sqrt{\pi} \cdot \frac{\Gamma\left(\frac{k}{2} + \frac{1}{2}\right)}{\Gamma\left(\frac{k}{2} + 1\right)}.$$

Using

$$\frac{\Gamma(n+1/2)}{\Gamma(n)} = \sqrt{n} \left(1 - \frac{1}{2^3 n} + \frac{1}{2^7 n^2} + \frac{5}{2^{10} n^3} - \frac{21}{2^{15} n^4} + O(n^{-5})\right)$$

([7], p. 576), we obtain for $k \geq 2$

$$\sqrt{\frac{2}{k+1}} < \frac{\Gamma\left(\frac{k}{2} + \frac{1}{2}\right)}{\Gamma\left(\frac{k}{2} + 1\right)} < \sqrt{\frac{2}{k}}.$$

Altogether, for $k \geq 2$

$$\sqrt{\frac{2\pi}{k+1}} < \int_0^{\pi} (\sin x)^k \, \mathrm{d}x \;=\; \Psi_k(\pi) \;<\; \sqrt{\frac{2\pi}{k}},$$

and consequently, $\int_0^\pi (\sin x)^k \, \mathrm{d}x = \Theta(1/\sqrt{k})$.

Appendix B. Proof of Lemma 12

Claim: "Let $X_1, X_2, \ldots$ denote random variables with bounded range and S the random variable defined by $S = \min\{s \mid X_1 + \cdots + X_s \geq g\}$ for a given $g > 0$. If $\mathsf{E}[S]$ exists and $\mathsf{E}[X_i \mid S \geq i] \leq u$ for $i \in \mathbb{N}$, then $\mathsf{E}[S] \geq g/u$."

Obviously $S \geq 1$, and the condition $S \geq i \geq 2$ is equivalent to $X_1 + \cdots + X_k < g$ for $1 \leq k < i$. Since the X_i are bounded, $\mathsf{E}[X_1 + \cdots + X_S]$ also exists if $\mathsf{E}[S]$ exists. The proof follows the one of Wald's equation (up to the point where the upper bound on $\mathsf{E}[X_i \mid S \geq i]$ is utilized rather than the original assumption that the X_i are i. i. d.).

$$\begin{aligned} g \;&\leq \mathsf{E}[X_1 + \cdots + X_S] \\ &= \sum_{t=1}^{\infty} \mathsf{P}\{S = t\} \cdot \mathsf{E}[X_1 + \cdots + X_t \mid S = t] \end{aligned}$$

$$= \sum_{t=1}^{\infty} \mathsf{P}\{S = t\} \cdot \sum_{i=1}^{t} \mathsf{E}[X_i \mid S = t]$$

$$= \sum_{t=1}^{\infty} \sum_{i=1}^{t} \mathsf{P}\{S = t\} \cdot \mathsf{E}[X_i \mid S = t]$$

since the series converges absolutely due to the boundedness of the X_i

$$= \sum_{i=1}^{\infty} \sum_{t=i}^{\infty} \mathsf{P}\{S = t\} \cdot \mathsf{E}[X_i \mid S = t]$$

$$= \sum_{i=1}^{\infty} \sum_{t=i}^{\infty} \mathsf{P}\{S = t \mid S \geq i\} \cdot \mathsf{P}\{S \geq i\} \cdot \mathsf{E}[X_i \mid S = t]$$

$$= \sum_{i=1}^{\infty} \mathsf{P}\{S \geq i\} \cdot \sum_{t=i}^{\infty} \mathsf{P}\{S = t \mid S \geq i\} \cdot \mathsf{E}[X_i \mid S = t]$$

since $t \geq i,\ S = t$ implies $S \geq i$

$$= \sum_{i=1}^{\infty} \mathsf{P}\{S \geq i\} \cdot \sum_{t=i}^{\infty} \mathsf{P}\{S = t \mid S \geq i\} \cdot \mathsf{E}[X_i \mid S = t \wedge S \geq i]$$

