UNIVERSITY OF DORTMUND

REIHE COMPUTATIONAL INTELLIGENCE

COLLABORATIVE RESEARCH CENTER 531

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

Analysis of a Simple Evolutionary Algorithm for the Minimization in Euclidian Spaces

Jens Jägersküpper

No. CI-140/02

Technical Report ISSN 1433-3325 October 2002

Secretary of the SFB 531 · University of Dortmund · Dept. of Computer Science/XI
44221 Dortmund · Germany

This work is a product of the Collaborative Research Center 531, "Computational Intelligence", at the University of Dortmund and was printed with financial support of the Deutsche Forschungsgemeinschaft.

Analysis of a Simple Evolutionary Algorithm for the Minimization in Euclidian Spaces

Jens Jägersküpper

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 expected runtimes and success probabilities are limited to discrete search spaces. In practice, however, EAs are mostly used for continuous optimization problems.
First results on the expected runtime of a simple, but fundamental EA minimizing a symmetric polynomial of degree two in $\mathbb{R}^n$ are presented. Namely, the so-called (1+1) evolution strategy ((1+1) ES) minimizing the Sphere function is investigated. A lower bound on the expected runtime is shown that is valid for any mutation adaptation using isotropically distributed mutation vectors. Furthermore, a matching upper bound on the expected runtime is proved when the well-known 1/5-rule is used to adapt the length of Gauss-mutation vectors. Consequently, the 1/5-rule in combination with Gauss-mutations indeed result in the (1+1) ES having asymptotically optimal expected runtime on Sphere.

1 Introduction

The optimization (here: minimization) of functions $f\colon S \to \mathbb{R}$ for some so-called search space S is one of the fundamental algorithmic problems. For discrete search spaces, like $\{0,1\}^n$, we get the many problems of combinatorial optimization like TSP, graph coloring, or vertex cover. For continuous search spaces, $\mathbb{R}^n$ for instance, we get the problems of mathematical optimization (often with constraints); here problems are defined by classes of functions (polynomials of degree d, k-times differentiable functions, etc.). Many problem-specific algorithms have been designed for each of the two scenarios. Furthermore, there is a theory on algorithms since algorithms can be analyzed and compared.

If not enough resources are on-hand to design a problem-specific algorithm or the knowledge about the problem instance is not sufficient, however, robust algorithms like randomized search heuristics are often a good alternative. Although frequently applied in practice by now, such heuristics have produced only little interest of theoreticians. The aim of this paper is to start the analysis of evolutionary algorithms on continuous search spaces with the tools and the purpose of classical algorithm theory.

We restrict ourselves to the simple (1+1) evolution strategy which—despite its simpleness—has been applied with surprising success. The notion "evolution strategy" is due to Rechenberg whose 1973 book *Evolutionsstrategie* can be seen as one of the starting points of evolutionary computation. The rough structure of the (1+1) ES minimizing the function $f\colon \mathbb{R}^n \to \mathbb{R}$ is given by the following infinite loop:

Initialization
Set the current search point $\boldsymbol{c} \in \mathbb{R}^n$ and the scaling factor $s \in \mathbb{R}$ to the given start values.

Evolution loop

1. Choose $\widetilde{\boldsymbol{m}} \in \mathbb{R}^n$ according to the given mutation distribution.
2. Set the (scaled) mutation vector: $\boldsymbol{m} := s \cdot \widetilde{\boldsymbol{m}} \in \mathbb{R}^n$
3. Generate the mutant: $\boldsymbol{x} := \boldsymbol{c} + \boldsymbol{m} \in \mathbb{R}^n$
4. The given *selection rule* determines (by f) whether $\boldsymbol{c} := \boldsymbol{x}$ or not.

A single execution of the loop is called a *step* of the (1+1) ES. Furthermore, *the mutant/the mutation is accepted* iff the mutant $\boldsymbol{x}$ is selected (to become the current search point $\boldsymbol{c}$). In this case, the step is called a *success*. The adaptation of the scaling factor s is controlled from outside the loop. Originally, the adaptation of the mutation vector's length is based on the (relative) frequency of successful steps (in a certain number of successive steps). That is, during the optimization the value of s is changed according to a given *adaptation rule* based on this frequency. Naturally, in applications a stopping criterion is needed to ensure termination, yet this is not the crucial aspect in our analysis. We investigate the (1+1) ES as an infinite process and are interested in random variables describing properties of the search process: Let X_f denote the first point in time, measured in the number of steps, when some good event happens, for instance the first point in time when $f(\boldsymbol{c}) \leq b$ for a fixed bound b. Then we are interested in $\mathsf{E}[X_f]$—if it exists—and in $\mathsf{P}\{X_f \leq t\}$, the probability that the good event happens within the first t steps. In particular, we are interested in asymptotic results with respect to the dimension of the search space $\mathbb{R}^n$.

The potentially most discussed function in the field of EAs for the search space $\mathbb{R}^n$ concerning their analysis is the simple, but fundamental Sphere function defined by $\text{Sphere}(\boldsymbol{x}) := x_1^2 + \cdots + x_n^2$ for $\boldsymbol{x} = (x_1, \ldots, x_n) \in \mathbb{R}^n$. Obviously, Sphere equals the L_2-norm squared, in other words, the square of the Euclidian distance from the origin, the minimum of Sphere. This function has a discrete counter part: OneMax: $\{0,1\}^n \to \mathbb{N}$ defined by $\text{OneMax}(\boldsymbol{a}) := a_1 + \cdots + a_n$. If OneMax is maximized, the number of ones is maximized, or equivalently, the Hamming distance from the optimum $(1, \ldots, 1) \in \{0,1\}^n$ is minimized. The discrete counter part of the (1+1) ES, the so-called (1+1) evolutionary algorithm ((1+1) EA), was thoroughly analyzed on OneMax (see Mühlenbein (1992) for an $O(n \log n)$ bound on the expected runtime). These results are easy to obtain, yet more sophisticated papers on the (1+1) EA have been published: The (1+1) EA has been investigated on linear functions (Droste, Jansen, and Wegener (2002)), on quadratic polynomials (Wegener and Witt (2002)) and on monotone polynomials (Wegener (2001)). Even the effect of recombination has been analyzed for the search space $\{0,1\}^n$ (Jansen and Wegener (2002, 2001)), and the number of papers focusing on algorithmic analyses is increasing.

The situation for continuous search spaces is different: Here, the vast majority of results on EAs are of empirical nature. Only very few papers focusing on algorithmic analyses have been published after the initial results from Rechenberg who emphasizes the importance of Sphere by stating that the minimization of Sphere models the minimization close to a minimum for many other functions (cf. his 1994 book *Evolutionsstrategie '94*). Extensive and detailed investigations are due to Beyer and can be found in his 2001 book *The theory of evolution strategies*. From an algorithmic point of view, however, these results are not exhaustive: Either only a model of the stochastic process is investigated such that the results have to be verified by experiments/simulations (as typically proceeded in physics) or simplifying assumptions in calculations are made without controlling the error connected. Sometimes for n, the number of dimensions, the limit $n \to \infty$ is taken to get rid of "unpleasant" terms in the calculations such that the significance of the results concerning the n-dependence is corrupted. Moreover, most results deal with the effects of a single step and are not strong enough to obtain results on the longtime behavior. As far as we know, no results containing a theorem on how the adaptation of the mutation vector's length affects the runtime of the (1+1) ES have been published yet.

The general description of the (1+1) ES given above captures simulated annealing for instance. The concrete (1+1) ES that is analyzed here on SPHERE is a randomized hill climber, namely, the mutant $\boldsymbol{x}$ replaces the current search point $\boldsymbol{c}$ iff $f(\boldsymbol{x}) \leq f(\boldsymbol{c})$. The distribution of the mutation vector is assumed to have the following property.

Definition 1. *For $\boldsymbol{m} \in \mathbb{R}^n$ let $|\boldsymbol{m}|$ denote its length/L_2-norm and $\widehat{\boldsymbol{m}} := \boldsymbol{m}/|\boldsymbol{m}|$ the normalized vector. The random mutation vector $\boldsymbol{m}$ is* isotropically distributed *iff $|\boldsymbol{m}|$ is independent of $\widehat{\boldsymbol{m}}$ and $\widehat{\boldsymbol{m}}$ is uniformly distributed upon the unit hypersphere $U := \{\boldsymbol{x} \in \mathbb{R} \mid |\boldsymbol{x}| = 1\}$. (That is, for $\boldsymbol{u}, \boldsymbol{v} \in U$ the probability density of $\{\widehat{\boldsymbol{m}} = \boldsymbol{u}\}$ equals the one of $\{\widehat{\boldsymbol{m}} = \boldsymbol{v}\}$.)*

In less formal words, if a mutation is isotropic, all directions are equiprobable and the (random) length of the mutation vector is independent of its direction. These two assumptions (on the selection rule and the mutation vector's distribution) are taken for the lower bound on the expected runtime of the (1+1) ES on SPHERE. Consequently, this result is valid for any adaptation of the length of isotropic mutation vectors.

