Rigorous Runtime Analysis of the (1+1) ES: 1/5-Rule and Ellipsoidal Fitness Landscapes

Jens Jägersküpper*

Dept. of Computer Science 2, Univ. Dortmund, 44221 Dortmund, Germany
jj@Ls2.cs.uni-dortmund.de

Abstract. We consider the (1+1) Evolution Strategy, a simple evolutionary algorithm for continuous optimization problems, using so-called Gaussian mutations and the 1/5-rule for the adaptation of the mutation strength. Here, the function $f\colon \mathbb{R}^n \to \mathbb{R}$ to be minimized is given by a quadratic form $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}$, where $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ is a positive definite diagonal matrix and $\boldsymbol{x}$ denotes the current search point. This is a natural extension of the well-known SPHERE-function ($\boldsymbol{Q} = \boldsymbol{I}$). Thus, very simple unconstrained quadratic programs are investigated, and the question is addressed how $\boldsymbol{Q}$ effects the runtime. For this purpose, quadratic forms

$$f(\boldsymbol{x}) \quad = \quad \xi \cdot \left({x_1}^2 + \cdots + {x_{n/2}}^2\right) + {x_{n/2+1}}^2 + \cdots + {x_n}^2$$

with $\xi = \omega(1)$, i. e. $1/\xi \to 0$ as $n \to \infty$, and $\xi = \mathrm{poly}(n)$ are investigated exemplarily. It is proved that the optimization very quickly stabilizes and that, subsequently, the runtime (defined as the number of f-evaluations) to halve the approximation error is $\Theta(\xi \cdot n)$. Though $\xi \cdot n = \mathrm{poly}(n)$, this result actually shows that the evolving search point indeed creeps along the "gentlest descent" of the ellipsoidal fitness landscape.

1 Introduction

Finding − or at least approximating − an optimum of a given function $f\colon S \to \mathbb{R}$ is one of the fundamental problems − in theory as well as in practice. Methods for solving continuous optimization problems, e. g. $S = \mathbb{R}^n$, are usually classified into first-order, second-order, and zeroth-order methods depending on whether they utilize the gradient (the first derivative) of the objective function, the gradient and the Hessian (the second derivative), or neither of the two.

> Note that here "continuous" relates to the search space rather than to f, and that, unlike in math programming, throughout this paper "n" denotes the number of dimensions of the search space and *not* the number of optimization steps; "d" generally denotes a distance in the n-dimensional search space.

A zeroth-order method is also called *derivative-free* or *direct search method.* Newton's method is the example of a second-order method; first-order methods

* Supported by the German Research Foundation (DFG) as part of the research center "Computational Intelligence" (SFB 531)

A.H. Wright et al. (Eds.): FOGA 2005, LNCS 3469, pp. 260–281, 2005.

can be (sub)classified into Quasi-Newton, steepest descent, and conjugate gradient methods. Classical zeroth-order methods try to approximate the gradient in order to plug this estimate into a first-order method. Finally, amongst the "modern" zeroth-order methods, evolutionary algorithms (EAs) come into play. EAs for continuous optimization, however, are usually subsumed under the term *evolution(ary) strategies (ESs)*. Obviously, in general we cannot expect zeroth-order methods to out-perform first-order methods or even second-order methods.

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

For discrete search spaces, essentially $\{0,1\}^n$, such a theory has been developed successfully since the mid-1990s (cf. Wegener (2001) and Droste et al. (2002)). Recently, first results for non-artificial but well-known problems have been obtained (namely for the maximum matching problem by Giel and Wegener (2003), for sorting and the shortest-path problem by Scharnow et al. (2002), and for the minimum-spanning tree by Neumann and Wegener (2004)).

The situation for continuous evolutionary optimization is different. Here, the vast majority of the results are based on empiricism, i. e., experiments are performed and their outcomes are interpreted, which leads to a theory in the sense of physics rather than computer science. Also convergence properties of EAs have been studied to a considerable extent (e. g. Rudolph (1997), Greenwood and Zhu (2001), Bienvenue and Francois (2003)). A lot of results have been obtained by analyzing a simplifying model of the stochastic process induced by the EA, for instance by letting the number of dimensions approach infinity. Unfortunately, such results rely on experimental validation as a justification for the simplifications/inaccuracies introduced by the modeling. In particular Beyer has obtained numerous results that focus on local performance measures (*progress rate*, *fitness gain*; cf. Beyer (2001b)), i. e., the effect of a single mutation (or, more generally, of a single transition from one generation to the next) is investigated. Best-case assumptions concerning the mutation adaptation in this single step then provide estimates of the maximum gain a single step may yield. However, when one aims at analyzing the (1+1) ES as an algorithm, rather than a model of the stochastic process induced, a different, more algorithmic approach is needed. In 2003 a first theoretical analysis of the expected runtime, given by the number of function evaluations, of the (1+1) ES using the 1/5-rule was presented (Jägersküpper, 2003). The function/fitness landscape considered therein is the well-know SPHERE-function, given by $\text{SPHERE}(\boldsymbol{x}) := \sum_{i=1}^{n} x_i^2 = \boldsymbol{x}^\top \boldsymbol{I}\boldsymbol{x}$, and the multi-step behavior that the (1+1) ES bears when using the 1/5-rule for the adaptation of the mutation strength is rigorously analyzed. As mentioned in the abstract, the present paper will extend this result to a broader class of functions. One may guess that an ellipsoidal landscape is similar to the ridge-function scenario (especially to the parabolic ridge). Beyer (2001a) focuses on local measures for this fitness landscape. However, since ridge functions are unbounded, i. e. there is no optimum, and there is no need for adaptation, from an algorithmic point of view – when one is interested in adaptation mechanisms and how they work – ellipsoidal fitness landscapes are more challenging.

Finally note that, regarding the approximation error, for unconstrained optimization it is generally not clear how the runtime can be measured (solely) with respect to the absolute error of the approximation. In contrast to discrete and finite problems, the initial error is generally not bounded, and hence, the question how many steps it takes to get into the ε-ball around an optimum does not make sense without specifying the starting conditions. Hence, we must consider the runtime with respect to the relative improvement of the approximation. Given that the (relative) progress that a step yields becomes steady-state, considering the number of steps/f-evaluations to halve the approximation error is a natural choice. For the SPHERE-function, Jägersküpper (2003) gives a proof that the 1/5-rule makes the (1+1) ES perform $\Theta(n)$ steps to halve the distance from the optimum and, in addition, that this is asymptotically the best possible w. r. t. isotropically distributed mutation vectors, i. e., for any adaptation of isotropic mutations, the expected number of f-evaluations is $\Omega(n)$ (moreover, for any constant $\varepsilon > 0$, $O(n^{1-\varepsilon})$ f-evaluations suffice only with an exponentially small probability).

