Probabilistic Runtime Analysis of $(1\overset{+}{,}\lambda)$ ES Using Isotropic Mutations

Jens Jägersküpper*
Dortmund University, Informatik 2, 44221 Dortmund, Germany
JJ@Ls2.cs.uni-dortmund.de

ABSTRACT

We consider the $(1+\lambda)$ ES and the $(1,\lambda)$ ES, which are simple evolutionary algorithms for minimization in $\mathbb{R}^n$, using isotropic mutations. General lower bounds on the number of mutations that are necessary to reduce the approximation error in the search space, i. e. the distance from the optimum (or from any other fixed point in the search space), are proved. Therefore, we generalize a lower-bound method recently introduced by Witt in a runtime analysis of the $(\mu+1)$ EA for the search space $\{0,1\}^n$, which was also already successfully applied in an analysis of a $(\mu+1)$ ES. Namely, we prove that both, the $(1+\lambda)$ ES as well as the $(1,\lambda)$ ES need $\Omega(n \cdot \lambda/\ln\lambda)$ function evaluations with an overwhelming probability to halve the approximation error in the search space – independently of how the isotropic mutations are adapted and of the function to be optimized.

On the other hand, for an upper bound we consider the following concrete scenario: the minimization of the well-known SPHERE-function using Gaussian mutation vectors adapted by the 1/5-rule. We prove that the $(1+\lambda)$ ES needs $O(n \cdot \lambda/\sqrt{\ln\lambda})$ SPHERE-evaluations with an overwhelming probability to halve the approximation error. Moreover, by some kind of reduction, we show that this upper bound also holds for the $(1,\lambda)$ ES.

Finally, the gap of size $O(\sqrt{\ln\lambda})$ between the lower bound and the upper bound is discussed.

Categories and Subject Descriptors

F.2 [**Analysis of Algorithms and Problem Complexity**]: Runtime Analysis; G.3 [**Probability and Statistics**]: Probabilistic Algorithms; G.1.6 [**Optimization**]: Sphere Function; I.2.8 [**Problem Solving, Control Methods, and Search**]: Evolution Strategies

*supported by the German Research Foundation (DFG) through the collaborative research center "Computational Intelligence" (SFB 531)

GECCO'06, July 8–12, 2006, Seattle, Washington, USA.

General Terms

Theory, Algorithms, Performance

Keywords

Runtime Analysis, Evolution Strategies, Sphere Function

1. INTRODUCTION

$(1\overset{+}{,}\lambda)$ evolutionary algorithms (EAs) are simple evolutionary algorithms. The "1" indicates that a (parent) population of size 1 is used; thus, cross-over is not possible and mutation is the only evolutionary search operator that is used to evolve a search point $\boldsymbol{c}$. The "λ" indicates that, in a single step, λ offspring of the current search point $\boldsymbol{c}$ are generated by λ independent mutations. When using *elitist selection* (indicated by the "+"), the best of these λ mutants replaces/becomes the current individual $\boldsymbol{c}$ if (and only if) it is at least as good as its parent. When *comma selection* is used (indicated by the ","), the best of the λ mutants replaces/becomes the current individual $\boldsymbol{c}$ irrespective of whether it is worse than its parent or not. We have just described a single step; this step is repeated in the so-called *evolution loop* until a stopping criterion is met. Fortunately, for the results we are aiming at here, we need not define a stopping criterion, yet consider an infinite evolution loop.

Evolutionary algorithms for optimization in the continuous search space $\mathbb{R}^n$, however, are commonly subsumed under the term *evolution(ary) strategies (ESs)* which was coined by Rechenberg and Schwefel; cf. Rechenberg (1973) and Schwefel (1995). Probabilistic analyses of the runtime of ESs like those by Jägersküpper (2003, 2005) for the (1+1) ES using Gaussian mutations adapted by the 1/5-rule succeeded only recently. Therein, unimodal functions, essentially the well-known SPHERE-function and positive definite quadratic forms, are considered. However, an almost uncountable number of experimental results exist. In attempting to find an explanation for the experimental findings obtained for search space dimensions usually ranging from 10 to 30, some works additionally present some calculations for the 1-dimensional search space $\mathbb{R}$. Unfortunately, such calculations cannot help us with an analysis of how the runtime grows with the dimension of the search space, i. e., with explaining the effect which an increasing search space dimensionality has on the performance of an ES on a given class of functions.

In the considerably developed theory on local performance measures (*progress rate, quality gain;* cf. Beyer (2001)), the progress which a single step yields (after the single-step gain has become steady-state) is analyzed, i. e., the optimization

process is (implicitly) assumed to stabilize w. r. t. the one-step progress. Moreover, the analytical challenges that the inherent randomness of the search process bears are circumvented by looking at a simplifying, in particular (more) deterministic model of the process – raising the need for experimental validation of the prediction quality of the behavior-describing equations (obtained for/in the simplified model).

In the present paper, however, rigorously tackling the randomness is the focus of the analysis.

Since the very first successful applications of EAs, one has desired to explain the principles of EAs from a theoretical point of view (e. g. the Schema Theorem by Holland (1975) and Rechenberg's (1973) investigations of the (1+1) ES). In the last decade, there has been a growing interest in theoretical runtime analysis of specific EAs on specific fitness functions and classes of functions, which is sometimes (somewhat confusingly) called "computational time complexity" of EAs. In this framework, it is examined by means of mathematical proofs how many evaluations of the fitness function are performed until the EA finds a global optimum. The hope is to identify practically relevant classes of functions where the EA behaves efficiently, i. e., where on average it takes only a small (polynomial) number of evaluations.

Probabilistic runtime analyses started with simple EAs, such as the (1+1) EA, on simple functions like OneMax, e. g. Droste et al. (2002). Nowadays, one is able to analyze the (expected) runtime of the (1+1) EA on practically relevant problems such as the maximum matching problem (Giel and Wegener, 2003), the minimum spanning tree problem (Neumann and Wegener, 2004), and simple scheduling problems (Witt, 2005). It has turned out that from our complexity-theoretical perspective, the (1+1) EA is surprisingly efficient on such problems.

Despite the successful analyses of the (1+1) EA, runtime analyses should also explain the utility of the ingredients of more complex EAs, e. g., populations and variation operators. Studies of the utility of populations were performed, amongst others, for a (1+λ) EA (Jansen and De Jong, 2002) and a (μ+1) EA (Witt, 2004). The impact of a crossover operator was also investigated (Storch and Wegener, 2003). All of these studies of more complex EAs, however, have been confined to discrete search spaces, more precisely, to pseudo-Boolean functions $f\colon \{0,1\}^n \to \mathbb{R}$. With respect to continuous search spaces, i. e. $f\colon \mathbb{R}^n \to \mathbb{R}$, most of the studies are experimental (as we already discussed above). As the probabilistic analyses of the simple (1+1) ES have reached a certain extend, a study of more complex EAs for continuous search spaces should again start with such simple functions like Sphere. Jägersküpper and Witt (2005) successfully followed this approach for an analysis of a (μ+1) ES.

In the following section, we look at the algorithm under consideration and at fitness-landscapes that are covered. In Section 3 some notions and results that are used are recapitulated. Subsequently, we present the lower-bound result in Section 4 and the upper-bound result in Section 5. Finally, we discuss those results and conclude in the last section.