For the upper bound on the (expected) runtime shown here the following type of mutation is considered which is common in practice.

Definition 2. *Let $\widetilde{\boldsymbol{m}} \in \mathbb{R}^n$ be $(N_1(0,1), \ldots, N_n(0,1))$-distributed (each component is independently standard normal distributed). A mutation is called* Gauss-mutation *iff the mutation vector's distribution equals the one of $\lambda \cdot \widetilde{\boldsymbol{m}}$, $\lambda \in \mathbb{R}^+$.*

In fact, Gauss-mutations are isotropically distributed (cf. Lemma 6). For the upper bound, the length of the Gauss-mutation vectors is adjusted by an instantiation of the well-know 1/5-rule for the adaptation of the scaling factor s (also due to Rechenberg):

> The rule that determines the scaling factor s within a run of the (1+1) ES is called *1/5-rule* iff it aims to ensure that in each step the mutant is accepted with probability 0.2 and this is done utilizing only the (relative) frequency of successful steps.

In Section 2 a closer look is taken at isotropic mutations and the geometric properties of SPHERE are discussed. Subsequently, these results are applied in Section 3 to obtain the probability density function of an isotropic mutation's spatial gain (in the search space $\mathbb{R}^n$) parallel to a fixed direction. In Section 4 this density function is used to estimate the success probability of a step, and in Section 5 the expected spatial gain in one step of the (1+1) ES is estimated. The main results described in the abstract are finally obtained in Section 6. All ideas are presented in this paper, yet some calculations are carried out in appendices. The following abbreviations will be used for better readability.

Definition 3. *A probability $p(n)$ is* exponentially small *in n iff 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 iff $\mathsf{P}\{\neg A(n)\}$ is exponentially small in n.*

If the parameter as to which a probability is (not) exponentially small is clear from the context naming it explicitly may be omitted.

2 Isotropic mutations and SPHERE

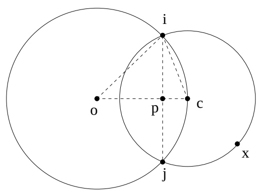
The special properties of isotropic mutations can be utilized to estimate *a step's success probability*, the probability that an acceptable mutant is generated in this step. Let $\boldsymbol{m} \in \mathbb{R}^n$ be generated by an isotropic mutation, and let $\boldsymbol{c} \in \mathbb{R}^n$ denote the current search point. The independence of $|\boldsymbol{m}|$ and $\widehat{\boldsymbol{m}}$ is crucial to the analysis presented here: It enables the application of the concept of "deferred decisions." We may assume that the mutation vector's length l is chosen according to $|\boldsymbol{m}|$'s distribution first such that the candidate search point is uniformly distributed upon the n-sphere formed by all points having distance l from $\boldsymbol{c}$. Let $S_{\boldsymbol{c},l} \subset \mathbb{R}^n$ denote this hyper-sphere and $A_{\boldsymbol{c}} := \{\boldsymbol{x} \in \mathbb{R}^n \mid \boldsymbol{x} \text{ would replace } \boldsymbol{c}\}$ the set of all acceptable points. As a hypersurface area is an $(n{-}1)$-volume,

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

Proposition 1. *Let G denote the spatial gain towards the optimum in a step of the (1+1) ES. If the mutation vector $\boldsymbol{m}$ is isotropically distributed then $\mathsf{E}[G] = \mathsf{E}[\,\mathsf{E}[G \mid |\boldsymbol{m}|\,]\,]$ in this step. If $g := \sup_l \mathsf{E}[G \mid |\boldsymbol{m}| = l\,]$ exists, then $\mathsf{E}[G] \le g$.*

In general, situations can occur in which $A \cap S \subset \mathbb{R}^n$ may be countable or actually empty. Furthermore, even if $A \cap S$ is uncountable, it may have zero $(n{-}1)$-volume—implying zero probability of success in the step concerned.

Crucial to the argumentation on SPHERE is that due to the selection rule the distance from the optimum is non-increasing since $\text{SPHERE}(\boldsymbol{x}) \le \text{SPHERE}(\boldsymbol{c}) \Leftrightarrow |\boldsymbol{x}| \le |\boldsymbol{c}|$. For $r \in \mathbb{R}^+$ the set $\{\boldsymbol{x} \in \mathbb{R}^n \mid \text{SPHERE}(\boldsymbol{x}) = r^2\}$ equals the hyper-sphere with radius r centered at the origin and $\{\boldsymbol{x} \in \mathbb{R}^n \mid \text{SPHERE}(\boldsymbol{x}) \le r^2\}$ the corresponding hyper-ball. Consequently, when SPHERE is minimized, $A \cap S$ has a positive $(n{-}1)$-volume iff $0 < |\boldsymbol{m}| < 2\,|\boldsymbol{c}|$, because $|\boldsymbol{c}|$ equals the distance from the origin $\boldsymbol{o} = (0, \ldots, 0)$ which is the optimum of SPHERE, and A equals the n-ball with radius $|\boldsymbol{c}|$ centered at the origin/optimum.



The n-sphere $\{\boldsymbol{c}' \in \mathbb{R}^n \mid |\boldsymbol{c}'| = |\boldsymbol{c}|\}$ will be called *fitness sphere* and the n-sphere $\{\boldsymbol{x} \in \mathbb{R}^n \mid |\boldsymbol{x} - \boldsymbol{c}| = |\boldsymbol{m}|\}$ *mutation sphere*. Let $I \subset \mathbb{R}^n$ denote the intersection of the two spheres. Since for all $\boldsymbol{i}, \boldsymbol{j} \in I$, $|\boldsymbol{i} - \boldsymbol{c}| = |\boldsymbol{j} - \boldsymbol{c}|$ and $|\boldsymbol{i}| = |\boldsymbol{j}|$ (implying that the two triangles defined by $\boldsymbol{i}, \boldsymbol{o}, \boldsymbol{c}$ resp. $\boldsymbol{j}, \boldsymbol{o}, \boldsymbol{c}$ are congruent), I is a subset of a hyperplane P which is orthogonal to the line segment $\overline{\boldsymbol{oc}}$. Let $\boldsymbol{p} \in P$ denote the point where this line intersects P. Then $|\boldsymbol{i} - \boldsymbol{p}| = |\boldsymbol{j} - \boldsymbol{p}|$ for all $\boldsymbol{i}, \boldsymbol{j} \in I$. Thus, in the flat $(n{-}1)$-subspace P, I forms a hyper-sphere centered at $\boldsymbol{p}$. In short, the mutation sphere's part lying inside the fitness sphere forms a hyper-spherical cap. Since the candidate search point is uniformly distributed upon the mutation sphere, the success probability of the mutation considered equals the relative share (regarding hypersurface area) of the mutation sphere lying inside the fitness sphere. Now, the interesting question is how this ratio depends on the *scaled distance from the optimum* $|\boldsymbol{c}|/|\boldsymbol{m}|$ and the number of dimensions n.

As "Pythagoras is right" in any dimension, for $|\boldsymbol{c} - \boldsymbol{p}| \le |\boldsymbol{c}|$ and any $\boldsymbol{i} \in I$, $|\boldsymbol{i}|^2 - |\boldsymbol{p}|^2 = |\boldsymbol{i} - \boldsymbol{p}|^2 = |\boldsymbol{i} - \boldsymbol{c}|^2 - |\boldsymbol{c} - \boldsymbol{p}|^2$. Since $|\boldsymbol{i}| = |\boldsymbol{c}|$ and $|\boldsymbol{i} - \boldsymbol{c}| = |\boldsymbol{m}|$, solving $|\boldsymbol{c}|^2 - |\boldsymbol{p}|^2 = |\boldsymbol{m}|^2 - |\boldsymbol{c} - \boldsymbol{p}|^2$ yields $|\boldsymbol{c} - \boldsymbol{p}| = |\boldsymbol{m}|^2/(2|\boldsymbol{c}|)$ for $|\boldsymbol{m}| \le \sqrt{2}|\boldsymbol{c}|$. Thus, the height of the mutation sphere's cap that is cut off by the fitness sphere equals $|\boldsymbol{m}| - |\boldsymbol{c} - \boldsymbol{p}| = |\boldsymbol{m}| - |\boldsymbol{m}|^2/(2|\boldsymbol{c}|)$ and the ratio of height to radius $|\boldsymbol{m}|$ equals $1 - |\boldsymbol{m}|/(2|\boldsymbol{c}|)$. As shown in Appendix A, in n-space, $n \ge 3$, the ratio (regarding hypersurface area) of a hyper-spherical cap (height h) to the hyper-sphere it is cut off (radius r) equals $\Psi_{n-2}(\arccos(1 - h/r))/\Psi_{n-2}(\pi)$ where $\Psi_k(\gamma) := \int_0^\gamma (\sin\beta)^k d\beta$.