The Algorithm

We will concentrate on the (1+1) evolution strategy *((1+1) ES)*, which dates back to the mid 1960s (cf. Rechenberg (1973) and Schwefel (1995)). This simple EA uses solely mutation due to a single-individual population, where here "individual" is just a synonym for "search point". Let $\boldsymbol{c} \in \mathbb{R}^n$ denote the current individual. Given a starting point, i. e. an initialization of $\boldsymbol{c}$, the (1+1) ES performs the following evolution loop:

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

Since a worse mutant (with respect to the function to be minimized) is always discarded, the (1+1) ES is a randomized hill climber, and the selection rule is called *elitist selection.* Fortunately, for the type of results we are after we need not define a reasonable stopping criterion. How the mutation vectors are generated must be specified, though. Originally, the mutation vector $\boldsymbol{m} \in \mathbb{R}^n$ is generated by firstly generating a *Gaussian mutation* vector $\widetilde{\boldsymbol{m}} \in \mathbb{R}^n$ each component of which is independently standard normal distributed; subsequently, this vector is scaled by the multiplication with a scalar $s \in \mathbb{R}_{>0}$, i. e. $\boldsymbol{m} = s \cdot \widetilde{\boldsymbol{m}}$. Gaussian mutations are the most common type of mutations (for the search space $\mathbb{R}^n$) and, therefore, will be considered here. The crucial property of a Gaussian mutation is that $\widetilde{\boldsymbol{m}}$, and with it $\boldsymbol{m}$, is isotropically distributed, i. e., $\boldsymbol{m}/\left|\boldsymbol{m}\right|$ is uniformly distributed upon the unit hypersphere and the length of the mutation, namely the random variable $|\boldsymbol{m}|$, is independent of the direction $\boldsymbol{m}/\left|\boldsymbol{m}\right|$.

The state of the art in mutation adaptation seems to be the *covariance matrix adaptation (CMA)* (Hansen and Ostermeier, 1996) where $s \cdot \mathbf{B} \cdot \widetilde{\boldsymbol{m}}$ makes up the mutation vector with a matrix $\mathbf{B} \in \mathbb{R}^{n \times n}$ which is also adapted. Unlike $\mathbf{B} = t \cdot \mathbf{I}$ for some scalar t, the mutation vector is not isotropically distributed.

The question that naturally arises is how the scaling factor s is to be chosen. Obviously, the smaller the approximation error, i. e., the closer $\boldsymbol{c}$ is to an optimum, the shorter $\boldsymbol{m}$ needs to be for a further improvement of the approximation to be possible. Unfortunately, the algorithm does not know about the current approximation error, but can utilize only the knowledge obtained by f-evaluations. Based on experiments and rough calculations for two function scenarios (namely Sphere and a corridor function), Rechenberg proposed the *1/5-(success-)rule.* The idea behind this adaptation mechanism is that in a step of the (1+1) ES the mutant should be accepted with probability 1/5. Hereinafter, a mutation that results in $f(\boldsymbol{c}') \leq f(\boldsymbol{c})$ is called *successful,* and hence, when talking about a mutation, *success probability* denotes the probability that the mutant $\boldsymbol{c}' = \boldsymbol{c} + \boldsymbol{m}$ is at least as good as $\boldsymbol{c}$. Obviously, when elitist selection is used, the success probability of a step equals the probability that the mutation is accepted in this step. If every step was successful with probability 1/5, we would observe that on the average one fifth of the mutations are successful. Thus, the 1/5-rule works as follows: the optimization process is observed for n steps without changing s; if more than one fifth of the steps in this observation phase have been successful, s is doubled, otherwise, s is halved.

Various implementations of the 1/5-rule can be found in the literature, yet in fact, one result of (Jägersküpper, 2003) is that the order of the runtime is indeed not affected as long as the observation lasts $\Theta(n)$ steps and the scaling factor s is multiplied by a constant greater than 1 resp. by a positive constant smaller than 1.

The Function Scenario

In this section we will have a closer look at the fitness landscape under consideration and preview isotropic mutations in this scenario. Note that, as minimization is considered, "function value" ("f-value") will be used rather than "fitness". Since the optimum function value is 0, the current approximation error is defined as $f(\boldsymbol{c})$, the f-value of the current individual. As mentioned in the abstract, we are going to consider the fitness landscapes induced by certain positive definite quadratic forms.

At first glance, one might guess that mixed terms like $3x_1x_2$ may crucially affect the fitness landscape induced by a positive definite quadratic form $\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}$. However, this is not the case. First note that w. l. o. g. we may assume $\boldsymbol{Q}$ to be symmetric (by balancing Q_{ij} with Q_{ji} for $i \neq j$). Furthermore, any symmetric matrix can by diagonalized since it has n eigen vectors. Namely, eigendecomposition yields $\boldsymbol{Q} = \boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^{-1}$ for a diagonal matrix $\boldsymbol{D}$ and an orthogonal matrix $\boldsymbol{R}$.

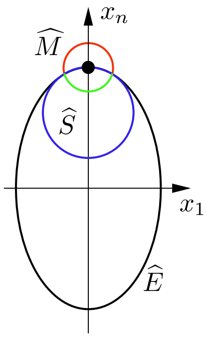
> Note that an orthogonal matrix $\boldsymbol{R}$ corresponds to a orthonormal transformation, which is merely a (possibly improper) rotation; then $\boldsymbol{R}^{-1}$ is the corresponding "anti-rotation".

Thus, the quadratic form equals $\boldsymbol{x}^\top \boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^{-1}\boldsymbol{x}$, and since $\boldsymbol{x}^\top \boldsymbol{R} = (\boldsymbol{R}^\top \boldsymbol{x})^\top$, we have $(\boldsymbol{R}^\top \boldsymbol{x})^\top \boldsymbol{D}(\boldsymbol{R}^{-1}\boldsymbol{x})$. As $\boldsymbol{R}^\top = \boldsymbol{R}^{-1}$ for an orthogonal matrix, the quadratic form equals $(\boldsymbol{R}^{-1}\boldsymbol{x})^\top \boldsymbol{D}(\boldsymbol{R}^{-1}\boldsymbol{x})$. Thus, investigating $\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}$ using the standard basis for $\mathbb{R}^n$ (given by $\boldsymbol{I}$) is the same as investigating $\boldsymbol{x}^\top \boldsymbol{D}\boldsymbol{x}$ using the orthonormal basis given by $\boldsymbol{R}$. Finally note that the inner product is independent of the orthonormal basis that we use (because $(\boldsymbol{R}\boldsymbol{x})^\top(\boldsymbol{R}\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{R}^\top \boldsymbol{R}\boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{R}^{-1}\boldsymbol{R}\boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{I}\boldsymbol{x} = \boldsymbol{x}^\top \boldsymbol{x}$). Consequently, we may assume that $\boldsymbol{Q}$ is a diagonal matrix each entry of which is positive. In other words, when talking about positive definite quadratic forms we are in fact talking about functions of the form $f_n(\boldsymbol{x}) = \sum_{i=1}^{n} \xi_i \cdot {x_i}^2$ with $\xi_i > 0$, and we may even assume $\xi_n \geq \cdots \geq \xi_1$.