2. ALGORITHM, FITNESS LANDSCAPE

As mentioned above, we consider the well-known Sphere function defined by $\text{Sphere}(\boldsymbol{x}) := \sum_{i=1}^{n} x_i^2 = |\boldsymbol{x}|^2 = \boldsymbol{x}^\top \boldsymbol{I}\boldsymbol{x}$, where $\boldsymbol{I}$ denotes the identity matrix and $|\boldsymbol{x}|$ the L^2-norm of the vector $\boldsymbol{x}$, i. e. its length in Euclidean space. It can easily be seen that the results are valid also for translations of this function, i. e., $f(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{x}^*)^\top \boldsymbol{I}(\boldsymbol{x}-\boldsymbol{x}^*)$ for some fixed minimum search point $\boldsymbol{x}^* \in \mathbb{R}^n$. Since we concentrate on the approximation error in the search space (which is defined as the distance from the optimum), the results are in fact valid for any unimodal function satisfying $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n\colon |\boldsymbol{x}-\boldsymbol{x}^*| < |\boldsymbol{y}-\boldsymbol{x}^*| \implies f(\boldsymbol{x}) < f(\boldsymbol{y})$, where $\boldsymbol{x}^* \in \mathbb{R}^n$ is the unique minimum. Note: For Sphere, reducing the approximation error in the search space by an α-fraction ($\alpha \in [0,1]$; for example $\alpha = 0.5$ for halving the distance) corresponds to a reduction of the Sphere-value by a $(2\alpha - \alpha^2)$-fraction (0.75 in the example so that the Sphere-value is reduced to 25 %).

As mentioned in the introduction, we will consider the 1/5-rule for the adaptation of the mutation strength, which was introduced by Rechenberg in the 1960s for the (1+1) ES. (It is a priori not clear how much sense this adaptation makes for (1$^+_,\lambda$) ESs. We will discuss this later.) The idea behind the 1/5-rule is that an isotropic mutation should result in an improvement with a probability of roughly 1/5. Therefore, the optimization is observed for $\Theta(n)$ steps. After each observation phase, the scaling factor σ for the adaptation of the length of the mutation vector (hereinafter called mutation strength) is decreased if less than 1/5 of the mutations in the respective observation phase have been successful, and otherwise, it is increased. Namely, σ is multiplied by a positive constant smaller resp. greater than 1. To keep the proofs as simple as possible, here the observation phase will last n steps and σ will be halved resp. doubled.

Commonly, in the ($\mu^+_,\lambda$)ES framework each individual consists of a search point and an associated mutation strength. As we have $\mu = 1$ and concentrate on 1/5-rule-like adaptation mechanisms, the mutation strength σ is not endogenous here. We use two counters, "g" and "b", to remember the number of "good" resp. "bad" mutations.

The (1+λ) ES for minimization of $f\colon \mathbb{R}^n \to \mathbb{R}$ we consider works as follows – when using Gaussian mutations adapted by the 1/5-rule: For a given initialization of the evolving search point $\boldsymbol{c} \in \mathbb{R}^n$ and the mutation strength $\sigma \in \mathbb{R}_{>0}$, an evolution loop is performed:

1. FOR (int $i := 1$; $i \le \lambda$; i++) DO
   (a) Create a new search point $\boldsymbol{y}^{[i]} := \boldsymbol{c} + \boldsymbol{m}$ with $\boldsymbol{m} := \sigma \cdot \widetilde{\boldsymbol{m}}$, where each component of $\widetilde{\boldsymbol{m}} \in \mathbb{R}^n$ is independently standard-normally distributed.
   (b) IF $f(\boldsymbol{y}^{[i]}) \le f(\boldsymbol{c})$ THEN $g := g + 1$ ELSE $b := b + 1$.
2. IF $b + g \ge n$ THEN
   (a) IF $g < (g+b)\cdot(1/5)$ THEN $\sigma := \sigma/2$ ELSE $\sigma := \sigma \cdot 2$.
   (b) $g := 0$. $b := 0$.
3. IF $\min_{i\in\{1,\dots,\lambda\}}\{f(\boldsymbol{y}^{[i]})\} \le f(\boldsymbol{c})$ THEN $\boldsymbol{c} := \operatorname{argmin}_{i\in\{1,\dots,\lambda\}}\{f(\boldsymbol{y}^{[i]})\}$.
4. GOTO 1.

It's understood that $\lambda = \text{poly}(n)$. In practice, obviously, the GOTO is conditioned on a stopping criterion. Fortunately, for the results we are aiming at, we need not define a reasonable stopping criterion. Rather we will consider a run of a (1$^+_,\lambda$) ES as an infinite stochastic process. We are interested in how fast $\boldsymbol{c}$ evolves.

We obtain the (1,λ) ES by dropping the IF-condition in instruction (3), implying that the best of the λ offspring always replaces (and becomes) the current search point. Unlike the (1+λ) ES, the (1,λ) ES may accept mutations that result in a worse (w. r. t. the function value) search point. (Obviously, a (1,1) ES results in pure random search and, thus, does not make much sense.)

Our choice (in the instructions (2)) to adapt σ after a step in which the nth mutation (after the previous adaptation) took place, i. e. after $\lceil n/\lambda \rceil$ iterations/steps, was somehow arbitrary. The number of steps between two subsequent σ-adaptations must merely be $\Theta(n/\lambda)$ if $\lambda = O(n)$, and/or $\Theta(1)$ otherwise. Furthermore, the constants for the σ-adaptation, which we chose to be 1/2 resp. 2 for notational convenience, can be chosen arbitrarily (strictly smaller resp. larger than 1, of course). Any such σ-adaptation is covered by the term "1/5-rule" here. (In fact, we may even consider a corresponding "1/4-rule" or a "1/6-rule.)

The lower bound on the runtime, which we define as the number of function evaluations, however, will be valid independently of the adaptation of σ. In fact, it will even be independent of the distribution of $|\widetilde{\boldsymbol{m}}|$. For instance, $|\widetilde{\boldsymbol{m}}|$ could be distributed according to a Cauchy distribution, rather than according to a χ-distribution (with n degrees of freedom) as the length of a Gaussian mutation. Moreover, the two σ-scaling-factors may even depend on n. The only restriction – besides the isotropy of $\widetilde{\boldsymbol{m}}$ – is that after each step, the (1$\dagger$,λ) ES* (as we will call this more general ES template) decides whether to keep σ unchanged, or to up-scale, or to down-scale the mutation strength.

For the Sphere scenario which is considered here this means that we are interested in how fast (number of function evaluations w. r. t. n, the dimensionality of the search space) the distance from the optimum/origin $\boldsymbol{o}$ is reduced. Note that for Sphere (and all other functions for which every plateau of constant fitness has zero n-volume), the function value of a mutant generated in instruction (1.a) differs with probability 1 from every function value seen before. As a consequence, there will always be exactly one best mutant in instruction (3). As another consequence, we obtain that on the Sphere function, the (1+λ) ES with $\lambda = 1$ resembles the (1+1) ES as considered by Jägersküpper (2003).

3. PRELIMINARIES

We say that a statement holds "for n large enough" if it is true for all $n \geq n'$, where n' is an absolute constant. Recall the following asymptotics, where $g, h\colon \mathbb{N} \to \mathbb{R}$ are positive for n large enough:
$g(n) = O(h(n))$ iff there exists a positive constant κ such that $g(n) \leq \kappa \cdot h(n)$ for n large enough,
$g(n) = \Omega(h(n))$ iff $h(n) = O(g(n))$,
$g(n) = \Theta(h(n))$ iff $g(n)$ is $O(h(n))$ as well as $\Omega(h(n))$,
$g(n) = \text{poly}(n)$ iff $g(n) = O(n^k)$ for a constant k.

As we are interested in how the runtime depends on n, the dimensionality of the search space, all asymptotics are w. r. t. to this parameter (unless stated differently).

A probability is exponentially small (in n) if it is upper bounded by $\exp(-\Omega(n^\varepsilon))$ for a constant $\varepsilon > 0$. An event $\mathcal{E}(n)$ happens with an overwhelming probability (w. o. p.) if $1 - \mathsf{P}\{\mathcal{E}(n)\}$ is exponentially small.

As mentioned above, we will consider isotropic mutations for $\mathbb{R}^n$. For $\boldsymbol{m} \in \mathbb{R}^n$ let $|\boldsymbol{m}|$ denote $\boldsymbol{m}$'s Euclidean length, i. e. its L^2-norm.