Since the relative height of the mutation sphere's cap equals $1-|\boldsymbol{m}|/(2\,|\boldsymbol{c}|)$, the ratio equals $\Psi_{n-2}\big(\arccos(\,|\boldsymbol{m}|/(2\,|\boldsymbol{c}|)\,)\big)\,/\,\Psi_{n-2}(\pi)$ in the situation above. This formula may be directly used to estimate a step's success probability, yet it can also be utilized to estimate the probability that an isotropic mutation hits an arbitrarily fixed cap of the mutation sphere.

3 Probability density of the mutation's spatial gain

Although the spatial gain towards the optimum in a step of the (1+1) ES on Sphere is the intermediate objective, for the moment we concentrate on the random variable G_m that corresponds to the mutation's spatial gain parallel to a fixed direction if the random variable $|\boldsymbol{m}|$ takes the value l. Since $\boldsymbol{m}$ is isotropically distributed, the mutation vector is uniformly distributed upon the hyper-sphere with radius l. Note that just the mutation is investigated, that is, the situation prior selection is examined. For instance, G_m equals m_1 if the spatial gain along the first axis is to be calculated.

Obviously, $\mathsf{P}\{G_m > g\} = \mathsf{P}\{G_m < -g\}$ and $\mathsf{P}\{G_m > l\} = 0 = \mathsf{P}\{G_m < -l\}$. Let S denote the hypersurface area of an n-sphere with radius l, and $C(h)$ the area of a cap that is cut off S and has height $h \in [0, l]$. For a fixed spatial gain $g \in [0, l]$, $\mathsf{P}\{G_m \geq g\} = C(l-g)/S$ where the cap's pole lies on the half line defined by the actual search point and the direction considered; the hyperplane that contains the boundary of the cap is orthogonal to the direction and has distance g from the current search point. Let $k = n-2 \geq 2$ be fixed. Since $h = l-g$, the formula for the ratio $C(h)/S$ in $(k+2)$-space yields for $\widehat{g} := g/l \in [0,1]$

$$\mathsf{P}\{G_m < g\} \;=\; 1 - \mathsf{P}\{G_m \geq g\} \;=\; 1 - \frac{C(l-g)}{S} \;=\; 1 - \frac{\Psi_k(\arccos \widehat{g})}{\Psi_k(\pi)}\,.$$

Thus, $F_k(x) := 1 - \Psi_k(\arccos x)/\Psi_k(\pi)$ for $x \in [0,1]$ is the probability distribution of G_m/l on $[0,1]$. Since Ψ_k is continuous, the probability density of $\{G_m/l = \widehat{g}\}$ for $\widehat{g} \in [0,1]$ equals $\frac{dF_k(x)}{dx}(\widehat{g}) = F_k'(\widehat{g})$,

$$F_k'(x) \;=\; \frac{-1}{\Psi_k(\pi)} \cdot \frac{d}{dx}\Psi_k(\arccos x) \;=\; \frac{-1}{\Psi_k(\pi)} \cdot \frac{d}{dx}\int_0^{\arccos x} (\sin\beta)^k \, d\beta\,.$$

Hence, the question is how the value of the integral on the right changes with x. As $\sin^k$ is continuous, let Sin_k denote its antiderivative such that $\mathrm{Sin}_k(0) = 0$. Then $\int_0^{\arccos x}(\sin y)^k\,dy \;=\; \mathrm{Sin}_k(\arccos x)$, and thus,

$$\frac{d}{dx}\mathrm{Sin}_k(\arccos x) \;=\; {\mathrm{Sin}_k}'(\arccos x)\cdot \arccos' x \;=\; (\sin\arccos x)^k \cdot \arccos' x\,.$$

Since $\sin\arccos x = \sqrt{1-x^2}$ and $\arccos' x = -1/\sqrt{1-x^2}$, finally

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

This proves that the probability density of $\{G_m = g\}$ for $g \in [0,l]$ equals

$$F_k'(g/l) \;=\; \frac{-1}{\Psi_k(\pi)}\cdot(-1)\cdot\left(1-(g/l)^2\right)^{(k-1)/2} \;=\; \Psi_k(\pi)^{-1}\,\left(1-(g/l)^2\right)^{(k-1)/2}$$

in $(k{+}2)$-space. Again, this is the density of the mutation's spatial gain parallel to an arbitrarily fixed direction—independently of the function optimized—if $|\boldsymbol{m}|$ takes the value l, not the one towards the optimum after selection.

4 Estimating a step's success probability

Before we use the density function just derived to estimate the spatial gain towards the optimum in a step of the (1+1) ES on SPHERE, we return to the success probability of a step, which has already been the subject in Section 2, and derive an alternative formula that enables us to estimate a step's success probability more comfortably.

If the random variable $|\boldsymbol{m}|$ takes the value l and distances are scaled by l^{-1} then $d := |\boldsymbol{c}|/l$ denotes the scaled distance from the optimum and the scaled height of the mutation sphere's cap cut off by the fitness sphere equals $1-1/(2d)$ (cf. Section 2). Consequently, the mutation is accepted iff the mutation's scaled spatial gain parallel to $\overline{\boldsymbol{co}}$ is at least $1/(2d)$—because then and only then the mutation hits this cap containing all acceptable search points.

Proposition 2. *Let the (1+1) ES minimize* SPHERE*, and let $\boldsymbol{c}, \boldsymbol{m} \in \mathbb{R}^n$, $n = k+2 \geq 4$, denote the current search point resp. the isotropically distributed mutation vector in some step. If $|\boldsymbol{m}|$ takes the value l, the success probability of this step equals $\Psi_k(\pi)^{-1} \cdot \int_{1/(2d)}^{1} \left(1-x^2\right)^{(k-1)/2} dx$, where $\Psi_k(\pi) = \int_0^\pi (\sin x)^k dx$ and $d = |\boldsymbol{c}|/l$.*

With respect to the $1/5$-rule, which will be used for the upper bound on the expected runtime, one might ask which length of the mutation vector results in a step of the (1+1) ES having success probability 0.2. This question can now be answered.

Lemma 1. *Let the (1+1) ES minimize* SPHERE*, and let $\boldsymbol{c}, \boldsymbol{m} \in \mathbb{R}^n$ denote the current search point resp. the isotropically distributed mutation vector in some step. This step's success probability is lower bounded by a constant greater than 0 and upper bounded by a constant smaller than $1/2$ iff $|\boldsymbol{m}| = \Theta(|\boldsymbol{c}|/\sqrt{n})$.*