As mentioned in the abstract, we exemplarily restrict ourselves to the following class of (sequences of) quadratic forms, where $n \in 2\mathbb{N}$ and $1/\xi \to 0$ as $n \to \infty$:

$$f_n(\boldsymbol{x}) \quad := \quad \xi \cdot \left({x_1}^2 + \cdots + {x_{n/2}}^2\right) + {x_{n/2+1}}^2 + \cdots + {x_n}^2$$

Hence, $f_n(\boldsymbol{x}) = \xi \cdot \textsc{Sphere}_{n/2}(\boldsymbol{y}) + \textsc{Sphere}_{n/2}(\boldsymbol{z})$ where $\boldsymbol{y} := (x_1, \ldots, x_{n/2})$ and $\boldsymbol{z} := (x_{n/2+1}, \ldots, x_n)$. Thus, the aim is to minimize the sum of two separate sphere functions, in $S_1 = \mathbb{R}^{n/2}$ resp. $S_2 = \mathbb{R}^{n/2}$, having weight ξ resp. 1, i. e., $f(\boldsymbol{x}) = \xi \cdot |\boldsymbol{y}|^2 + |\boldsymbol{z}|^2$, where $|\cdot|$ denotes the length of a vector in Euclidean space (Euclidean norm). Recall that the mutation vector $\boldsymbol{m}$ equals $s \cdot \widetilde{\boldsymbol{m}}$. As each component of $\widetilde{\boldsymbol{m}}$ is independently standard normal distributed, $\boldsymbol{m_1} := (m_1, \ldots, m_{n/2})$ and $\boldsymbol{m_2} := (m_{n/2+1}, \ldots, m_n)$ are two independent $(n/2)$-dimensional Gaussian mutations which are respectively scaled by the same factor s. Obviously, $\boldsymbol{m_1}$ only affects $\boldsymbol{y}$, whereas $\boldsymbol{m_2}$ only affects $\boldsymbol{z}$, and thus, the f-value of the mutant equals $\xi \cdot |\boldsymbol{y} + \boldsymbol{m_1}|^2 + |\boldsymbol{z} + \boldsymbol{m_2}|^2$.



Let $d_1 := |\boldsymbol{y}|$ and $d_2 := |\boldsymbol{z}|$ denote the distance from the origin/optimum in S_1 resp. S_2. Since Gaussian mutations as well as SPHERE are invariant with respect to rotations of the coordinate system, we may rotate S_1 and S_2 such that we are located at $(d_1, 0, \ldots, 0) \in S_1$ resp. $(0, \ldots, 0, d_2) \in S_2$. In other words, we may assume w. l. o. g. that the current search point is located at $(d_1, 0, \ldots, 0, d_2) \in \mathbb{R}^n$, i. e., that it lies in the x_1-x_n-plane. In fact, we have just described a projection $\widehat{\ }$: $\mathbb{R}^n \to \mathbb{R}^2$. Note that due to the properties of f and Gaussian mutations this projection only conceals irrelevant information, i. e., all information relevant to the analysis is preserved. Thus, we can concentrate on the 2D-projection as depicted in the figure. For some arguments, however, it is crucial to keep in mind that this projection is based on the fact that the current search point, and also its mutant, can be assumed to lie in the x_1-x_n-plane w. l. o. g. (obviously, for the mutant to lie in this plane, S_1 and S_2 must almost surely (a. s.) be re-rotated).

In the next section some of the results presented in (Jägersküpper, 2003), which will be used here, will be shortly restated. In Section 3 the crucial properties of a single mutation in the considered fitness landscape are discussed, and in the subsequent section we will have a closer look at the adaptation, i. e., the multi-step behavior of the (1+1) ES will be analyzed for the considered function class/fitness landscape. We end with some concluding remarks in Section 5.

2 Preliminaries

In this section some notions and notations are introduced. Furthermore, the results obtained for the SPHERE-scenario in (Jägersküpper, 2003) that we will use are recapitulated; for more details cf. (Jägersküpper, 2002).

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

A statement $Z(n)$ holds ***for n large enough*** *if $(\exists\, n_0 \in \mathbb{N})(\forall\, n \geq n_0)\, Z(n)$.*

Recall the following asymptotics: $g(n) = O(h(n))$ 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 both $O(h(n))$ and $\Omega(h(n))$; for $h(n), g(n) > 0$, we have $g(n) = o(h(n))$ iff $g(n)/h(n) \to 0$ as $n \to \infty$ and $g(n) = \omega(h(n))$ iff $h(n) = o(g(n))$. 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).

Let $\boldsymbol{c} \in \mathbb{R}^n - \{\boldsymbol{0}\}$ denote a search point and $\boldsymbol{m}$ a scaled Gauss mutation. Note that $\text{SPHERE}(\boldsymbol{c}) = |\boldsymbol{c}|^2$ (recall that $|\boldsymbol{c}|$ is the L^2-norm (Euclidian length) of $\boldsymbol{c}$). The analysis of the (1+1) ES for SPHERE has shown that for n large enough

$$\mathsf{P}\{|\boldsymbol{c}+\boldsymbol{m}| \leq |\boldsymbol{c}| \mid |\boldsymbol{m}| = \ell\} \geq \varepsilon \text{ for a constant } \varepsilon \in (0, \tfrac{1}{2}) \iff \ell = O(|\boldsymbol{c}| / \sqrt{n}),$$

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

$$\mathsf{P}\{|\boldsymbol{c}+\boldsymbol{m}| \leq |\boldsymbol{c}| \mid |\boldsymbol{m}| = \ell\} \leq \varepsilon \text{ for a constant } \varepsilon \in (0, \tfrac{1}{2}) \iff \ell = \Omega(|\boldsymbol{c}| / \sqrt{n}),$$

in other words, the mutant obtained by an isotropic mutation of $\boldsymbol{c}$ is closer to a predefined point, here again the origin, with a constant probability strictly smaller than 1/2 iff the length of the mutation vector is at least an $\Omega(1/\sqrt{n})$-fraction of the distance between $\boldsymbol{c}$ and this point. (The actual constant ε respectively correlates with the constant in the O-notation resp. in the Ω-notation.)