Definition 1. *A given probability distribution over $\mathbb{R}^n$ is isotropic if (and only if) $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n\colon |\boldsymbol{x}| = |\boldsymbol{y}| \Rightarrow \mathsf{D}(\boldsymbol{x}) = \mathsf{D}(\boldsymbol{y})$, where "$\mathsf{D}(\boldsymbol{x})$" denotes the distribution's density at $\boldsymbol{x}$.*

This implies two very useful properties for an isotropic mutation $\boldsymbol{m}$:

- the normalized vector $\boldsymbol{m}/|\boldsymbol{m}|$ is uniformly distributed upon the unit hyper-sphere $\{\boldsymbol{x} \in \mathbb{R}^n \mid |\boldsymbol{x}| = 1\}$ and
- the random length $|\boldsymbol{m}|$ is independent of the random direction $\boldsymbol{m}/|\boldsymbol{m}|$.

Each component of a Gaussian mutation vector $\widetilde{\boldsymbol{m}} \in \mathbb{R}^n$ is independently standard-normally distributed. Thus, the density at $\boldsymbol{x} \in \mathbb{R}^n$ equals

$$\prod_{i=1}^{n} \frac{\exp(-{x_i}^2/2)}{\sqrt{2\pi}} = \frac{\exp(\sum_{i=1}^{n} -{x_i}^2/2)}{\sqrt{2\pi}} = \frac{\exp(-|\boldsymbol{x}|^2/2)}{\sqrt{2\pi}},$$

implying that a Gaussian mutation is in fact isotropic. It is easy to see that a scaled Gaussian mutation $\sigma \cdot \widetilde{\boldsymbol{m}}$ with $\sigma > 0$ is also isotropically distributed. Scaled Gaussian mutations are commonly used, for instance within Rechenberg's 1/5-rule or Schwefel's σ-self-adaptation.

We should note, however, that within the more sophisticated *covariance matrix adaptation (CMA)* due to Hansen and Ostermeier (1996), $\sigma \cdot \mathbf{B} \cdot \widetilde{\boldsymbol{m}}$ makes up the mutation vector with a matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$ which is also adapted. Thus, unless $\mathbf{B} = s \cdot \mathbf{I}$ for some scalar s, the mutation vector is not isotropically distributed. Consequently, CMA is not covered by the results presented here (at least not in general).

However, the crucial property of an isotropic mutation is that we can first choose a random direction and subsequently sample a point on the already chosen half line according to the distribution of $|\boldsymbol{m}|$; or we can (and will) do it vice versa: First we sample the mutation's length ℓ according to the random variable $|\boldsymbol{m}|$; subsequently, the mutant is uniformly distributed upon the hyper-sphere with radius ℓ centered at the mutated search point.

Let G_ℓ denote the spatial gain of an isotropically distributed mutation vector $\boldsymbol{m}$ parallel to an arbitrarily fixed direction, given that (the random variable) $|\boldsymbol{m}|$ takes the value $\ell > 0$; note that $G_\ell \in [-\ell, \ell]$. In other words, when $\boldsymbol{c} \in \mathbb{R}^n$ is mutated and $\boldsymbol{c}' := \boldsymbol{c} + \boldsymbol{m}$ is the mutant, then the absolute value of G_ℓ equals the distance between the two hyper-planes containing $\boldsymbol{c}$ resp. $\boldsymbol{c}'$ that are perpendicular to the predefined direction, respectively. Jägersküpper (2003) shows that for $n \geq 4$ the density of the random variable G_ℓ at $g \in [-\ell, \ell]$ equals

$$(1/\ell) \cdot \left(1 - (g/\ell)^2\right)^{(n-3)/2} / \Psi, \tag{1}$$

where $\mathsf{P}\{G_\ell \in [-\ell, \ell]\} = 1$, i. e. $\Psi := \int_{-1}^{1} (1 - x^2)^{(n-3)/2}\,\mathrm{d}x$ lies in the interval $\sqrt{2\pi}/\sqrt{n - [1.5 \pm 0.5]}$ (normalization).

With this density function, the probability of $\boldsymbol{c}'$ being closer to some fixed point in the search space than $\boldsymbol{c}$ has been estimated. For notational convenience, we consider w. l. o. g. the distance from the origin $\boldsymbol{o}$ so that we can use $|\boldsymbol{c}|$ for this distance. Obviously, if $\ell > 2|\boldsymbol{c}|$ then $|\boldsymbol{c}'| > |\boldsymbol{c}|$; if $\ell \leq 2|\boldsymbol{c}|$, however, $|\boldsymbol{c}'| \leq |\boldsymbol{c}|$ if and only if the spatial gain parallel to $\overrightarrow{\boldsymbol{c}\,\boldsymbol{o}}$ is at least $\ell^2/(2|\boldsymbol{c}|)$ (cf. Jägersküpper (2003)) Thus, for instance, the (conditional) probability (given $|\boldsymbol{m}| = \ell$) that

the mutant is at least as close to the optimum/origin as its parent equals

$$\begin{aligned} & \mathsf{P}\big\{|\boldsymbol{c}'| \le |\boldsymbol{c}| \text{ given that } |\boldsymbol{m}| = \ell\big\} \\ = \quad & \frac{1}{\Psi} \cdot \int_{\min\{1\,,\,\ell/(2|\boldsymbol{c}|)\}}^{1} \big(1-x^2\big)^{(n-3)/2}\,\mathrm{d}x \qquad (2) \end{aligned}$$

(since $|\boldsymbol{c}'| \le |\boldsymbol{c}|$ is impossible if $\ell > 2|\boldsymbol{c}|$). The density of G_ℓ will be considered again and in more detail in Section 5.

4. LOWER BOUND

As mentioned before, we want to obtain a general lower bound on the number of steps to reduce the distance from a fixed point in the search space $\mathbb{R}^n$, for instance from the (or a fixed) optimum of the function to be optimized. In particular, we want to know how this number depends, on the one hand, on n (the search space's dimensionality) and on the offspring-population size λ, on the other hand.

The idea behind this bound is the "curse of dimensionality" in $\mathbb{R}^n$. Therefore, first consider the search space $\{0,1\}^n$ and the standard mutation operator (namely, each bit is independently flipped with probability $1/n$). When we repeatedly mutate a search point without doing selection, then each point in the search space is hit infinitely often as the number of mutations approaches infinity.[1] In particular, the number of steps it takes the random search to visit a certain point is finite. Now consider $\mathbb{R}^n$ for $n \ge 3$. Lets start with a fixed point and repeatedly add an isotropically distributed vector (with an arbitrary distribution of the length that might be degenerate, but not concentrated at 0) to this point. Despite the fact that our starting point is never exactly hit again, even the probability to get close to our starting point tends to zero as the dimensionality increases, even if the number of mutations approaches infinity.

Obviously, the search of a $(1^+\!,\lambda)\,\mathrm{ES}$ is not purely random, yet guided by selection – unless a flat fitness landscape is given. Selection, however, merely means that search paths that do not seem promising are no longer followed (pruned). One may imagine that also these search paths would be followed – in addition to the promising ones, of course.