Proof. That $\sqrt{2\pi}/\sqrt{k+1} < \Psi_k(\pi) < \sqrt{2\pi}/\sqrt{k}$ for $k \geq 2$ can be found in Appendix B, and thus, $\Psi_k(\pi) = \Theta(1/\sqrt{k})$. Let $n = k+2 \geq 4$ and $d = |\boldsymbol{c}|/|\boldsymbol{m}| = |\boldsymbol{c}|/\Theta(|\boldsymbol{c}|/\sqrt{k}) = \Theta(\sqrt{k})$. The spatial gain parallel to $\overline{\boldsymbol{co}}$ (cf. G_m in Section 3) is negative resp. positive with probability $1/2$, respectively; if it is negative, a success is precluded. With probability

$$\begin{aligned}
\Psi_k(\pi)^{-1} \int_0^{1/(2d)} (1-x^2)^{(k-1)/2} dx \quad &> \quad \Psi_k(\pi)^{-1}\,(2d)^{-1}\,(1-(2d)^{-2})^{(k-1)/2} \\
&= \quad \Psi_k(\pi)^{-1}\,\Theta(1/\sqrt{k})\,(1-\Theta(1/k))^{(k-1)/2} \\
&= \quad \Psi_k(\pi)^{-1}\,\Theta(1/\sqrt{k})\,\Theta(1) \quad = \quad \Theta(1)
\end{aligned}$$

it is positive, but still the mutant lies outside the fitness sphere such that the mutation is rejected. Finally, with probability

$$\begin{aligned}
\Psi_k(\pi)^{-1} \int_{1/(2d)}^{1} (1-x^2)^{(k-1)/2} dx \quad &> \quad \Psi_k(\pi)^{-1} \int_{1/(2d)}^{1/d} (1-x^2)^{(k-1)/2} dx \\
&> \quad \Psi_k(\pi)^{-1}\,(1/d-(2d)^{-1})\,(1-(1/d)^2)^{(k-1)/2} \\
&= \quad \Psi_k(\pi)^{-1}\,(2d)^{-1}\,(1-\Theta(1/k))^{(k-1)/2} \\
&= \quad \Psi_k(\pi)^{-1}\,\Theta(1/\sqrt{k})\,\Theta(1) \quad = \quad \Theta(1)
\end{aligned}$$

the mutant lies inside the fitness sphere and is accepted. □

As $|\boldsymbol{m}| = \Theta(|\boldsymbol{c}|/\sqrt{n})$ is equivalent to the mutation sphere's cap that is cut off by the fitness sphere having scaled height $1-1/(2d) = 1-\Theta(1/\sqrt{n})$, the result can be read as follows: An isotropic mutation $\boldsymbol{m} \in \mathbb{R}^n$ hits a cap having height

$|\boldsymbol{m}| \cdot \bigl(1 - \Theta(1/\sqrt{n})\bigr)$ with a probability in $[a, b] \subset (0, 1/2)$ for two constants a and b which depend on the constants in the Θ-notation.

It is clear that a step's success probability approaches $1/2$ as $|\boldsymbol{m}| / |\boldsymbol{c}| \to 0$; the interesting question is how the success probability changes when the mutation vector gets too long. In fact, it turns out that a mutation is rejected w. o. p. if $|\boldsymbol{m}| = \Omega(|\boldsymbol{c}| \cdot n^{\varepsilon - 1/2})$ for a positive constant ε in the step considered.

5 Expected spatial gain in one step

As Lemma 1 implies that the length of the mutation vector would be in $\Theta(|\boldsymbol{c}|/\sqrt{n})$ if the 1/5-rule was able to ensure a success probability of exactly 0.2 in the step considered, the mutation's expected spatial gain towards the optimum in this situation is of particular interest and is estimated in the following.

Lemma 2. *Let the (1+1) ES minimize* Sphere, *and let $\boldsymbol{c}, \boldsymbol{m} \in \mathbb{R}^n$ denote the current search point resp. the isotropically distributed mutation vector in some step. If $|\boldsymbol{m}| = \Theta(|\boldsymbol{c}|/\sqrt{n})$, this step's spatial gain is in $\Omega(|\boldsymbol{m}|/\sqrt{n}) = \Omega(|\boldsymbol{c}|/n)$ with probability $\Theta(1)$, and thus, the expected decrease in distance to the optimum is also in $\Omega(|\boldsymbol{m}|/\sqrt{n}) = \Omega(|\boldsymbol{c}|/n)$ in this step.*

Proof. The condition yields $d = |\boldsymbol{c}|/|\boldsymbol{m}| = \sqrt{n}/\lambda$ with $\lambda = \Theta(1)$; d denotes the scaled distance from the optimum. Let C denote the mutation sphere's cap that is cut off by the fitness sphere. Then the scaled height of C equals $1 - 1/(2d) = 1 - \lambda/(2\sqrt{n}) = 1 - \Theta(1/\sqrt{n})$, and the candidate search point belongs to this cap with a constant probability in $(0, 1/2)$ according to Lemma 1.

Let $B \subset C$ denote the cap with height $1 - 1/d = 1 - \lambda/\sqrt{n}$ such that its pole coincides with the one of C. Then each point belonging to B is at least $1/d - 1/(2d) = 1/(2d)$ scaled 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 $1/(2d) = \Theta(1/\sqrt{n})$ scaled distance units if the candidate search point is in B. This happens with probability $\Theta(1)$ because the scaled height of B equals $1 - \Theta(1/\sqrt{n})$ like the one of C. Since a negative spatial gain is precluded, the expected decrease in distance to the optimum is lower bounded by $\Theta(1) \cdot |\boldsymbol{m}| \cdot \Omega(1/\sqrt{n}) = \Omega(|\boldsymbol{m}|/\sqrt{n}) = \Omega(|\boldsymbol{c}|/n)$. □

Consequently, if the 1/5-rule is capable of adjusting the mutation vector's length such that the success probability is close to 0.2, the distance to the optimum is expected to decrease by an $\Omega(1/n)$-fraction. Naturally, one might ask if some other mutation strength causes an expected spatial gain that is in $\omega(|\boldsymbol{c}|/n)$. In fact, we will show that the expected spatial gain towards the optimum is in $O(|\boldsymbol{c}|/n)$ regardless of the adaptation of the mutation vector's length—as long as isotropic mutations are used. As a consequence, the 1/5-rule indeed tries to adjust the length of the mutation vector to have optimal order $\Theta(|\boldsymbol{c}|/\sqrt{n})$ such that the expected spatial gain towards the optimum has maximal order $\Theta(|\boldsymbol{c}|/n)$.

Obviously, the spatial gain of a step equals 0 if the mutation is rejected (the mutant is not selected) and is upper bounded by the mutation's spatial gain parallel to $\overline{\boldsymbol{co}}$, otherwise. As mentioned above, when Sphere is minimized, a mutation is accepted iff the spatial gain of the mutation parallel to $\overline{\boldsymbol{co}}$ is at least $|\boldsymbol{m}|/(2d)$, $d = |\boldsymbol{c}|/|\boldsymbol{m}|$. Using the probability density of a mutation's spatial gain (parallel to a fixed direction) obtain in Section 3 (F_k' on page 5), the expected spatial gain of a step is upper bounded by $|\boldsymbol{m}| \cdot \int_{1/(2d)}^{1} x \, F_k'(x)\, dx$. Since

$$\int x\,(1 - x^2)^{(k-1)/2}\, dx \quad = \quad \frac{(1 - x^2)^{(k+1)/2}}{-(k+1)}\,,$$

$$\mathsf{E}[\text{gain}] \quad \le \quad \frac{|\boldsymbol{m}|}{\Psi_k(\pi)} \int_{1/(2d)}^{1} x\,(1-x^2)^{(k-1)/2} dx \quad = \quad \frac{|\boldsymbol{m}| \cdot (1-(2d)^{-2})^{(k+1)/2}}{\Psi(\pi)\cdot(k+1)}.$$

As $\Psi_k(\pi) > \sqrt{2\pi}/\sqrt{k+1}$, finally

$$\mathsf{E}[\text{gain}] \quad < \quad \frac{|\boldsymbol{m}|}{\sqrt{2\pi(k+1)}} \cdot \left(1-(2d)^{-2}\right)^{(k+1)/2}$$

in $(k+2)$-space. Therefore, $\mathsf{E}[\text{gain}] = O(|\boldsymbol{m}|/\sqrt{n})$ even if $d = |\boldsymbol{c}|/|\boldsymbol{m}| \to \infty$.

Furthermore, this inequality enables the proof that the expected spatial gain is in $O(|\boldsymbol{c}|/n)$ for any choice of the mutation vector's length.

Lemma 3. *Let the (1+1) ES minimize* Sphere*, and let $\boldsymbol{c} \in \mathbb{R}^n$ denote the current search point. If the mutation vector is generated isotropically, the expected spatial gain towards the optimum is in $O(|\boldsymbol{c}|/n)$.*

Proof. To prove this claim it must be shown that even if the mutation vector's length is chosen such that the expected spatial gain is maximized, this expected gain is in $O(|\boldsymbol{c}|/k)$, $k = n-2 \ge 4$. When distances are scaled by $|\boldsymbol{m}|^{-1}$, the analogous question is which scaled distance from the optimum maximizes the ratio of expected scaled gain to scaled distance.