The expected length of $\boldsymbol{m}$ equals $s \cdot \mathsf{E}[|\widetilde{\boldsymbol{m}}|] = s \cdot \sqrt{n} \cdot (1 - \Theta(1/n))$ since $|\widetilde{\boldsymbol{m}}|$ is χ-distributed (with n degrees of freedom). Moreover, with $\bar{\ell} := \mathsf{E}[|\boldsymbol{m}|]$ we have $\mathsf{P}\big\{\big||\boldsymbol{m}| - \bar{\ell}\big| \geq \delta \cdot \bar{\ell}\big\} \leq \delta^{-2}/(2n-1)$ for $\delta > 0$, in other words, there is only small deviation in the length of a mutation; e. g., with probability $1 - O(1/n)$ the mutation vector's actual length differs from its expected length by no more than $\pm 1\%$.

Concerning the mutation adaptation by the 1/5-rule for Sphere, we know that there exists a constant $p_h \in (0, 1/5)$ such that if the success probability of the mutation in the first step of an observation phase is smaller than p_h, then w. o. p. less than 1/5 of the steps in this phase are successful so that the scaling factor is halved. Analogously, a constant $p_d \in (1/5, 1/2)$ exists such that if the first step of a phase is successful with probability at least p_d, then w. o. p. more than 1/5 of the steps in this phase are successful so that s is doubled. This can be used to show that the 1/5-rule in fact ensures that each step is successful with a probability in $[a, b] \subset (0, 1/2)$ for two constants a, b.

Let $\Delta := |\boldsymbol{c}| - |\boldsymbol{c}'|$ denote the spatial gain towards the origin, the optimum of Sphere, in a step. For Sphere, a mutation is accepted (by elitist selection) iff $\Delta \geq 0$. Consequently, negative gains are zeroed out so that the expected spatial gain of a step is $\mathsf{E}\big[\Delta \cdot \mathbb{1}_{\{\Delta \geq 0\}}\big]$. We know that $\mathsf{E}\big[\Delta \cdot \mathbb{1}_{\{\Delta \geq 0\}}\big]$ is $O(\bar{\ell}/\sqrt{n})$ and – however the scaling factor is chosen/adapted – also $O(|\boldsymbol{c}| / n)$. Furthermore, $\mathsf{E}\big[\Delta \cdot \mathbb{1}_{\{\Delta \geq 0\}} \mid s = \Theta(|\boldsymbol{c}| / n)\big]$ is $\Omega(\bar{\ell}/\sqrt{n})$ and $\Omega(|\boldsymbol{c}| / n)$, i. e., the distance from the optimum is expected to decrease by an $\Theta(1/n)$-fraction if s is chosen/adapted appropriately. Furthermore, in this situation for any constant $\kappa > 0$ the distance decreases (at least) by an κ/n-fraction with probability $\Omega(1)$.

Recall that $\bar{\ell} = s \cdot \sqrt{n} \cdot (1 - \Theta(1/n))$. Thus, when scaled Gaussian mutations are used, "$s = \Theta(|\boldsymbol{c}| / n)$" is equivalent to "$\bar{\ell} = \Theta(|\boldsymbol{c}| / \sqrt{n})$" which is again equivalent to "$\exists$ constant $\varepsilon > 0$ such that for n large enough $\mathsf{P}\{\Delta \geq 0\} \in [\varepsilon, 1/2 - \varepsilon]$" since $\mathsf{P}\big\{|\boldsymbol{m}| = \Theta(\bar{\ell})\big\} = 1 - O(1/n)$. The equivalance of these three events/conditions will be of great help in the argumentation.

3 Gain in a Single Step

In this section we now take a closer look at the properties of a Gaussian mutation in the ellipsoidal fitness landscape under consideration. Since $\xi = \omega(1)$, $\xi > 1$ for n large enough, and therefore, we assume $\xi > 1$ in the following. Furthermore,

"f" will also be used as an abbreviation of the f-value of the current individual and "f'" stands for the mutant's f-value.

Recall that $f = \xi\, d_1^2 + d_2^2$ (for the current search point) and $f' = \xi\, {d'_1}^2 + {d'_2}^2$ (for its mutant), where $d'_1 := |\boldsymbol{y} + \boldsymbol{m_1}|$ and $d'_2 := |\boldsymbol{z} + \boldsymbol{m_2}|$. The crucial point to the analysis is the answer to the question how d_1, d_2 and the scaling factor s – and with it $|\boldsymbol{m}|$ – relate when the success probability of a step, i. e. the probability that the mutant is accepted, is about 1/5. In other words, how does the length of the mutation vector depend on d_1 and d_2, and how do d_1 and d_2 relate. Since $\nabla\widehat{f}(d_1, d_2) = (\xi\, 2\, d_1, 2\, d_2)^\top$, for a search point satisfying $d_1/d_2 = 1/\xi$ an infinitesimal change of d_1 has the same effect on f as an infinitesimal change of d_2. Though the length of a mutation is not infinitesimal, this may be taken as an indicator that the ratio d_1/d_2 will stabilize when using isotropic mutations, and indeed, it turns out that the process stabilizes w. r. t. $d_1/d_2 = \Theta(1/\xi)$. In this section, we will see that near the gentlest descent in our ellipsoidal fitness landscape, namely for $d_1/d_2 = O(1/\xi)$, a mutation succeeds with a constant probability greater than 0 but smaller than 1/2 iff the scaling factor s is $\Theta((\sqrt{f}/n)/\xi)$. Furthermore, asymptotically tight bounds on the expected f-gain of a single step in such a situation will be obtained. Therefore, we will show that a mutation of a search point $\boldsymbol{c}$ for which $d_1/d_2 = O(1/\xi)$ with a mutation using a scaling factor $s = \Theta((\sqrt{f}/n)/\xi)$ in the ellipsoidal fitness landscape is "similar" to the mutation of a search point $\boldsymbol{x}$ in the Sphere scenario with $\text{Sphere}(\boldsymbol{x}) = \Theta(f/\xi^2)$ (using the same scaling factor).