In the following we modify the $(1^+\!,\lambda)\,\mathrm{ES}^*$ such that we end up with a search procedure that is independent of the function to be optimized and, thus, purely random: Consider the $(1^+\!,\lambda)\,\mathrm{ES}^*$ after initialization, i. e., an initial starting point and an initial mutation strength are given. In the first step, λ mutants are generated by respectively adding $\sigma \cdot \widetilde{\boldsymbol{m}}$ to the starting point. In contrast to the original $(1^+\!,\lambda)\,\mathrm{ES}^*$, we now do *not* select one of the $\lambda(+1)$ individuals, yet keep all $1+\lambda$ search points as a population P_1. After the first step, σ may be up- or down-scaled – depending on the individuals' function values. Thus, to also get rid of this function-dependency, each of the $1+\lambda$ points in P_1 is mutated 3 times: once without changing σ, once with an up-scaled σ, and once with a down-scaled mutation strength. Again we keep all $(1+\lambda)\cdot 3\lambda$ newly generated individuals. Consequently, we have $(1+\lambda) + (1+\lambda)\cdot 3\lambda = (1+\lambda)(1+3\lambda)$ individuals after the second step in the population P_2. Repeating this procedure, after i iterations a population P_i is generated which contains

$$(1+\lambda)(1+3\lambda)^{i-1} \le (1+3\lambda)^i = \exp(\ln(1+3\lambda)\cdot i)$$

individuals. The crucial point is that P_i is built without any dependency on the function to be optimized, and that all search paths of the original $(1^+\!,\lambda)\,\mathrm{ES}^*$ emerge in this modified search procedure with the same probability density: Let $S \subset \mathbb{R}^n$ denote some (measurable) set. Then, the probability that P_i hits S, namely $\mathsf{P}\{S \cap P_i \ne \emptyset\}$, is an upper bound on the probability that the search point evolved within i iterations of the original $(1^+\!,\lambda)\,\mathrm{ES}^*$ is in S.

Now, if we knew that the probability that an individual in P_i hits S is very small, say, upper bounded by $\mathrm{e}^{-\xi n}$ for some $\xi > 0$, then the probability that P_i contains at least one point from S would be bounded above by

$$\#P_i \cdot \mathrm{e}^{-\xi n} = \frac{\exp(\ln(1+3\lambda)\cdot i)}{\exp(\xi n)} = \exp(\ln(1+3\lambda)\cdot i - \xi n)$$

(note that the probability of a union of events is upper bounded by the sum of the single-event probabilities even, and in particular, if the events are not independent, which is obviously the case here). Then we could choose i such that $\xi n - \ln(1+3\lambda)\cdot i \ge n\cdot\xi/2$, i. e., $i \le (\xi/2)\cdot n/\ln(1+3\lambda)$. We would obtain that $(\xi/2)\cdot n/\ln(1+3\lambda)$ steps suffice only with a probability of at most $\mathrm{e}^{-n\cdot\xi/2}$. In other words, if ξ was $\Omega(1)$, then $\Omega(n/\ln\lambda)$ steps (i. e. $\lambda\cdot\Omega(n/\ln\lambda)$ mutations in the original $(1^+\!,\lambda)\,\mathrm{ES}^*$) would be necessary w. o. p. (namely with probability of $1-\mathrm{e}^{-\Omega(n)}$) to hit S. Obviously, we would finally chose S to be the hyper-ball exactly consisting of all points having at most halve the distance from the optimum as our starting point.

After sketching the proof, we begin filling in the details by noting some rather obvious properties of isotropically distributed vectors.

PROPOSITION 1. *Let $\boldsymbol{x} \in \mathbb{R}^n$ be isotropically distributed and $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ a fixed orthogonal matrix (i. e., $\boldsymbol{M}^\top\boldsymbol{M} = \boldsymbol{I}$). Then the distribution of $\boldsymbol{M}\boldsymbol{x}$ equals the one of $\boldsymbol{x}$.*

PROOF. The multiplication with $\boldsymbol{M}$ corresponds to an ortho*norm*al transformation, and thus, $\boldsymbol{x} \mapsto \boldsymbol{M}\boldsymbol{x}$ is a bijection in $\mathbb{R}^n$ that preserves the inner product, implying that $|\boldsymbol{M}\boldsymbol{x}| = |\boldsymbol{x}|$. Thus, each vector in $\mathbb{R}^n$ has exactly one unique pre-image, and both vectors have the same length. Finally, vectors of the same length have equal density due to the isotropy. □

In other words, an isotropic distribution over $\mathbb{R}^n$ is invariant w. r. t. orthonormal transformations and, in particular, w. r. t. rotations of the orthonormal system.

LEMMA 1. *Let the vectors $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ be independently (not necessarily identically) isotropically distributed. Then $\boldsymbol{z} := \boldsymbol{x} + \boldsymbol{y}$ is also isotropically distributed.*

The proof can be found in the appendix. By induction, we directly obtain

COROLLARY 1. *Let the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k \in \mathbb{R}^n$ be independently (not necessarily identically) isotropically distributed. Then $\boldsymbol{y} := \boldsymbol{x}_1 + \cdots + \boldsymbol{x}_k$ is also isotropically distributed.*

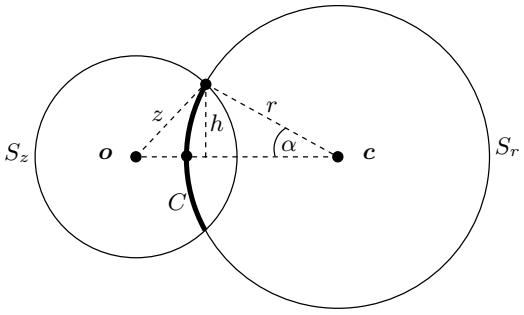
All the individuals in P_i originate from the initial starting point by adding isotropically distributed vectors to it. Due

[1] this is also true, just for instance, for pure (i. e. selectionless) random local search, where in each step one uniformly chosen bit is flipped

to our modifications – for which we pay with an exponential growth of the population – all those vectors are independently distributed. The preceding corollary tells us, that each $\boldsymbol{x} \in P_i$ is isotropically distributed around the starting point. We do not know the distribution of $\boldsymbol{x}$'s distance from the starting point, though. Yet this does not matter as the next lemma will tell us.

LEMMA 2. *Let $\boldsymbol{s} \in \mathbb{R}^n \setminus \{\boldsymbol{o}\}$ be fixed and $\boldsymbol{m} \in \mathbb{R}^n$ be isotropically distributed. Then $\mathsf{P}\{|\boldsymbol{s}+\boldsymbol{m}| \le |\boldsymbol{s}|/2\}$ is at most $\mathrm{e}^{-0.647\,(n-1)}(2\pi n)^{-1/2} \cdot (1+\Theta(n^{-2}))$.*

PROOF. Let $d := |\boldsymbol{s}|$ and $z := d/2$. Furthermore, let $r := |\boldsymbol{m}|$ be the length of $\boldsymbol{m}$. Henceforth, we assume r to take some fixed value maximizing $\mathsf{P}\{|\boldsymbol{s}+\boldsymbol{m}| \le z \mid |\boldsymbol{m}| = r\}$. According to the law of total probability, this leads to an upper bound on $\mathsf{P}\{|\boldsymbol{s}+\boldsymbol{m}| \le z\}$.



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

Let L denote the line segment connecting $\boldsymbol{o}$ to $\boldsymbol{s}$ and let $\boldsymbol{c} := L \cap C$. Then $\boldsymbol{c} \in C$ is the center of the cap C. Note that all points in R, the boundary of the cap, have equal distance from L. Let h denote this distance and let $\alpha := \arcsin(h/r)$. Then $r\alpha$ is the distance of $\boldsymbol{c}$ from R, the boundary of the cap, within the $(n-1)$-space C, namely the spherical distance w. r. t. the hyper-sphere $S_r \supset C$.

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

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

Hence, $\alpha \le 1-\varepsilon$ for some constant $\varepsilon > 0$ will imply that $\mathsf{P}\{|\boldsymbol{s}+\boldsymbol{m}| \le z\} = \mathrm{e}^{-\Omega(n)}$. To show that α is indeed bounded away from 1, we have to estimate h/r. A geometric argument (in the next paragraph) yields $h/r \le z/d$. Since $z/d = 0.5$ and $h, r > 0$, $\arcsin(h/r) \le \arcsin(0.5) < \mathrm{e}^{-0.647}$. This will imply the lemma.