Let $d = |\boldsymbol{c}|/|\boldsymbol{m}|$ denote the scaled distance from the optimum. Applying the upper bound on the expected spatial gain from above yields

$$\mathsf{E}[\text{scaled gain}] \,/\, d \quad < \quad \frac{1}{\sqrt{2\pi\,(k+1)}} \cdot \underbrace{\left(1-(2d)^{-2}\right)^{(k+1)/2} / \, d}_{=:\ w_k(d)} .$$

Hence, an upper bound on $\mathsf{E}[\text{scaled gain}]\,/d$ can be derived by maximizing the function w_k. This is done in Appendix C with the result $w_k(d) \le w_k(\sqrt{k}/2)$ for $d > 0$. Since $\mathsf{E}[\text{scaled gain}]\,/d < w_k(d)/\sqrt{2\pi(k+1)}$ and

$$w_k(d) \;\le\; w_k(\sqrt{k}/2) \;=\; (2/\sqrt{k})\cdot(1-1/k)^{(k-1)/2} \;=\; (2/\sqrt{k})\cdot\Theta(1) \;=\; \Theta(1/\sqrt{k}),$$

finally $\mathsf{E}[\text{scaled gain}]\,/d < \Theta(1/\sqrt{k})/\sqrt{2\pi(k+1)} = \Theta(1/k)$. Hence, even if the the scaled distance from the optimum $d = |\boldsymbol{c}|/|\boldsymbol{m}|$ is optimal,

$$\mathsf{E}[\text{gain}] \;=\; \frac{|\boldsymbol{c}|\cdot|\boldsymbol{m}|\cdot\mathsf{E}[\text{scaled gain}]}{|\boldsymbol{c}|} \;=\; \frac{|\boldsymbol{c}|\cdot\mathsf{E}[\text{scaled gain}]}{d} \;=\; |\boldsymbol{c}|\cdot O(1/k). \qquad \square$$

That the expected spatial gain is in $O(|\boldsymbol{m}|/\sqrt{n})$ and even in $O(|\boldsymbol{c}|/n)$ regardless of the mutation vector's length does not preclude that the spatial gain has a greater order with a certain probability. In fact, the results obtained enable the proof that the spatial gain is $o(|\boldsymbol{m}|\cdot n^{\varepsilon-1/2})$ w. o. p. for any positive constant ε.

6 Expected gain in multiple steps / Expected runtime

As the (1+1) ES (in general) doesn't optimize, but actually approximates Sphere, it is not evident what the term "runtime" means. If runtime is defined as the number of steps necessary to halve the distance from the/an optimum, then linear runtime aligns with linear convergence, for instance.

Obviously, the runtime depends on the mutation adaptation the (1+1) ES uses when minimizing Sphere; but a lower bound on the (expected) runtime does not, as optimal mutation adaptation can be assumed. Since the expected spatial gain in one step is maximized if the distribution of $|\boldsymbol{m}|$ is concentrated on a length depending on the current distance $|\boldsymbol{c}|$, intuition says that the following "greedy" mutation adaptation is theoretically optimal: In each step $\boldsymbol{m}$'s isotropic distribution is chosen such that $|\boldsymbol{m}|$ is concentrated at the value maximizing the expected spatial gain in this single step.

Lemma 4. *Let the (1+1) ES minimize* Sphere, *and let $\boldsymbol{c} \in \mathbb{R}^n$ denote the current search point. If the mutation vector is generated isotropically, the number of steps necessary to obtain an expected spatial gain of $\Theta(|\boldsymbol{c}|)$ is in $\Omega(n)$.*

Proof. As shown in Section 5, if for a given search point $\boldsymbol{y} \in \mathbb{R}^n$ the mutation strength is chosen such that the expected spatial gain is maximized, it is in $\Theta(|\boldsymbol{y}|/n)$. Due to the symmetry and scalability properties of Sphere, the maximal expected spatial gain equals $\xi \cdot |\boldsymbol{y}|/n$ for some $\xi = \Theta(1)$ which is independent of $|\boldsymbol{y}|$. Let $D_{r,M}$ denote the random variable that corresponds to the distance from the optimum after a step in which some $\boldsymbol{y}$ with $|\boldsymbol{y}| = r$ is mutated using the distribution M for the mutation vector's length. The results in this paper yield that for fixed radius of the fitness sphere r, $\mathsf{E}[D_{r,M}]$ is minimal if M is concentrated at a specific length. Consequently, $\arg\min_M \mathsf{E}[D_{r,M}]$ exists, and moreover, $\min_M \mathsf{E}[D_{r,M}] = r - \xi \cdot r/n = r \cdot (1 - \xi/n)$.

Now the situation is examined in which the total expected spatial gain of s consecutive steps is to be maximized. Naturally, in the last step the mutation strength is chosen in the way that the spatial gain (in this last step) is maximized. Since the maximal expected spatial gain is monotone in the distance from the optimum, it is by no means obvious that being greedy in each step maximizes the total expected spatial gain; but intuition will prove right in this case.

Let D_{s-1} denote the random variable that corresponds to the distance from the optimum after the first $s-1$ steps and f_{s-1} its density (function). Naturally, f_{s-1} depends on the initial distance from the optimum and the choice of the mutation strength in the first $s-1$ steps. Furthermore, let D_s denote the random variable that corresponds to the distance from the optimum after the s-th step. Since in this last step the mutation strength is chosen in the way that the resulting expected distance from the optimum is minimized,

$$\begin{aligned}
\mathsf{E}[D_s] &= \int_{-\infty}^{\infty} \min_M \mathsf{E}[D_{r,M}] \cdot f_{s-1}(r)\, dr &&= \int_{-\infty}^{\infty} r \cdot (1 - \xi/n) \cdot f_{s-1}(r)\, dr \\
&= (1 - \xi/n) \cdot \int_{-\infty}^{\infty} r \cdot f_{s-1}(r)\, dr &&= (1 - \xi/n) \cdot \mathsf{E}[D_{s-1}]
\end{aligned}$$

(the third equality is due to ξ's independence of r). Consequently, $\mathsf{E}[D_s]$ indeed takes its minimum if $\mathsf{E}[D_{s-1}]$ is minimal. Hence, to maximize the expected spatial gain in s steps the expected spatial gain in the first $s-1$ steps is to be maximized. Starting with $s = 2$, by induction it follows that choosing the mutation strength in each step such that the expected spatial gain is maximal, respectively, also maximizes the total expected spatial gain. After s greedy steps the expected distance from the optimum—which is now proved minimal—equals $|\boldsymbol{c}| \cdot (1 - \xi/n)^s$ due to linearity of expectation. Finally, $(1 - \xi/n)^s \leq 1 - \varepsilon$ for $\xi \in \Theta(1)$ and a constant $\varepsilon \in (0,1)$ implies $s = \Omega(n)$. □

This result raises the conjecture that the expected number of steps to obtain a predefined spatial gain of $\Theta(|\boldsymbol{c}|)$—for instance, to halve the distance from the optimum—is in $\Omega(n)$ even if the mutation adaptation works theoretically perfect. This can be proved with the following modification of Wald's equation.

Lemma 5. *Let $X_1, X_2, \ldots$ denote random variables with bounded range*[1] *and T the random variable defined by $T = \min\{\, t \mid X_1 + \cdots + X_t \geq g\}$ for a given $g > 0$. If $\mathsf{E}[T]$ exists and $\mathsf{E}[X_i \mid T \geq i] \leq u$ for $i \in \mathbb{N}$ and $u > 0$, then $\mathsf{E}[T] \geq g/u$.*

[1] that is, $\inf X_i$ and $\sup X_i$ exist for $i \in \mathbb{N}$

Proof. Obviously $T \geq 1$, and for $i \geq 2$ the condition $T \geq i$ is equivalent to $X_1 + \cdots + X_k < g$ for $1 \leq k < i$. Since the X_i are bounded, $\mathsf{E}[X_1 + \cdots + X_T]$ also exists if $\mathsf{E}[T]$ exists. The proof follows the one of Wald's equation (up to the point where the upper bound on $\mathsf{E}[X_i \mid T \geq i]$ is utilized rather than the original assumption that the X_i are independent and identically distributed).