We start our analysis at a point $\boldsymbol{c}$ with $\widehat{\boldsymbol{c}} = (0, \phi)$, i. e. $d_1 = 0$ and $d_2 = \phi$, so that $f = \phi^2$. Consequently, $\widehat{\boldsymbol{c}}$ is located at a point with gentlest descent w. r. t. all points with f-value ϕ^2, and hence, the curvature of the 2D-curve given by the projection $\widehat{E}$ of the n-ellipsoid $E := \{\boldsymbol{x} \mid f(\boldsymbol{x}) = f(\boldsymbol{c})\} \subset \mathbb{R}^n$, is maximum at $\widehat{\boldsymbol{c}}$. By a simple application of differential geometry, we get that the curvature of this 2D-curve at $\widehat{\boldsymbol{c}}$ equals ξ/ϕ. Consequently, the radius of the osculating circle ($\widehat{S}$ in the figure) equals ϕ/ξ. As this circle $\widehat{S}$ actually lies in the x_1-x_n-plane, it is the equator of an n-sphere S with radius ϕ/ξ (the center of which lies on the x_n-axis, just like the current search point $\boldsymbol{c}$). Note that this sphere lies completely inside E such that $S \cap E = \{\boldsymbol{c}\}$. Thus, the probability that a mutation hits inside S is a lower bound on the probability that $f' \leq f$, i. e.,

$$\begin{aligned}
\mathsf{P}\{f' \leq f\} &= \mathsf{P}\{\boldsymbol{c} + \boldsymbol{m} \text{ lies inside } E\} \\
&\geq \mathsf{P}\{\boldsymbol{c} + \boldsymbol{m} \text{ lies inside } S\} \\
&= \mathsf{P}\Big\{|\boldsymbol{x} + \boldsymbol{m}| \leq |\boldsymbol{x}| \text{ for some } \boldsymbol{x} \text{ with } |\boldsymbol{x}| = \text{radius of } \widehat{S} = \phi/\xi\Big\} \\
&= \mathsf{P}\big\{\text{Sphere}(\boldsymbol{x} + \boldsymbol{m}) \leq \text{Sphere}(\boldsymbol{x}) \mid \text{Sphere}(\boldsymbol{x}) = (\phi/\xi)^2\big\}.
\end{aligned}$$

In fact, our argumentation yields that the above (in)equalities hold for any fixed length ℓ of the mutation vector $\boldsymbol{m}$, i. e., if the probabilities are conditioned on the event $\{|\boldsymbol{m}| = \ell\}$, respectively. Since ℓ is arbitrary here and the radius of S is independent of ℓ, they remain valid when this condition is dropped.

For an upper bound on the probability that a mutation hits inside E, consider a mutation (vector) having length $\ell < 2\phi$ (since for $\ell \geq 2\phi$, E lies inside M).

Let $M = \{\boldsymbol{x} \mid |\boldsymbol{c} - \boldsymbol{x}| = \ell\} \subset \mathbb{R}^n$ denote the mutation sphere consisting of all potential mutants. Then $\widehat{M}$ is a circle (cf. the figure above) with radius ℓ centered at $\widehat{\boldsymbol{c}}$. (Note that, though $\boldsymbol{c}' = \boldsymbol{c} + \boldsymbol{m}$, given $|\boldsymbol{m}| = \ell$, is uniformly distributed upon M, $\widehat{\boldsymbol{c}'}$ is *not* uniformly distributed upon $\widehat{M}$). Now consider the curvature at a point in $\widehat{E} \cap \widehat{M} = \{\boldsymbol{z}_1, \boldsymbol{z}_2\}$ (there are exactly two points of intersection since $0 < \ell < 2\phi$). Simple differential geometry shows that the curvature at $\boldsymbol{z}_i$ is $\kappa_\ell = \Theta(\xi/\phi)$ if $\ell = O(\phi/\xi)$. As the curvature at any point of $\widehat{E}$ that lies inside $\widehat{M}$ is greater than κ_ℓ (since $\xi > 1$), $\widehat{\boldsymbol{c}}$ as well as $\boldsymbol{z}_i$ lie inside the osculating circle at $\boldsymbol{z}_{3-i}$ which has radius $r_\ell := 1/\kappa_\ell = \Theta(\phi/\xi)$ if $\ell = O(\phi/\xi)$. Thus, there is also a circle with radius r_ℓ passing through $\widehat{\boldsymbol{c}}$ such that $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ lie inside this circle. Therefore, the circle passing through $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, and $\hat{\boldsymbol{c}}$ has a radius smaller than r_ℓ, and again, this circle actually lies in the x_1-x_n-plane of the search space and is the image of the n-sphere having this circle as an equator. Hence,

$$\begin{aligned} & \mathsf{P}\{f' \le f \mid |\boldsymbol{m}| = \ell\} \\ \le\ & \mathsf{P}\big\{\textsc{Sphere}(\boldsymbol{x} + \boldsymbol{m}) \le \textsc{Sphere}(\boldsymbol{x}) \mid \textsc{Sphere}(\boldsymbol{x}) = (\alpha\,\phi/\xi)^2, |\boldsymbol{m}| = \ell\big\} \end{aligned}$$

where $\alpha = \Theta(1)$ if $\ell = O(\phi/\xi)$. (Besides, $\alpha \searrow 1$, i. e. $r_\ell \searrow \phi/\xi$, as $\ell \searrow 0$.)

Recall that we assumed $\widehat{\boldsymbol{c}} = (0, \phi) \in \mathbb{R}^2$, i. e. $d_1 = 0$ and $d_2 = \phi$, in the above argumentation. The estimates we have made for the bounds on the probability of a mutation hitting inside the n-ellipsoid E, however, remain valid as long as $d_1/d_2 = O(1/\xi)$: Since ξ/ϕ is the maximum curvature of $\widehat{E}$, there is always a circle $\widehat{S}$ with radius ϕ/ξ lying inside $\widehat{E}$ such that $\widehat{S} \cap \widehat{E} = \{\widehat{\boldsymbol{c}}\}$, and since $\widehat{S}$ is in fact an equator of an n-sphere S, S lies completely inside E such that $S \cap E = \{\boldsymbol{c}\}$. For the upper bound, we must merely consider the $\boldsymbol{z_i}$ at which the curvature is smaller, and indeed, it turns out that as long as $d_1/d_2 = O(1/\xi)$ and $\ell = O(\phi/\xi)$, κ_ℓ remains $\Theta(\xi/\phi)$.

Hence, when $f(\boldsymbol{c}) = \phi^2$ such that $\boldsymbol{c}$ satisfies $d_1/d_2 = O(1/\xi)$, we are in a situation resembling (w. r. t. the success probability of a mutation) the minimization of Sphere at a point having distance $\Theta(\phi/\xi)$ from the optimum/origin. Concerning the 1/5-rule, we then know (cf. Section 2) that

$$\begin{aligned} & \exists \text{ constant } \varepsilon > 0 \text{ such that for } n \text{ large enough } \mathsf{P}\{f' \le f\} \in [\varepsilon, 1/2 - \varepsilon] \\ & \iff s = \Theta((\phi/\xi)/n) \iff \bar{\ell} = \Theta((\phi/\xi)/\sqrt{n}) \end{aligned}$$

where ε correlates with the two multiplicative constants within the Θ-notation.