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

We still have to show that the hypersurface area $A(C)$ can be bounded by the volume of an $(n-1)$-dimensional ball of radius $\alpha\, r$. Intuitively, we are confronted with the error that is introduced by mapping the area of a part of a sphere (e. g., the area of a continent) onto a plane (e. g., a map of the continent). Of course, the larger the area compared to the sphere, the greater the error is. Formally, for $\alpha \le \pi/2$, the hypersurface area of C can be expressed as

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

(cf. Appendix B in Ericson and Zinoviev (2001)). Since $\sin\beta \le \beta$ for $\beta \ge 0$, we obtain (by estimating the first integral) that

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

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

This lemma, which formalizes the curse of dimensionality mentioned above, enables us to obtain the lower bound:

THEOREM 1. *Let a $(1^+,\lambda)$ ES* optimize an arbitrary function f, and let $\boldsymbol{y} \in \mathbb{R}^n$ be some fixed point (for instance an optimum). Assume the initial search point has distance $d > 0$ from $\boldsymbol{y}$. Then, for n large enough, the number of f-evaluations until $|\boldsymbol{c}-\boldsymbol{y}| \le d/2^b$, where $b\colon \mathbb{N} \to \mathbb{N}$ such that $b = \mathrm{poly}(n)$, for the first time is w. o. p. larger than $b \cdot \lambda \cdot 0.64n/\ln(1+3\lambda)$, which is $\Omega(b \cdot n \cdot \lambda/\ln\lambda)$ for $\lambda \ge 2$.*

PROOF. First note that $\mathrm{e}^{-0.647\,(n-1)}/\Theta(\sqrt{n}) \le \mathrm{e}^{-0.647\,n}$ for n large enough. Thus, we can choose $\xi := 0.647$ in the reasoning preceding Proposition 1. Rather than splitting ξ into $\xi/2+\xi/2$ as done there, we split 0.647 into $0.64+0.007$, and obtain an upper bound of $\mathrm{e}^{-0.007n}$ on the probability to halve the distance. Finally, we add up the b "failure probabilities" to obtain an upper bound of $b \cdot \mathrm{e}^{-0.007n} = \mathrm{e}^{-\Omega(n)}$ (because $b = \mathrm{poly}(n)$) on the probability that (at least) one of the b halvings needs only $0.64n/\ln(1+3\lambda)$ (or even fewer) steps. □

For SPHERE, reducing the distance by a factor of $1/\sqrt{2}$ corresponds to halving the SPHERE value. Choosing $z := d/\sqrt{2}$ rather than $d/2$ in the proof of Lemma 2 and noting that $\arcsin(1/\sqrt{2}) < \mathrm{e}^{-0.241}$ directly yields:

THEOREM 2. *Let a $(1^+,\lambda)$ ES* minimize* SPHERE *and let $b\colon \mathbb{N} \to \mathbb{N}$ such that $b = \mathrm{poly}(n)$. Then (unless the initial search point is optimal) for n large enough, the number of mutations until the* SPHERE*-value is reduced to a 2^{-b}-fraction (of the initial one) is larger than $b \cdot \lambda \cdot 0.24n/\ln(1+3\lambda)$ (which is $\Omega(b \cdot n \cdot \lambda/\ln\lambda)$ for $\lambda \ge 2$) w. o. p.*

5. UPPER BOUND

Naturally, the upper bound on the runtime crucially depends on the σ-adaptation which is actually used. We consider the 1/5-rule, a deterministic (non-endogenous) adaptation. Jägersküpper (2003) proves that the (1+1) ES using Gaussian mutations adapted by the 1/5-rule is able to keep $\sigma = \Theta(|\boldsymbol{c}|/n)$ for an arbitrary polynomial number of steps. The proof can easily be adapted for the $(1+\lambda)$ ES. Although we cannot repeat the whole proof here due to the page limit, we want to give the main ideas behind the reasoning:

The length of a Gaussian mutation differs from its mean by $\pm 10\%$ only with probability $1 - \Theta(1/n)$. As a consequence, when $\sigma = \Theta(|\boldsymbol{c}|/n)$, which results in an expected spatial gain towards the optimum of optimum order $\Theta(|\boldsymbol{c}|/n)$, we expect all but a constant number of mutations in an observation phase (of n steps) to have a length of actually $\Theta(|\boldsymbol{c}|/\sqrt{n})$. Then, by Chernoff bounds, w. o. p. in more than 90% of the steps in a phase (where σ is kept constant) $|\boldsymbol{m}|$ differs by no more than $\pm 10\%$. Furthermore, for a fixed σ, the success probability of a mutation, i. e. the probability that the mutant is closer to the optimum than its parent, drops as the distance from the optimum gets smaller; cf. Equation (2). Hence, during an observation phase the steps' success probabilities cannot increase. As a consequence, if the success probability in the first step of a phase is small, say 0.1, then we expect at most 10% of the phase's steps to be successful. By Chernoff bounds, less than 20% are actually successful so that σ is halved after the phase – resulting in an increase of the success probabilities in the next phase. Even though the success probability is already small at the beginning of such a phase, each of the steps yields a gain of maximum order $\Theta(|\boldsymbol{c}|/n)$ with probability $\Omega(1)$. Again by Chernoff bounds, the number of steps in the phase each of which actually yields a gain of $\Theta(|\boldsymbol{c}|/n)$ is $\Omega(n)$ w. o. p., so that, finally, the total gain of the phase is w. o. p. a constant fraction of the distance from the optimum at the beginning of the phase.

The treatment of a phase with a large success probability like 0.4 is analog, in fact, even a little simpler since the distance from the optimum cannot increase (recall: elitist selection). All in all, we have that – for Sphere – the 1/5-rule is able to make the right decisions (concerning the scaling of σ). In particular, the progress of a phase cannot become that large (w. o. p. in a polynomial number of steps) that the mutation strength is no longer of optimal order $\Theta(|\boldsymbol{c}|/n)$.

For the $(1+\lambda)$ ES we consider the interesting case $\lambda \geq 2$ and $\lambda = O(n)$ (otherwise σ is adapted in each step anyway). The mutation strength is kept unchanged for $\lceil n/\lambda \rceil$ iterations, i. e., after fewer than $2n$ mutations adaptation takes place. As we have seen in the lower bound on the runtime in Section 4, the progress (between two sequent adaptations) cannot be that large that the 1/5-rule would fail to follow.

Hence, we investigate the (expected) gain of a step of the $(1+\lambda)$ ES next. Therefore, we consider again the random variable G_ℓ, the density of which is given in Formula (1). Let J denote the hyper-plane containing $\boldsymbol{c}$ and being perpendicular to line passing through $\boldsymbol{c}$ and the optimum/origin. Then G_ℓ correspond to the "distance" of $\boldsymbol{c} + \boldsymbol{m}$ from J, where $\boldsymbol{m}$ is isotropically distributed with $|\boldsymbol{m}| = \ell$; note that this "distance" is negative if the mutant lies in the half-space (w. r. t. J) that does not contain the optimum. The analysis of the (1+1) ES utilizes that $\mathsf{P}\{G_\ell \geq \alpha \cdot \ell/\sqrt{n}\} = \Omega(1)$ for any constant α. Since $\ell = \Theta(|\boldsymbol{c}|/\sqrt{n})$ (at least w. o. p. for any polynomial number of steps) due to the 1/5-rule (as we have recapitulated above), each step yields a spatial gain of $\alpha \cdot \Theta(|\boldsymbol{c}|/\sqrt{n})/\sqrt{n}) = \alpha \cdot \Theta(|\boldsymbol{c}|/n)$ with probability $\Omega(1)$.

In the $(1^{+}\lambda)$ ES, however, we consider the best of λ such mutations. Since $(1-\Theta(1/\lambda))^\lambda = \Theta(1)$, we want to know for which $g \in [0, \ell]$ we have $\mathsf{P}\{G_\ell \geq g\} = \Theta(1/\lambda)$ because such a gain g would be realized by the best of the λ mutations with probability $\Omega(1)$. Therefore, we need a better approximation of the integral over G_ℓ's density (Formula (1) in Section 3).