$$
\begin{aligned}
g &\leq \mathsf{E}[X_1 + \cdots + X_T] \\
&= \sum_{t=1}^{\infty} \mathsf{P}\{T = t\} \cdot \mathsf{E}[X_1 + \cdots + X_t \mid T = t] \\
&= \sum_{t=1}^{\infty} \mathsf{P}\{T = t\} \cdot \sum_{i=1}^{t} \mathsf{E}[X_i \mid T = t] \\
&= \sum_{t=1}^{\infty} \sum_{i=1}^{t} \mathsf{P}\{T = t\} \cdot \mathsf{E}[X_i \mid T = t] \\
&\ \text{since the series converges absolutely due to the boundedness of the } X_i \\
&= \sum_{i=1}^{\infty} \sum_{t=i}^{\infty} \mathsf{P}\{T = t\} \cdot \mathsf{E}[X_i \mid T = t] \\
&= \sum_{i=1}^{\infty} \sum_{t=i}^{\infty} \mathsf{P}\{T = t \mid T \geq i\} \cdot \mathsf{P}\{T \geq i\} \cdot \mathsf{E}[X_i \mid T = t] \\
&= \sum_{i=1}^{\infty} \mathsf{P}\{T \geq i\} \cdot \sum_{t=i}^{\infty} \mathsf{P}\{T = t \mid T \geq i\} \cdot \mathsf{E}[X_i \mid T = t] \\
&\ \text{since } t \geq i,\ T = t \text{ implies } T \geq i \\
&= \sum_{i=1}^{\infty} \mathsf{P}\{T \geq i\} \cdot \sum_{t=i}^{\infty} \mathsf{P}\{T = t \mid T \geq i\} \cdot \mathsf{E}[X_i \mid T = t \wedge N \geq i] \\
&\ \text{since } t < i \text{ implies } \mathsf{P}\{T = t \mid T \geq i\} = 0 \\
&= \sum_{i=1}^{\infty} \mathsf{P}\{T \geq i\} \cdot \sum_{t=1}^{\infty} \mathsf{P}\{T = t \mid T \geq i\} \cdot \mathsf{E}[X_i \mid T = t \wedge T \geq i] \\
&= \sum_{i=1}^{\infty} \mathsf{P}\{T \geq i\} \cdot \mathsf{E}[X_i \mid T \geq i] \\
&\leq \sum_{i=1}^{\infty} \mathsf{P}\{T \geq i\} \cdot u \\
&= \mathsf{E}[T] \cdot u \qquad \square
\end{aligned}
$$

Theorem 1. *Let the (1+1) ES minimize* SPHERE *using isotropic mutations, and let $\boldsymbol{c} \in \mathbb{R}^n$ denote the current search point. Independently of the mutation adaptation used the expected number of steps necessary to to obtain a spatial gain of $\Theta(|\boldsymbol{c}|)$ is in $\Omega(n)$.*

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 T denote the (random) number of steps the (1+1) ES needs to realize the postulated gain. As SPHERE is monotone in the distance from the optimum, $X_i \geq 0$, and since every accepted mutant is at most $|\boldsymbol{c}|$ distance units away from the optimum, $X_i \leq |\boldsymbol{c}|$ and $\mathsf{E}[X_i \mid T \geq i] = O(|\boldsymbol{c}|/n)$ according to Lemma 3. If $\mathsf{E}[T]$ exists, then $0 < g := \sup_i \mathsf{E}[X_i \mid T \geq i] = O(|\boldsymbol{c}|/n)$, and consequently, $\mathsf{E}[T] \geq \Theta(|\boldsymbol{c}|)/g = \Omega(n)$ according to Lemma 5.

If the series that corresponds to $\mathsf{E}[T]$ does not converge absolutely, one may informally argue that "$\mathsf{E}[T] = \infty = \Omega(n)$" since T is positive. $\square$

Moreover, it can be proved that the number of steps necessary to obtain a spatial gain of $\Theta(|c|)$ is in $\omega(n^{1-\varepsilon})$ w. o. p. for any positive constant ε.

The lower bound on the expected runtime holds independently of the mutation adaptation applied since theoretically optimal adaptation is assumed. An upper bound, however, crucially depends on the mutation adaptation used. Next, a matching upper bound on the expected runtime of (1+1) ES on SPHERE using Gauss-mutations and the following instantiation of the 1/5-rule to adapt the mutation vector's length is shown.

Every n steps the relative success frequency of the last n steps is evaluated. If it is smaller than 0.2 the scaling factor is halved, otherwise doubled.

The following properties of Gauss-mutations are proved in Appendix D.

Lemma 6. *A Gauss-mutation $\boldsymbol{m} \in \mathbb{R}^n$ is isotropically distributed, and moreover, $l := \mathsf{E}[|\boldsymbol{m}|]$ exists and $\mathsf{P}\{\big||\boldsymbol{m}| - l\big| \geq \delta \cdot l\} \leq \delta^{-2}/(n - 1/2)$.*

Let $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_n$ denote independent copies of $\boldsymbol{m}$. Then for any constant $\lambda < 1$, two positive constants $a(\lambda)$ and $b(\lambda)$ exist such that for the cardinality of $I := \{i \mid a(\lambda) \cdot l \leq |\boldsymbol{m}_i| \leq b(\lambda) \cdot l\}$ w. o. p. $\#I \geq \lambda n$.

As the adaptation of the scaling factor is done every n steps, the run of the (1+1) ES is virtually partitioned into phases each of which lasts n steps such that $\mathsf{E}[|\boldsymbol{m}|]$ is constant in each phase. Let s_i denote the scaling factor used throughout the i-th phase and l_i the corresponding expected length of the mutation vectors. Furthermore, let r_i denote the relative frequency of successful steps in the i-th phase and d_i the distance from the optimum at the beginning of this phase; hence, $d_i - d_{i+1}$ equals the spatial gain in the i-th phase. Furthermore, let p_i denote the success probability in the first step of the i-th phase. Note these simple, but crucial facts that are due to the symmetry/scalability of SPHERE: During a phase the steps' success probabilities are non-increasing (as the distance from the optimum is non-increasing), and $p_i > p_j$ iff $d_i/s_i > d_j/s_j$.

Lemma 7. *Let the (1+1) ES minimize* SPHERE *using Gauss-mutations and the 1/5-rule defined above. If $r_i \geq 0.2$ and $r_{i+1} < 0.2$ resp. if $r_i < 0.2$ and $r_{i+1} \geq 0.2$, then $d_{i+2} = d_i - \Omega(d_i)$ w. o. p., that is, the distance from the optimum is reduced by a constant fraction w. o. p. in these two phases.*

Proof. Assume the opposite (assumption A1): $d_{i+2} = d_i - o(d_i)$. As A1 implies $d_{i+2} = \Theta(d_{i+1}) = \Theta(d_i)$ (in addition to $d_{i+2} \leq d_{i+1} \leq d_i$), the order of distance from the optimum does not change in the two phases. As $s_{i+1} = 2s_i$ resp. $s_{i+1} = s_i/2$, $l_{i+1} \in \{2\,l_i, l_i/2\}$, in other words, $\mathsf{E}[|\boldsymbol{m}|]$ varies only by a factor of 2 in the two phases.

Assume that $\mathsf{E}[|\boldsymbol{m}|] \neq \Theta(d_i/\sqrt{n}))$ in the two phases (assumption A2). Then Lemma 6 yields that w. o. p. $|\boldsymbol{m}| \neq \Theta(d_i/\sqrt{n})$ in $0.9n$ steps in each of the two phases. According to Lemma 1, the success probability is either in $o(1)$ or lower bounded by $1/2 - o(1)$ in each of these steps. By Chernoff-bounds, the probability of at least $0.2n$ successful steps in the one phase and fewer than $0.2n$ in the other is exponentially small either way. Thus, w. o. p. $\mathsf{E}[|\boldsymbol{m}|] = \Theta(d_i/\sqrt{n})$ in the two phases such that w. o. p. $\neg$A2.

Finally, it must be shown that $\mathsf{E}[|\boldsymbol{m}|] = \Theta(d_i/\sqrt{n})$ in the two phases implies that w. o. p. $\neg$A1. For a search point with distance $\Theta(d_i)$ from the optimum the spatial gain is in $\Omega(d_i/n)$ with probability $\Theta(1)$ if $|\boldsymbol{m}| = \Theta(d_i/\sqrt{n})$ by Lemma 2. As $|\boldsymbol{m}| = \Theta(\mathsf{E}[|\boldsymbol{m}|]) = \Theta(d_i/\sqrt{n})$ w. o. p. in $0.9n$ steps in each of the two phases according to Lemma 6, the number of steps in each of which the spatial gain is in $\Omega(d_i/n)$ is w. o. p. in $\Theta(n)$ in each of the two phases by Chernoff-bounds. Consequently, the total spatial gain is in $\Omega(d_i)$ w. o. p. in each of the two phases—implying that w. o. p. $\neg$A1. $\square$

A run of the (1+1) ES is considered as a sequence of phases; this sequence is notated as follows: A phase in which $r_i < 0.2$, such the scaling factor is halved, is symbolized by "$\div$" and "$\times$" symbolizes a phase in which $r_j \geq 0.2$, such that the scaling factor is doubled. Specifying sequences of phases by regular expressions[2] over the alphabet $\{\times, \div\}$, the lemma above deals with $\times\div$ resp. $\div\times$ subsequences and implies that the (1+1) ES converges linearly w. o. p. for polynomially many phases if the corresponding sequence is in $\{\div\times, \times\div\}^*$. To obtain the main result, the argumentation in the proof of Lemma 7 must be extended to subsequences $\div\times^b$ and $\times\div^b$ for $b \geq 3$.