Thus, we are now going to investigate the gain of a step when $f = \phi^2$ and $s = \Theta((\phi/\xi)/n)$. As we have seen above, there exists an n-sphere S with radius $r = \phi/\xi$ lying completely in E such that $S \cap E = \{\boldsymbol{c}\}$. Again owing to the results for Sphere, we know that a mutation having length $\ell = \Theta(r/\sqrt{n})$ hits with probability $\Omega(1)$ a hyperspherical cap $C \subset M$ containing all points of M that are at least $\Omega(r/n)$ closer to the center of S than $\boldsymbol{c}$. Consequently, with probability $\Omega(1)$ the mutant lies inside E such that its distance from E is $\Theta(r/n)$, i. e. $\Theta((\phi/\xi)/n)$. If we pessimistically assume that this spatial gain were realized along the gentlest descent of f, i. e. $d_1 = 0$ and $d_1' = 0$ so that $d_2' = d_2 - \Theta((\phi/\xi)/n)$, we obtain that with probability $\Omega(1)$

$$\begin{aligned}
f' &\le (\phi - \Theta((\phi/\xi)/n)\,)^2 \\
&= \phi^2 - 2\alpha\phi^2/(\xi n) + \alpha^2\phi^2/(\xi n)^2 \text{ for some } \alpha = \Theta(1) \\
&= \phi^2 - \underbrace{\alpha(2-\alpha/(\xi n))}\,\phi^2/(\xi n) \\
&= \phi^2 - \qquad \Theta(1) \qquad \phi^2/(\xi n) \\
&= f - \Theta(f/(\xi n)).
\end{aligned}$$

Let $\boldsymbol{c}'' := \arg\min\{f(\boldsymbol{c}),\, f(\boldsymbol{c}')\}$ denote the search point that gets selected by elitist selection. Since mutants with a larger f-value are rejected, i. e. $f'' \le f$, this implies for the expected f-gain of a step

$$\mathsf{E}\Big[f'' \mid s = \Theta((\sqrt{f}/n)/\xi)\Big] \;=\; f - \Omega(f/(\xi n)).$$

Due to the pessimistic assumptions, this lower bound on the f-gain just derived is valid only for $s = \Theta((\sqrt{f}/n)/\xi))$, yet it holds independently of the ratio d_1/d_2. A spatial gain of $\Theta(f/(\xi n))$ could result in a much larger f-gain, though. If $d_1/d_2 = O(1/\xi)$, however, the f-gain is also $O(f/(\xi n))$ as we will see. Therefore, let $d_1 = \alpha \cdot \phi/\xi$ with $\alpha = O(1)$ and still $f = \xi \cdot d_1^2 + d_2^2 = \phi^2$. Owing to the argumentation for the upper bound on the success probability of a step, we know that there is an n-sphere S with radius $r = \Theta(\phi/\xi)$ such that $\boldsymbol{c} \in S$ and $I := M \cap E \in S$, where I is the boundary of the hyperspherical cap $C \subset M$ lying inside E. Owing to the results for SPHERE, we know that $\mathsf{E}\big[\mathrm{dist}(\boldsymbol{c}', I) \cdot \mathbb{1}_{\{\boldsymbol{c}' \in C\}}\big] = O(r/n)$ for any choice of the scaling factor, i. e., even if the length of the mutation vector were magically chosen such that the expected distance of the selected search point $\boldsymbol{c}''$ from the center of S is minimized. In other words, we know that if a mutation hits inside E, its expected distance from E is $O(r/n) = O((\phi/\xi)/n)$ anyway. Thus, if we optimistically assume that the spatial gain were realized completely in S_1, i. e. completely on the ξ-weighted SPHERE$_{n/2}$, (so that $d_2' = d_2$, implying $d_2'' = d_2$), we obtain

$$\begin{aligned}
\mathsf{E}\big[\xi\, d_1''^2 + d_2''^2 \mid d_1/d_2 = O(1/\xi)\big] &\ge \xi\left(d_1 - O((\phi/\xi)/n)\right)^2 + d_2^2 \\
&= \xi\left(\alpha\phi/\xi - O((\phi/\xi)/n)\right)^2 + d_2^2 \\
&\ge \xi\left((\alpha\phi/\xi)^2 - 2\alpha(\phi/\xi)\cdot O((\phi/\xi)/n)\right) + d_2^2 \\
&= \xi\, d_1^2 - O(\phi^2/(\xi n)) + d_2^2
\end{aligned}$$

and hence,

$$\mathsf{E}[f'' \mid d_1/d_2 = O(1/\xi)] \;=\; \phi^2 - O(\phi^2/(\xi n)) \;=\; f - O(f/(\xi n)).$$

This upper bound on the expected f-gain of a step holds only for $d_1/d_2 = O(1/\xi)$, yet independently of (the distribution of) $|\boldsymbol{m}|$, which is converse to the lower bound. However, altogether we have proved the following:

Lemma 1. *Consider a step of the (1+1) ES. If $d_1/d_2 = O(1/\xi)$ in this step, then there exists a constant $\varepsilon > 0$ such that for n large enough $\mathsf{P}\{f' \le f\} \in [\varepsilon, 1/2-\varepsilon]$ iff $s = \Theta((\sqrt{f}/n)/\xi)$.*

If $d_1/d_2 = O(1/\xi)$ and $s = \Theta((\sqrt{f}/n)/\xi)$ in this step, then $\mathsf{E}[f - f''] = \Theta((f/n)/\xi)$, and furthermore, $f - f'' = \Omega((f/n)/\xi)$ with probability $\Omega(1)$.

4 Multi-step Behavior

The results just obtained imply that if $d_1/d_2 = O(1/\xi)$ during a phase of n steps (an observation phase of the 1/5-rule) and $s = \Theta((\sqrt{f}/n)/\xi)$, i.e. $\mathsf{P}\{f' \leq f\} \in [\varepsilon, 1/2 - \varepsilon]$ for a constant $\varepsilon > 0$, at the beginning of this phase, then we expect $\Theta(n)$ steps each of which reduces the f-value by $\Theta(f/(\xi n))$. By Chernoff bounds, there are $\Omega(n)$ such steps w. o. p., and thus, the f-value, and with it the approximation error, is reduced w. o. p. by an $\Theta(1/\xi)$-fraction in this phase. Consequently, after $\Theta(\xi)$ consecutive phases, w. o. p. the approximation error is halved – if during all these phases $d_1/d_2 = O(1/\xi)$. Since, up to now, the argumentation completely bases on the results for SPHERE, even the argumentation on the 1/5-rule can be adopted, which directly yields the following result (cf. Theorem 2 in (Jägersküpper, 2003) or Theorem 3 in (Jägersküpper, 2002)):

Theorem 1. *If $d_1/d_2 = O(1/\xi)$ in the complete optimization process and the initialization satisfies $s = \Theta((\sqrt{f(\boldsymbol{c})}/n)/\xi)$, then w. o. p. the number of steps/ f-evaluations to reduce the initial f-value/approximation error to a 2^{-t}-fraction, $t = poly(n)$, is $\Theta(t \cdot \xi \cdot n)$.*

Obviously, the assumption "$d_1/d_2 = O(1/\xi)$ in the complete optimization process" lacks any justification and is, therefore, objectionable. It must be replaced by a much weaker assumption on the starting conditions only. Thus, the crucial point in the analysis is the question why should the ratio d_1/d_2 remain $O(1/\xi)$ (once this is the case). This crucial question will be tackled in the remainder of this paper.