since $t < i$ implies $\mathsf{P}\{S = t \mid S \geq i\} = 0$

$$= \sum_{i=1}^{\infty} \mathsf{P}\{S \geq i\} \cdot \sum_{t=1}^{\infty} \mathsf{P}\{S = t \mid S \geq i\} \cdot \mathsf{E}[X_i \mid S = t \wedge S \geq i]$$

$$= \sum_{i=1}^{\infty} \mathsf{P}\{S \geq i\} \cdot \mathsf{E}[X_i \mid S \geq i]$$

$$\leq \sum_{i=1}^{\infty} \mathsf{P}\{S \geq i\} \cdot u$$

$$= \mathsf{E}[S] \cdot u \quad \square$$

References

[1] G.B. Arfken, Mathematical Methods for Physicists, 3rd ed., Academic Press, San Diego, 1990.

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

[3] A. Bienvenue, O. Francois, Global convergence for evolution strategies in spherical problems: Some simple proofs and difficulties, Theoretical Computer Science 306 (2003) 269–289.

[4] S. Droste, T. Jansen, K. Tinnefeld, I. Wegener, A new framework for the valuation of algorithms for black-box optimization, in: Foundations of Genetic Algorithms 7, FOGA 2002, Morgan Kaufmann, San Francisco, 2003, pp. 253–270.

[5] S. Droste, T. Jansen, I. Wegener, On the analysis of the (1+1) evolutionary algorithm, Theoretical Computer Science 276 (2002) 51–82.

[6] O. Giel, I. Wegener, Evolutionary algorithms and the maximum matching problem, in: Proc. of the 20th Int'l Symposium on Theoretical Aspects of Computer Science, STACS, in: LNCS, vol. 2607, Springer, 2003, pp. 415–426.

[7] R.L. Graham, D.E. Knuth, O. Patashnik, Concrete Mathematics: A Foundation for Computer Science, Addison-Wesley, 1989.

[8] J. Jägersküpper, Analysis of a simple evolutionary algorithm for minimization in Euclidean spaces, in: Proceedings of the 30th Int'l Colloquium on Automata, Languages and Programming, ICALP, in: LNCS, vol. 2719, Springer, 2003, pp. 1068–1079.

[9] M.G. Kendall, A Course in the Geometry of *n* Dimensions, Charles Griffin & Co. Ltd., London, 1961.

[10] H. Mühlenbein, How genetic algorithms really work: Mutation and hillclimbing, in: Parallel Problem Solving from Nature 2, PPSN, North-Holland, Amsterdam, 1992, pp. 15–25.

[11] A.S. Nemirovsky, D.B. Yudin, Problem Complexity and Method Efficiency in Optimization, Wiley, New York, 1983.

[12] I. Rechenberg, Evolutionsstrategie, Frommann-Holzboog, Stuttgart, Germany, 1973.

[13] I. Rechenberg, Evolutionsstrategie '94. Frommann-Holzboog, Stuttgart, Germany, 1994.

[14] G. Rudolph, Convergence Properties of Evolutionary Algorithms, Verlag Dr. Kovač, Hamburg, 1997.

[15] J. Scharnow, K. Tinnefeld, I. Wegener, Fitness landscapes based on sorting and shortest paths problems, in: Parallel Problem Solving from Nature 7, PPSN, in: LNCS, vol. 2439, Springer, 2002, pp. 54–63.

[16] H.-P. Schwefel, Evolution and Optimum Seeking, Wiley, New York, 1995.

[17] I. Wegener, Theoretical aspects of evolutionary algorithms, in: Proceedings of the 28th Int'l Colloquium on Automata, Languages and Programming, ICALP, in: LNCS, vol. 2076, Springer, 2001, pp. 64–78.

[18] I. Wegener, Towards a theory of randomized search heuristics, in: Proceedings of the 28th Int'l Symposium on Mathematical Foundations of Computer Science, MFCS, in: LNCS, vol. 2747, Springer, 2003, pp. 125–141.