Theorem 2. *Let the (1+1) ES minimize* SPHERE *in* $\mathbb{R}^n$ *using Gauss-mutations and the 1/5-rule defined above. Let the* i*-th phase be the first one such that* $r_i < 0.2$ *and* $r_{i+1} \geq 0.2$ *or such that* $r_i \geq 0.2$ *and* $r_{i+1} < 0.2$*. Beginning with the* i*-th phase the (1+1) ES converges linearly for any polynomial number of phases w. o. p. That is,* $d_{i+t+1} \leq 2^{-\Omega(t)} \cdot d_i$ *w. o. p. for* $t \in poly(n)$ *phases each of which lasts* n *steps.*

Proof. The sequence of the phases considered begins either with $\times\div$ or with $\div\times$. Subsequences in $\{\div\times, \times\div\}^*$ are covered by Lemma 7.

First we investigate the subsequence $\div\times^b$ of phases for $b \geq 3$. Assume these are the phases $j, \dots, j+b$ such that the phases $j+a$, $2 \leq a \leq b$, are not covered by Lemma 7. Nevertheless, the proof of this lemma yields that w. o. p. $l_j = \Theta(d_j/\sqrt{n})$ and that this w. o. p. results in $d_{j+1} = d_j - \Omega(d_j)$. We show that the conditions in the $(j{+}a)$-th phase are w. o. p. similar to the ones in the j-th phase. Since $s_{j+a} \geq s_j$, also $l_{j+a} \geq l_j$. Furthermore, $d_{j+a} \leq d_j$, and thus, "$l_j = \Theta(d_j/\sqrt{n})$ w. o. p." implies that w. o. p. $l_{j+a} = \Omega(d_{j+a}/\sqrt{n})$. That also w. o. p. $l_{j+a} = O(d_{j+a}/\sqrt{n})$ can be proved by showing that the assumption $l_{j+a} = \omega(d_{j+a}/\sqrt{n})$ leads to an exponentially small probability for $r_{j+a} \geq 0.2$ (cf. the proof of Lemma 7). Hence, w. o. p. $l_{j+a} = \Theta(d_{j+a}/\sqrt{n})$ implying that w. o. p. $d_{j+a+1} = d_{j+a} - \Omega(d_{j+a})$ (again cf. the proof of Lemma 7).

Now the subsequence $\times\div^b$ of phases is investigated for $b \geq 3$. Again, $2 \leq a \leq b$ and the first phase of the sequence is the j-th one in the run of the (1+1) ES. The proof of Lemma 7 yields that w. o. p. $l_{j+1} = \Theta(d_{j+1}/\sqrt{n})$ and that this w. o. p. results in $d_{j+2} = d_{j+1} - \Omega(d_{j+1})$. If $p_{j+a} \geq p_{j+1}$, then $d_{j+a}/s_{j+a} \geq d_{j+1}/s_{j+1}$ which is equivalent to $l_{j+a}/d_{j+a} \leq l_{j+1}/d_{j+1}$. In this case, the $(j{+}a)$-th phase resembles the $(j{+}1)$-th. Namely, "$l_{j+1} = \Theta(d_{j+1}/\sqrt{n})$ w. o. p." implies that w. o. p. $l_{j+a} = O(d_{j+a}/\sqrt{n})$. Furthermore, $r_{j+a} < 0.2$ implies that w. o. p. also $l_{j+a} = \Omega(d_{j+a}/\sqrt{n})$ (cf. the proof of Lemma 7). Consequently, if $p_{j+a} \geq p_{j+1}$ then w. o. p. $l_{j+a} = \Theta(d_{j+a}/\sqrt{n})$ implying that w. o. p. $d_{j+a+1} = d_{j+a} - \Omega(d_{j+a})$ (again cf. the proof of Lemma 7). Finally, the (sub-) case $p_{j+a} < p_{j+1}$ is investigated. Note that $s_{j+a} = s_{j+1}/2^{a-1}$, and thus, $p_{j+a} < p_{j+1}$ iff $d_{j+a} < d_j/2^{a-1}$. Altogether, either w. o. p. the distance from the optimum is reduced by a constant fraction in the $(j{+}a)$-th phase or in the preceding $a{-}1$ phases the distance from the optimum is at least halved $a{-}1$ times. By an accounting-method argument, in the latter case $d_{j+a+1} \leq \lambda^a d_j$ for a constant $\lambda \in (1/2, 1)$ even if the $(j{+}a)$-th phase yields no spatial gain. (Remember that negative spatial gain is precluded.)

All in all, the assumption that the sequence of the $t \in \text{poly}(n)$ phases considered starts with $\times\div$ or with $\div\times$ yields that after the tn steps $d_{i+t+1} \leq \lambda^t d_i$ w. o. p. for a constant $\lambda \in (0, 1)$. □

Finally, the theorem just proved together with Theorem 1 yield the bound on the expected runtime, the expected number of steps the (1+1) ES needs to realize a predefined reduction of the distance from the optimum in the search space.

[2] For the set W of words W^* denotes the Kleene closure.

Theorem 3. *Let the (1+1) ES minimize* SPHERE *using Gauss-mutations and the 1/5-rule described in this section. If for the initial search point $\boldsymbol{a} \in \mathbb{R}^n$ and the initial scaling factor s_1, $|\boldsymbol{a}|/s_1 = \Theta(n)$ then the expected number of steps to obtain a search point $\boldsymbol{c}$ such that $|\boldsymbol{c}| \leq |\boldsymbol{a}| \cdot 2^{-t}$ for $t \in poly(n)$ is in $\Theta(t \cdot n)$.*

Proof. The lower bound $\Omega(t \cdot n)$ follows immediately from Theorem 1. The assumption on the starting values ensure that $l_1 = \Theta(\sqrt{n}) \cdot s_1 = \Theta(d_1/\sqrt{n})$ (see Appendix D for the expected length $\Theta(\sqrt{n})$ of the unscaled mutation vector). In other words, the expected length of the mutation vector has optimal order (cf. Lemma 2). In particular, this assumption on the starting conditions imply the starting conditions used in the proof of Theorem 2 (which have been ensured there by the assumption that the sequence of phases starts with $\times\div$ or $\div\times$). Hence, Theorem 2 yields that the number of phases such that the expected distance from the optimum is smaller than $|\boldsymbol{a}| \cdot 2^{-t}/2$ is in $O(t)$. By Markov's inequality, $\mathsf{P}\{|\boldsymbol{c}| \leq |\boldsymbol{a}| \cdot 2^{-t}\} \geq 1/2$ after these $O(t)$ phases. If this is not the case, the distance from the optimum is not greater than $|\boldsymbol{a}|$ such that again with probability at least $1/2$, $|\boldsymbol{c}| \leq |\boldsymbol{a}| \cdot 2^{-t}$ 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 in $O(t \cdot n)$. □

For different 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 utilizing the results that have been presented.

- Rechenberg (1973, 1994); Beyer (2001); Arfken (1990); Motwani and Raghavan (1995) -

Bibliography

Arfken, G. B. (1990). *Mathematical Methods for Physicists*. Academic Press, San Diego, CA, 3rd edition.

Beyer, H.-G. (2001). *The theory of evolution strategies*. Springer, Berlin.

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

Graham, R. L., Knuth, D. E., and Patashnik, O. (1989). *Concrete Mathematics: A Foundation for Computer Science*. Addison-Wesley, Reading, MA.

Jansen, T. and Wegener, I. (2001). Real royal road functions – where crossover provably is essential. *Proc. of the 3rd Genetic and Evolutionary Computation Conference (GECCO 2001)*, 375–382.

Jansen, T. and Wegener, I. (2002). The analysis of evolutionary algorithms – a proof that crossover really can help. Algorithmica 34, 47–66.

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

Motwani, R. and Raghavan, P. (1995). *Randomized Algorithms*. Cambridge University Press.

Mühlenbein, H. (1992). How genetic algorithmis really work. Mutation and hillclimbing. R. Männer and R. Manderick, editors, *Proc. of the 2nd Parallel Problem Solving from Nature (PPSN II)*. North-Holland, Amsterdam, The Netherlands, 15–25.