Let $\Delta_1 := d_1 - d_1'$ and $\Delta_2 := d_2 - d_2'$ denote the spatial gain of the mutant towards the origin in S_1 resp. S_2. Then d_1'/d_2' for the mutant is smaller than d_1/d_2 for its parent iff $\Delta_1/d_1 > \Delta_2/d_2$. Unfortunately, Δ_1 and Δ_2 correlate because $\boldsymbol{m_1}$ and $\boldsymbol{m_2}$ use the same scaling factor s, and furthermore, we must take selection into account since only certain combinations of Δ_1 and Δ_2 will be accepted. To see which combinations become accepted note that

$$f' = \xi\,(d_1 - \Delta_1)^2 + (d_2 - \Delta_2)^2 = \xi d_1^2 - \xi 2 d_1 \Delta_1 + \xi \Delta_1^2 + d_2^2 - 2 d_2 \Delta_2 + \Delta_2^2\,,$$

and hence,

$$f' \leq f \iff f' - f \leq 0 \iff -\xi 2 d_1 \Delta_1 + \xi \Delta_1^2 - 2 d_2 \Delta_2 + \Delta_2^2 \leq 0.$$

Let α be defined by $\alpha/\xi = d_1/d_2$. Then the latter inequality is equivalent to

$$-2\alpha d_2 \Delta_1 + \xi \Delta_1^2 - 2 d_2 \Delta_2 + \Delta_2^2 \leq 0 \iff -\alpha\Delta_1 + \frac{\xi \Delta_1^2}{2 d_2} \leq \Delta_2 - \frac{\Delta_2^2}{2 d_2}$$

$$\iff -\alpha\Delta_1 \left(1 - \frac{\Delta_1}{2 d_1}\right) \leq \Delta_2 \left(1 - \frac{\Delta_2}{2 d_2}\right) \qquad \text{(using } d_2 = \xi \cdot d_1/\alpha\text{)}$$

Thus, when using elitist selection, the mutant is accepted iff the last inequality holds. Note that whenever a mutation satisfying $-\alpha\Delta_1 > \Delta_2$ is accepted, then

$$1 - \frac{\Delta_1}{2 d_1} < 1 - \frac{\Delta_2}{2 d_2} \quad\Leftrightarrow\quad \frac{\Delta_1}{d_1} > \frac{\Delta_2}{d_2} \quad\Leftrightarrow\quad \Delta_1 > \frac{d_1}{d_2}\Delta_2 \quad\Leftrightarrow\quad \Delta_1 > \frac{\alpha}{\xi}\Delta_2,$$

implying that $\Delta_1 > 0$ and $\Delta_2 < 0$, and consequently, such a step surely results in $d_1''/d_2'' < d_1/d_2$, i.e. $\alpha'' < \alpha$. Hence, in the following we may concentrate on the accepted mutations for which $-\alpha\Delta_1 \le \Delta_2$.

So, let us assume for a moment that the mutant replaces/becomes the current individual iff $-\alpha\Delta_1 \le \Delta_2$. As Δ_{3-i}, $i \in \{1,2\}$, is random, $\mathsf{E}\big[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}}\big]$ is a random variable taking the value $\mathsf{E}\big[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le x\}}\big]$ whenever Δ_2 happens to take the value x. We are interested in $\mathsf{E}\big[\mathsf{E}\big[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}}\big]\big] = \mathsf{E}[d_i - d_i'']$, the expected reduction of the distance from the optimum in S_i in a step, and $\mathsf{E}[d_1'']/\mathsf{E}[d_2''] \le d_1/d_2$, i.e. we "expect" $\alpha'' \le \alpha$, iff

$$\begin{array}{rcl} \mathsf{E}\big[\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}}\big]\big]/d_1 & \ge & \mathsf{E}\big[\mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}}\big]\big]/d_2 \\ \iff \xi \cdot \mathsf{E}\big[\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}}\big]\big] & \ge & \alpha \cdot \mathsf{E}\big[\mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}}\big]\big]. \end{array}$$

In order to prove that this inequality holds for $\alpha = O(1)$, we aim at a lower bound on $\mathsf{E}\big[\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}}\big]\big]$ and an upper bound on $\mathsf{E}\big[\mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}}\big]\big]$ in the following. Note that

$$\begin{array}{rcl} \mathsf{E}\big[\mathsf{E}\big[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}}\big]\big] & = & \mathsf{E}\big[\mathsf{E}\big[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_i < 0\}}\big] \cdot \mathbb{1}_{\{\Delta_{3-i} < 0\}}\big] + \\ & & \mathsf{E}\big[\mathsf{E}\big[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_i < 0\}}\big] \cdot \mathbb{1}_{\{\Delta_{3-i} \ge 0\}}\big] + \\ & & \mathsf{E}\big[\mathsf{E}\big[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_i \ge 0\}}\big] \cdot \mathbb{1}_{\{\Delta_{3-i} < 0\}}\big] + \\ & & \mathsf{E}\big[\mathsf{E}\big[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_i \ge 0\}}\big] \cdot \mathbb{1}_{\{\Delta_{3-i} \ge 0\}}\big] \end{array}$$

and that $\mathsf{E}\big[\mathsf{E}\big[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_i < 0\}}\big] \cdot \mathbb{1}_{\{\Delta_{3-i} < 0\}}\big] = 0$ since the three indicator inequalities describe the empty set. Since $\Delta_1, \Delta_2 \ge 0$ implies $-\alpha\Delta_1 \le \Delta_2$,

$$\begin{array}{rcl} \mathsf{E}\big[\mathsf{E}\big[\Delta_i \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}} \mathbb{1}_{\{\Delta_i \ge 0\}}\big] \mathbb{1}_{\{\Delta_{3-i} \ge 0\}}\big] & = & \mathsf{E}\big[\mathsf{E}\big[\Delta_i \mathbb{1}_{\{\Delta_i \ge 0\}}\big] \cdot \mathbb{1}_{\{\Delta_{3-i} \ge 0\}}\big] \\ & = & \mathsf{E}\big[\Delta_i \mathbb{1}_{\{\Delta_i \ge 0\}}\big] \cdot \mathsf{P}\{\Delta_{3-i} \ge 0\}. \end{array}$$