Namely, we have for $\sqrt{n}/3 \geq \beta = \Omega(1)$ and $n \geq 9$

$$\begin{aligned}
&\int_{\beta/\sqrt{n}}^{1} (1-x^2)^{(n-3)/2}\,\mathrm{d}x \\
\geq\ &\int_{\beta/\sqrt{n}}^{2\beta/\sqrt{n}} (1-x^2)^{(n-3)/2}\,\mathrm{d}x \\
>\ &\frac{\beta}{\sqrt{n}} \cdot (1-(2\beta)^2/n)^{(n-3)/2}\,\mathrm{d}x \\
&\text{using } (1-1/k)^{k-1} > \mathrm{e}^{-1} \text{ yields} \\
>\ &\frac{\beta}{\sqrt{n}} \cdot \exp\left(-\frac{(n-3)/2}{n/(2\beta)^2 - 1}\right) \\
=\ &\frac{\beta}{\sqrt{n}} \cdot \exp\left(-2\beta^2 \frac{n-3}{n-4\beta^2}\right) \\
\geq\ &\frac{\beta}{\sqrt{n}} \cdot \exp\left(-4\beta^2\right)
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\int_{\beta/\sqrt{n}}^{1} (1-x^2)^{(n-3)/2}\,\mathrm{d}x \\
\leq\ &\sum_{i=1}^{\lfloor\sqrt{n}/\beta\rfloor} \frac{\beta}{\sqrt{n}} \cdot \left(1-(i\beta/\sqrt{n})^2\right)^{(n-3)/2} \\
=\ &\frac{\beta}{\sqrt{n}} \cdot \sum_{i=1}^{\lfloor\sqrt{n}/\beta\rfloor} \left(1-(i\beta)^2/n\right)^{(n-3)/2} \\
&\text{using } (1-1/k)^{k} < \mathrm{e}^{-1} \text{ yields} \\
<\ &\frac{\beta}{\sqrt{n}} \cdot \sum_{i=1}^{\infty} \exp\left(-\frac{(n-3)/2}{n/(i\beta)^2}\right) \\
\leq\ &\frac{\beta}{\sqrt{n}} \cdot \sum_{i=1}^{\infty} \exp(-(i\beta)^2/3) \\
<\ &\frac{\beta}{\sqrt{n}} \cdot \exp(-\beta^2/3) \cdot \frac{1}{1-\exp(-\beta^2)} \\
=\ &\frac{\beta}{\sqrt{n}} \cdot \exp(-\beta^2/3) \cdot O(1),
\end{aligned}$$

where the last inequality follows because the summands of the series drop by a factor of

$$\frac{\exp(-(i+1)^2\beta^2/3)}{\exp(-\ \ i^2\ \ \beta^2/3)} = \exp(-(2i+1)\beta^2/3) \overset{i\geq 1}{\leq} \exp(-\beta^2).$$

As a consequence, for $\beta = \Omega(1)$ but $\beta \leq \sqrt{n}/3$

$$\int_{\beta/\sqrt{n}}^{1} (1-x^2)^{(n-3)/2}\,\mathrm{d}x = \frac{\beta}{\sqrt{n}} \cdot \mathrm{e}^{-\Theta(\beta^2)}.$$

(Note that the integral's value is obviously bounded by $\mathrm{e}^{-\Omega(n)}$ for $\beta \in [\sqrt{n}/3, \sqrt{n}\,]$.) Thus, for $\sqrt{n}/3 \geq \beta = \Omega(1)$

$$\mathsf{P}\{G_\ell \geq \ell \cdot \beta/\sqrt{n}\} = \frac{\beta}{\sqrt{n}} \cdot \mathrm{e}^{-\Theta(\beta^2)}/\Psi = \beta \cdot \mathrm{e}^{-\Theta(\beta^2)}. \quad (3)$$

Let G_ℓ^λ denote the maximum of λ independent copies of G_ℓ. Since $\beta \cdot \mathrm{e}^{-\Theta(\beta^2)} = \Theta(1/\lambda)$ for $\beta = \Theta(\sqrt{\ln \lambda})$ as well as $1-(1-\Theta(1/\lambda))^\lambda = \Omega(1)$, for any constant α we have $\mathsf{P}\{G_\ell^\lambda \geq \alpha \cdot \ell \cdot \sqrt{\ln \lambda}/\sqrt{n})\} = \Omega(1)$.

Thus – when $\sigma = \Theta(|\boldsymbol{c}|/n)$ as discussed above – a step yields a spatial gain of $\Omega(\sqrt{\ln \lambda} \cdot |\boldsymbol{c}|/n)$ with probability $\Omega(1)$. Recall that this gain was $\Omega(|\boldsymbol{c}|/n)$ for the (1+1) ES. Thus, doing $\lambda \geq 2$ scaled Gaussian mutations instead of one in a single step yields a factor of $\Theta(\sqrt{\ln \lambda})$ for (the lower bound on) the gain of a single step (that is realized with a constant probability) only – compared to the factor of λ in the number of necessary function evaluations. However, exactly – up to the additional $\Theta(\sqrt{\ln \lambda})$-factor – re-doing the analysis presented by Jägersküpper (2003) for the upper bound on the runtime of the (1+1) ES, finally results in the upper bound for the (1+λ) ES:

THEOREM 3. *Let the (1+λ) ES minimize* SPHERE *using Gaussian mutations adapted by the 1/5-rule. Assume that $\sigma = \Theta(|\boldsymbol{c}|/n)$ after initialization. Then the number of steps until the approximation error is reduced to a 2^{-b}-fraction, $b\colon \mathbb{N} \to \mathbb{N}$ s. t. $b = \mathrm{poly}(n)$, is $O(b \cdot n/\sqrt{\ln \lambda})$ w. o. p. (the number of* SPHERE*-evaluations is $O(b \cdot n \cdot \lambda/\sqrt{\ln \lambda})$ w. o. p.).*

So what about the (1,λ) ES? Reconsider the density of G_ℓ (Formula (1) in Section 3): It is symmetric. In other words, for any $g \in [0, \ell]$, we have $\mathsf{P}\{G_\ell \geq g\} = \mathsf{P}\{G_\ell \leq -g\}$; let p denote this probability. For G_ℓ^λ, however, we have $\mathsf{P}\{G_\ell^\lambda \geq g\} = 1-(1-p)^\lambda$ apposed to $\mathsf{P}\{G_\ell^\lambda \leq -g\} = p^\lambda$. Since $1-(1-p)^\lambda \geq 3 \cdot p^\lambda$ for $\lambda \geq 2$ (since $0 \leq p \leq 0.5$, see appendix), a positive spatial gain of at least $g \geq 0$ is at least thrice as probable as a negative gain of at most $-g \leq 0$, for *any* $g \geq 0$. In particular, this implies that $\mathsf{E}[G_\ell^\lambda] \geq \mathsf{E}[G_\ell^\lambda \cdot \mathbb{1}_{\{G_\ell^\lambda \geq 0\}}]/2$, where the indicator variable $\mathbb{1}_{\{G_\ell^\lambda \geq 0\}}$ zeroes out negative gains. Thus, for $\lambda \geq 2$, the difference between elitist and comma selection gets lost in asymptotic notation we use.