Rechenberg, I. (1973). *Evolutionsstrategie*. Frommann-Holzboog, Stuttgart, Germany.

Rechenberg, I. (1994). *Evolutionsstrategie '94*. Frommann-Holzboog, Stuttgart, Germany.

Wegener, I. (2001). Theoretical aspects of evolutionary algorithms. *Proc. of the 28th Int'l Colloquium on Automata, Languages and Programming (ICALP),* LNCS 2076, 64–78.

Wegener, I. and Witt, C. (2002). On the analysis of a simple evolutionary algorithm on quadratic pseudo-boolean functions. Journal of Discrete Algorithms. To appear.

A Hypersurface area of a hyper-spherical cap

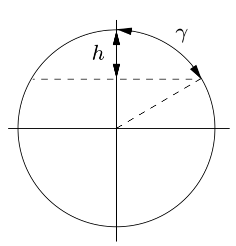
In the following, polar/spherical coordinates are used. Therefore, let r denote the distance from the origin, α the azimuthal angle with range $[0, 2\pi)$ and $\beta_3, \ldots, \beta_n$ the remaining angles with range $[0, \pi]$. The connection to a Cartesian orthonormal system is the following. For some given $\boldsymbol{x} \in \mathbb{R}^n$ let $\boldsymbol{x}'$ denote its orthogonal projection onto the x_1-x_2-plane. If $\boldsymbol{x}'$ equals the origin, $\alpha := 0$, otherwise α is the angle between (the positive half of) the x_1-axis and the line segment $\overline{\boldsymbol{o}\,\boldsymbol{x}'}$ (which lies inside the x_1-x_2-plane). Furthermore, for $i \in \{3, \ldots, n\}$, β_i is the angle between (the positive half of) the x_i-axis and the line segment $\overline{\boldsymbol{o}\,\boldsymbol{x}}$. As a consequence, representation by spherical coordinates is well-defined.

Let ρ denote an arbitrary permutation on $\{3, \ldots, n\}$. Fixing r in n-space, but none of the angles defines a point set $S_r^{(n)}$ forming an n-sphere 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. Kendall (1961)).

Thus, the hypersurface area of an n-sphere with radius $r \geq 0$ equals

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

Each of the $n-1$ factors in parentheses corresponds (one-to-one) to one dimension of the infinitesimal hypersurface element at the point $(r, \alpha, \beta_3, \ldots, \beta_n)$, which is illustratively generated by simultaneously changing all $n-1$ angles by $d\alpha, d\beta_3, \ldots, d\beta_n$, respectively. Re-grouping the factors, solving the α-integral $\left(\int_0^{2\pi} d\alpha = 2\pi\right)$ and defining $\Psi_i(\gamma) := \int_0^{\gamma} (\sin \beta)^i \, d\beta$ yields $r^{n-1} \cdot 2\pi \cdot \prod_{i=1}^{n-2} \Psi_i(\pi)$ for the area of an n-sphere with radius r.



The area of an n-dimensional spherical cap is calculated by adjusting β_n's upper limit appropriately. In the figure on the right, the interdependence between the upper limit (γ) on β_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_n-plane if $\alpha = 0$). In the unit circle $h = 1 - \cos \gamma$ for $\gamma \in [0, \pi]$. Thus, the area of a hyper-spherical cap with radius r and height $r \cdot (1 - \cos \gamma) \in [0, 2r]$ equals $r^{n-1} \cdot 2\pi \cdot \left(\prod_{i=1}^{n-3} \Psi_i(\pi)\right) \cdot \Psi_{n-2}(\gamma)$.

Consequently, the ratio of the hypersurface area of a hyper-spherical cap to the hypersurface area of the hyper-sphere it is cut off equals $\Psi_{n-2}(\gamma)/\Psi_{n-2}(\pi)$ where the angle $\gamma = \arccos(1 - h/r) \in [0, \pi]$ corresponds to the cap's height.

B Tight bounds for $\Psi_k(\pi) = \int_0^\pi (\sin x)^k\, dx$

By the definition of the beta function (cf. Arfken (1990)), namely

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

for $k \in \mathbb{N}$

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

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

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

Furthermore, using the given answer to exercise 9.60 in [Graham, Knuth, and Patashnik (1989)], namely

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

for $k \geq 2$

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

Altogether, for $k \geq 2$

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

and therefore, $\Psi_k(\pi) = \Theta(1/\sqrt{k})$.

C Maximizing w_k

The derivative of

$$w_k(x) \quad = \quad \left(1 - \frac{1}{4x^2}\right)^{(k+1)/2} \cdot \frac{1}{x}$$

equals

$$\begin{aligned} w_k'(x) \quad &= \quad \frac{-1}{x^2}\left(1 - \frac{1}{4x^2}\right)^{(k-1)/2} + \frac{1}{x}\left(\frac{1}{2x^3}\,\frac{k-1}{2}\left(1 - \frac{1}{4x^2}\right)^{(k-1)/2-1}\right) \\ &= \quad \frac{-1}{x^2}\left(1 - \frac{1}{4x^2}\right)^{(k-1)/2-1}\left(\left(1 - \frac{1}{4x^2}\right)^1 - \frac{k-1}{4x^2}\right)\,, \end{aligned}$$

and consequently, if $x > 0$,

$$w_k'(x) \;=\; 0 \quad \Leftrightarrow \quad \left(1 - \frac{1}{4x^2}\right) - \frac{k-1}{4x^2} \;=\; 0 \quad \Leftrightarrow \quad x \;=\; \frac{\sqrt{k}}{2}\,.$$

Since $w_k(x) > 0$ for $x > 0$ and $\lim w_k(x) = 0$ as $x \to \infty$, the unique extremum of the function w_k at $\sqrt{k}/2$ is in fact a maximum.

D Proof of Lemma 6

Claim. "A Gauss-mutation $\boldsymbol{m} \in \mathbb{R}^n$ is isotropically distributed, and moreover, $l := \mathsf{E}[|\boldsymbol{m}|]$ exists and $\mathsf{P}\{|\,|\boldsymbol{m}| - l\,| \geq \delta \cdot l\} \leq \delta^{-2}/(n-1/2)$.

If $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_n$ denote independent copies of $\boldsymbol{m}$, then for any constant $\lambda < 1$, two positive constants $a(\lambda)$ and $b(\lambda)$ exist such that for the cardinality of $I := \{i \mid a(\lambda) \cdot l \leq |\boldsymbol{m}_i| \leq b(\lambda) \cdot l\}$ w. o. p. $\#I \geq \lambda n$."

Proof. Let $\widetilde{\boldsymbol{m}}$ be $(N_1(0,1), \ldots, N_n(0,1))$-distributed (cf. Definition 2). Then $|\widetilde{\boldsymbol{m}}|$ is χ-distributed (with n degrees of freedom; cf. Arfken (1990)), 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,

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

It can be shown that $\mathsf{Var}[|\widetilde{\boldsymbol{m}}|]$ approaches $1/2$ from below as $n \to \infty$, and thus, $\mathsf{Var}[|\widetilde{\boldsymbol{m}}|] \leq 1/2$ and $\mathsf{E}[|\widetilde{\boldsymbol{m}}|]^2 \geq (n-1/2)/2$.

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

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

for any $\delta > 0$. Since $\mathsf{E}[|\boldsymbol{m}|] = \lambda \cdot \mathsf{E}[|\widetilde{\boldsymbol{m}}|]$ and $\mathsf{Var}[|\boldsymbol{m}|] = \lambda^2 \cdot \mathsf{Var}[|\widetilde{\boldsymbol{m}}|]$ for $\lambda \in \mathbb{R}^+$, applying this bound to $|\boldsymbol{m}|$ yields (as $l = \mathsf{E}[|m|]$)

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

Furthermore, it can be shown that $\boldsymbol{x}$ remains $(N_1(0,1), \ldots, N_n(0,1))$-distributed when switching to another Cartesian coordinate system by an arbitrary orthogonal transformation (cf. Beyer (2001)). In other words, $\boldsymbol{m}$'s distribution is invariant with respect to the rotation of the coordinate system; this implies the isotropy of $\boldsymbol{m}$'s distribution.

Finally, the situation of n iid Gauss-mutations: Since $|\boldsymbol{m}| \neq \Theta(\mathsf{E}[|\boldsymbol{m}|])$ only with probability $O(1/n)$, $\mathsf{E}[\#I] = n - O(1)$ and Chernoff-bounds can be applied. □