However, as mentioned in Section 3, for SPHERE a spatial gain of $\ell^2/(2|\boldsymbol{c}|)$ is necessary for a mutation to be successful. For small λ like 2, the factor 1/2 that we loose (at least in the analysis) by switching from elitists selection to comma selection is crucial. We have to choose λ large enough such that $\mathsf{E}[G_\ell^\lambda] \geq (1+\varepsilon) \cdot \ell^2/(2|\boldsymbol{c}|)$ for a constant $\varepsilon > 0$. A simple calculation (see appendix) yields that we can choose $\lambda = O(1)$. Then not only the expected drift away from the hyper-plane J (which contains the parent), yet also the expected drift towards the optimum is of the same order as for elitist selection. In particular, this drift ensures that the search never (w. o. p. for any polynomial number of steps) increases the distance from optimum by an ε-fraction again, where the constant ε can be chosen arbitrarily small. This implies that the 1/5-rule keeps $\sigma = \Theta(|\boldsymbol{c}|/n)$ also for comma selection. Hence, we obtain

THEOREM 4. *A constant κ exists such that for $\lambda \geq \kappa$ Theorem 3 also holds when using comma selection (instead of elitist selection).*

6. DISCUSSION AND OUTLOOK

Naturally, one may ask why the upper and the lower bound on the runtime do not meet – they differ by a factor of $O(\sqrt{\ln \lambda})$ for $\lambda \geq 2$. The model-based result on the progress-rate by Beyer (2001, Formula 3.116) indicates that the lower bound "w. o. p. $\Omega(n/\ln \lambda)$ iterations" is tight, i. e., that an (1+λ) ES using appropriately adapted isotropic mutations should be able to halve the approximation error in $O(n/\ln \lambda)$ iterations (i. e., $\lambda \cdot O(n/\ln \lambda)$ SPHERE-evaluations should suffice) with a high probability.

If so, two possibilities are left: Either the analysis presented here resulting in the upper bound is too weak, or the 1/5-rule fails to adapt the mutation strength σ appropriately. And indeed, the 1/5-rule fails. The reason is that this rule tries to maximize the expected gain of a single (isotropic) mutation. Yet in fact, the gain of the best of λ mutants should be maximized. Therefore, a slightly larger mutation strength seems appropriate – and necessary. This does not mean, however, that a "success frequency"-based rule cannot work at all: For the (1+λ) ES, a slight modification of the 1/5-rule indeed results in optimal performance: Therefore, assume $\lambda = O(n^{1-\varepsilon})$ for a constant $\varepsilon > 0$. The (1+λ) ES with modified 1/5-rule reads

1. FOR (int $i := 1$; $i \leq \lambda$; i++) DO
   Create a new search point $\boldsymbol{y}^{[i]} := \boldsymbol{c} + \boldsymbol{m}$ with $\boldsymbol{m} := \sigma \cdot \widetilde{\boldsymbol{m}}$, where each component of $\widetilde{\boldsymbol{m}} \in \mathbb{R}^n$ is independently standard-normally distributed.
2. IF $\min_{i \in \{1,\ldots,\lambda\}}\{f(\boldsymbol{y}^{[i]})\} \leq f(\boldsymbol{c})$
   THEN $g := g+1$ and $\boldsymbol{c} := \mathrm{argmin}_{i \in \{1,\ldots,\lambda\}}\{f(\boldsymbol{y}^{[i]})\}$
   ELSE $b := b+1$.
3. IF $\lambda \cdot (b+g) \geq n$ THEN
   (a) IF $g < (g+b) \cdot (1/5)$
   THEN $\sigma := \sigma/2$ ELSE $\sigma := \sigma \cdot 2$.
   (b) $g := 0$. $b := 0$.
4. GOTO 1.

The number of steps for observation/between two sequent σ-adaptations (namely $\lceil n/\lambda \rceil$) is $\Omega(n^\varepsilon)$. The *number of steps* in which the best of the lambda mutants is at least as good as its parent is counted – rather than the number of mutants. Recall that there is a certain $\alpha' = \Theta(1)$ such that $\mathsf{P}\{G_\ell^\lambda \geq \alpha' \cdot \ell \cdot \sqrt{\ln \lambda}/\sqrt{n})\} = 1/5$ as we have seen in the previous section. Recall also that for a mutation with $|\boldsymbol{m}| = \ell$, $|\boldsymbol{c}+\boldsymbol{m}| \leq |\boldsymbol{c}| \Leftrightarrow G \geq \ell^2/(2\,|\boldsymbol{c}|)$. Solving $\alpha' \cdot \ell \cdot \sqrt{\ln \lambda}/\sqrt{n} = \ell^2/(2\,|\boldsymbol{c}|)$ for ℓ yields $\ell' := 2 \cdot |\boldsymbol{c}| \cdot \alpha' \cdot \sqrt{\ln \lambda}/\sqrt{n}$. Thus, if isotropic mutations with length ℓ' were used, then a step would succeed (i. e., the best of the λ mutants is at least as good as the parent) with probability 1/5. Interestingly, ℓ' is by a factor of $\Theta(\sqrt{\ln \lambda})$ larger than the expected length of a scaled Gaussian mutation when using the original 1/5-rule.

So, ℓ' is such that $\mathsf{P}\{G_{\ell'}^\lambda \geq \ell'^2/(2\,|\boldsymbol{c}|)\} = 1/5$. Moreover, $\ell'^2/(2\,|\boldsymbol{c}|) = 2 \cdot |\boldsymbol{c}| \cdot \alpha'^2 \cdot \ln \lambda/n = \Theta(\ln \lambda \cdot |\boldsymbol{c}|/n)$. The results from the previous section tell us that in fact $\mathsf{P}\{G_{\ell'}^\lambda \geq \beta \cdot \ln \lambda \cdot |\boldsymbol{c}|/n\} = \Omega(1)$ for any constant β. In particular, we can choose β such that (given $|\boldsymbol{m}| = \ell'$) $\mathsf{P}\{|\boldsymbol{c}+\boldsymbol{m}| \leq |\boldsymbol{c}| - \ln \lambda \cdot |\boldsymbol{c}|/n\} = \Omega(1)$, i. e., the distance from the optimum is reduced by a $(\ln \lambda/n)$-fraction with probability $\Omega(1)$ – if isotropic mutations of length ℓ' were used.

If there was no page limit, we would now show that the modified 1/5-rule in fact ensures $\sigma = \Theta(\sqrt{\ln \lambda} \cdot |\boldsymbol{c}|/\sqrt{n})$ (at least w. o. p. for any polynomial number of steps) so that $\mathsf{E}[\sigma \cdot \widetilde{\boldsymbol{m}}] = \Theta(\ell') = \Theta(\sqrt{\ln \lambda} \cdot |\boldsymbol{c}|/n)$ for scaled Gaussian mutations – which is precisely by a factor $\Theta(\sqrt{\ln \lambda})$ larger than with the original 1/5-rule. Then we could again re-do the

analysis presented by Jägersküpper (2003), and we would finally obtain

THEOREM 5. *Let the (1+λ) ES, where $\lambda = O(n^{1-\varepsilon})$ for a constant $\varepsilon > 0$, minimize* SPHERE *using Gaussian mutations adapted by the* modified *1/5-rule. Assume that after initialization $\sigma = \Theta(\ln\lambda \cdot |\boldsymbol{c}|/n)$. Then the number of steps until the approximation error is reduced to a 2^{-b}-fraction, $b\colon \mathbb{N} \to \mathbb{N}$ such that $b = \mathrm{poly}(n)$, is $O(b{\cdot}n/\ln\lambda)$ w. o. p. (the number of* SPHERE*-evaluations is $O(b \cdot n \cdot \lambda/\ln\lambda)$ w. o. p.).*

It remains an open question, however, whether there is a "simple" (deterministic?) σ-adaptation (for isotropic mutations) that makes also the (1,λ) ES achieve optimal performance on SPHERE. The question may also read: Can we prove that it does achieve optimal performance.

References

Beyer, H.-G. (2001): *The Theory of Evolution Strategies*. Springer.

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

Ericson, T., Zinoviev, V. (2001): *Codes on Euclidean Spheres*. Elsevier, Amsterdam.

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

Hansen, N., Ostermeier, A. (1996): *Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation*. In *Proc. IEEE Int'l Conference on Evolutionary Computation (ICEC)*, 312–317.

Holland, J. H. (1975): *Adaption in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor, MI, USA.

Jägersküpper, J. (2003): *Analysis of a simple evolutionary algorithm for minimization in Euclidean spaces*. In *Proc. 30th Int'l Colloquium on Automata, Languages and Programming (ICALP)*, vol. 2719 of *LNCS*, 1068–79, Springer.

Jägersküpper, J. (2005): *Rigorous runtime analysis of the (1+1) ES: 1/5-rule and ellipsoidal fitness landscapes*. In *Foundations of Genetic Algorithms: 8th Int'l Workshop, Revised Selected Papers (FOGA)*, vol. 3469 of *LNCS*, 260–281, Springer.

Jägersküpper, J., Witt, C. (2005): *Rigorous runtime analysis of a (μ+1) ES for the sphere function*. In *Proc. Genetic and Evolutionary Computation Conference (GECCO)*, 849–856, ACM Press.

Jansen, T., De Jong, K. A. (2002): *An analysis of the role of offspring population size in EAs*. In *Proc. Genetic and Evolutionary Computation Conference (GECCO)*, 238–246, Morgan Kaufmann.

Neumann, F., Wegener, I. (2004): *Randomized local search, evolutionary algorithms, and the minimum spanning tree problem*. In *Proc. Genetic and Evolutionary Computation Conference (GECCO)*, vol. 3102 of *LNCS*, 713–724, Springer.

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

Schwefel, H.-P. (1995): *Evolution and Optimum Seeking*. Wiley, New York.

Storch, T., Wegener, I. (2003): *Real royal road functions for constant population size*. In *Proc. Genetic and Evolutionary Computation Conference (GECCO '03)*, vol. 2724 of *LNCS*, 1406–17, Springer.

Witt, C. (2004): *An analysis of the (μ+1) EA on simple pseudo-boolean functions*. In *Proc. Genetic and Evolutionary Computation Conference (GECCO)*, vol. 3102 of *LNCS*, 761–773, Springer.

Witt, C. (2005): *Worst-case and average-case approximations by simple randomized search heuristics*. In *Proc. 22nd Annual Symposium on Theoretical Aspects of Computer Science (STACS)*, vol. 3404 of *LNCS*, 44–56, Springer.

Yao, X., Liu, Y., Lin, G. (1999): *Evolutionary programming made faster*. IEEE Transactions on Evolutionary Computation, 3(2):82–102.

APPENDIX

Proof of Lemma 1

The density of (the distribution of) $\boldsymbol{z}$ at a point $\boldsymbol{b} \in \mathbb{R}^n$ is

$$\mathsf{D}_{\boldsymbol{z}}(\boldsymbol{b}) \;=\; \int_{\boldsymbol{a}\in\mathbb{R}^n} \mathsf{D}_{\boldsymbol{y}}(\boldsymbol{b}-\boldsymbol{a}) \cdot \mathsf{D}_{\boldsymbol{x}}(\boldsymbol{a})\,\mathrm{d}\boldsymbol{a}.$$

We must merely show that $\mathsf{D}_{\boldsymbol{z}}(\boldsymbol{M}\boldsymbol{b}) = \mathsf{D}_{\boldsymbol{z}}(\boldsymbol{b})$ for any orthogonal matrix $\boldsymbol{M} \in \mathbb{R}^{n\times n}$ since, by choosing $\boldsymbol{M}$ appropriately, any vector of length $|\boldsymbol{b}|$ is covered. Recall that $\boldsymbol{x} \mapsto \boldsymbol{M}\boldsymbol{x}$ actually defines a bijection in $\mathbb{R}^n$ that preserves the vectors' lengths. We have

$$\begin{aligned}
\mathsf{D}_{\boldsymbol{z}}(\boldsymbol{M}\boldsymbol{b}) &= \textstyle\int_{\boldsymbol{a}\in\mathbb{R}^n} \mathsf{D}_{\boldsymbol{y}}(\boldsymbol{M}\boldsymbol{b}-\boldsymbol{a}) \cdot \mathsf{D}_{\boldsymbol{x}}(\boldsymbol{a})\,\mathrm{d}\boldsymbol{a} \\
&= \textstyle\int_{\boldsymbol{a}\in\mathbb{R}^n} \mathsf{D}_{\boldsymbol{y}}(\boldsymbol{M}\boldsymbol{b}-\boldsymbol{M}\boldsymbol{a}) \cdot \mathsf{D}_{\boldsymbol{x}}(\boldsymbol{M}\boldsymbol{a})\,\mathrm{d}\boldsymbol{a} \\
&\quad \text{because } \{\boldsymbol{M}\boldsymbol{a} \mid \boldsymbol{a} \in \mathbb{R}^n\} = \mathbb{R}^n \\
&= \textstyle\int_{\boldsymbol{a}\in\mathbb{R}^n} \mathsf{D}_{\boldsymbol{y}}(\boldsymbol{M}(\boldsymbol{b}-\boldsymbol{a})) \cdot \mathsf{D}_{\boldsymbol{x}}(\boldsymbol{M}\boldsymbol{a})\,\mathrm{d}\boldsymbol{a} \\
&= \textstyle\int_{\boldsymbol{a}\in\mathbb{R}^n} \mathsf{D}_{\boldsymbol{y}}(\boldsymbol{b}-\boldsymbol{a}) \cdot \mathsf{D}_{\boldsymbol{x}}(\boldsymbol{a})\,\mathrm{d}\boldsymbol{a} \;=\; \mathsf{D}_{\boldsymbol{z}}(\boldsymbol{b})
\end{aligned}$$

because $\mathsf{D}_{\boldsymbol{y}}(\boldsymbol{M}(\boldsymbol{b}-\boldsymbol{a})) = \mathsf{D}_{\boldsymbol{y}}(\boldsymbol{b}-\boldsymbol{a})$ and $\mathsf{D}_{\boldsymbol{x}}(\boldsymbol{M}\boldsymbol{a}) = \mathsf{D}_{\boldsymbol{x}}(\boldsymbol{a})$ as we have already seen in the proof of Proposition 1. □

Proof of an inequality

We prove $1-(1-p)^\lambda \geq 3\cdot p^\lambda$ for $\lambda \geq 2$ and $0 \leq p \leq 0.5$, starting with $\lambda = 2$.

$$\begin{aligned}
1-(1-p)^2 &\geq 3p^2 \\
\iff \quad 2p-p^2 &\geq 3p^2 \\
\iff \quad 2p &\geq (2p)^2
\end{aligned}$$

For $\lambda \geq 3$, we have $3p^\lambda = p^{\lambda-2}\cdot 3p^2$ as well as

$$\begin{aligned}
1-(1-p)^\lambda &= (1-p)^{\lambda-2}\big((1-p)^{2-\lambda} - (1-p)^2\big) \\
&\geq (1-p)^{\lambda-2}\big(\quad 1 \quad - (1-p)^2\big),
\end{aligned}$$

and hence, we merely have to show that $(1-p)^{\lambda-2} \geq p^{\lambda-2}$, which in fact holds since $p \in [0,1/2]$.

Moreover, if $0 \leq p \leq 1/2-\varepsilon$ for a constant $\varepsilon > 0$, then for any constant c, we can choose λ large enough such that

$$(1-p)^{\lambda-2} \geq (1/2+\varepsilon)^{\lambda-2} \geq c\cdot(1/2-\varepsilon)^{\lambda-2} \geq c\cdot p^{\lambda-2}$$

(and consequently $1-(1-p)^\lambda \geq 3c\cdot p^\lambda$). Thus, if g is such that $\mathsf{P}\{G_\ell \geq g\} \leq 1/2-\Omega(1)$, namely $g = \Omega(\ell/\sqrt{n})$, then $\mathsf{P}\{G_\ell^\lambda \geq g\} \geq 3\cdot c\cdot \mathsf{P}\{G_\ell^\lambda \leq -g\}$ for λ large enough.