As we need a lower bound on $\mathsf{E}\big[\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}}\big]\big]$, we may pessimistically assume that $\Delta_1 = -x/\alpha$ whenever Δ_2 happens to equal x. By this assumption,

$$\begin{array}{rl} & \mathsf{E}\big[\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_1 < 0\}}\big] \cdot \mathbb{1}_{\{\Delta_2 \ge 0\}}\big] \\ \ge & -\mathsf{E}\big[\mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_2 \ge 0\}}\big] \cdot \mathbb{1}_{\{\Delta_1 < 0\}}\big]/\alpha, \end{array}$$

$$\begin{array}{rl} & \mathsf{E}\big[\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_1 \ge 0\}}\big] \cdot \mathbb{1}_{\{\Delta_2 < 0\}}\big] \\ \ge & -\mathsf{E}\big[\mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_2 < 0\}}\big] \cdot \mathbb{1}_{\{\Delta_1 \ge 0\}}\big]/\alpha. \end{array}$$

All in all, we have

$$\begin{array}{rcl} \mathsf{E}\big[\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}}\big]\big] & \ge & \mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \ge 0\}}\big] \cdot \mathsf{P}\{\Delta_2 \ge 0\} \\ & & -\,\mathsf{E}\big[\mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_2 \ge 0\}}\big] \cdot \mathbb{1}_{\{\Delta_1 < 0\}}\big]/\alpha \\ & & -\,\mathsf{E}\big[\mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_2 < 0\}}\big] \cdot \mathbb{1}_{\{\Delta_1 \ge 0\}}\big]/\alpha, \end{array}$$

$$\begin{aligned}\mathsf{E}\big[\mathsf{E}\big[\Delta_2\cdot\mathbb{1}_{\{-\alpha\Delta_1\le\Delta_2\}}\big]\big] \;&=\; \mathsf{E}\big[\Delta_2\cdot\mathbb{1}_{\{\Delta_2\ge0\}}\big]\cdot\mathsf{P}\{\Delta_1\ge0\}\\ &\quad+\mathsf{E}\big[\mathsf{E}\big[\Delta_2\cdot\mathbb{1}_{\{-\alpha\Delta_1\le\Delta_2\}}\cdot\mathbb{1}_{\{\Delta_2\ge0\}}\big]\cdot\mathbb{1}_{\{\Delta_1<0\}}\big]\\ &\quad+\mathsf{E}\big[\mathsf{E}\big[\Delta_2\cdot\mathbb{1}_{\{-\alpha\Delta_1\le\Delta_2\}}\cdot\mathbb{1}_{\{\Delta_2<0\}}\big]\cdot\mathbb{1}_{\{\Delta_1\ge0\}}\big].\end{aligned}$$

Recall that we want to show that for some $\alpha=O(1)$

$$\xi\cdot\mathsf{E}\big[\mathsf{E}\big[\Delta_1\cdot\mathbb{1}_{\{-\alpha\Delta_1\le\Delta_2\}}\big]\big] \;\ge\; \alpha\cdot\mathsf{E}\big[\mathsf{E}\big[\Delta_2\cdot\mathbb{1}_{\{-\alpha\Delta_1\le\Delta_2\}}\big]\big],$$

and note that $\mathsf{E}\big[\Delta_1\cdot\mathbb{1}_{\{\Delta_1\ge0\}}\big]\cdot\mathsf{P}\{\Delta_2\ge0\}$ and $\mathsf{E}\big[\Delta_2\cdot\mathbb{1}_{\{\Delta_2\ge0\}}\big]\cdot\mathsf{P}\{\Delta_1\ge0\}$ are of the same order when $\mathsf{P}\{\Delta_2\ge0\}$ and $\mathsf{P}\{\Delta_1\ge0\}$ are $\Omega(1)$, respectively. Consequently, since $\xi=\omega(1)$, for the above inequality to hold for n large enough, it would be sufficient that

$$\begin{aligned}&\mathsf{E}\big[\mathsf{E}\big[\Delta_2\cdot\mathbb{1}_{\{-\alpha\Delta_1\le\Delta_2\}}\cdot\mathbb{1}_{\{\Delta_2\ge0\}}\big]\cdot\mathbb{1}_{\{\Delta_1<0\}}\big]\\ +\,&\mathsf{E}\big[\mathsf{E}\big[\Delta_2\cdot\mathbb{1}_{\{-\alpha\Delta_1\le\Delta_2\}}\cdot\mathbb{1}_{\{\Delta_2<0\}}\big]\cdot\mathbb{1}_{\{\Delta_1\ge0\}}\big] \;\le\; 0\end{aligned} \tag{1}$$

because then we would have

$$\begin{aligned}\mathsf{E}\big[\mathsf{E}\big[\Delta_1\cdot\mathbb{1}_{\{-\alpha\Delta_1\le\Delta_2\}}\big]\big] \;&\ge\; \mathsf{E}\big[\Delta_1\cdot\mathbb{1}_{\{\Delta_1\ge0\}}\big]\cdot\mathsf{P}\{\Delta_2\ge0\} \quad\text{and}\\ \mathsf{E}\big[\mathsf{E}\big[\Delta_2\cdot\mathbb{1}_{\{-\alpha\Delta_1\le\Delta_2\}}\big]\big] \;&\le\; \mathsf{E}\big[\Delta_2\cdot\mathbb{1}_{\{\Delta_2\ge0\}}\big]\cdot\mathsf{P}\{\Delta_1\ge0\}.\end{aligned}$$

Concerning the expected spatial gain in S_2, however, we are going to use the trivial upper bound $\mathsf{E}\big[\mathsf{E}\big[\Delta_2\cdot\mathbb{1}_{\{-\alpha\Delta_1\le\Delta_2\}}\big]\big]\le\mathsf{E}\big[\Delta_2\cdot\mathbb{1}_{\{\Delta_2\ge0\}}\big]$, and thus, we concentrate on a lower bound on the expected spatial gain in S_1 in the following. Therefore, we prove next that inequality (1) holds for $\alpha=O(1)$ at least if the actual length of $\boldsymbol{m_2}$ differs by no more than a constant factor from $\bar{\ell_1}$, the expected length of $\boldsymbol{m_1}$.

Lemma 2. *If* $\mathsf{P}\{\Delta_1\ge0\}=\Omega(1)$ *and* $|\boldsymbol{m_2}|=\Theta(\bar{\ell_1})$*, there exists a constant* α^* *such that for* n *large enough inequality (1) on this page holds for all* $\alpha\ge\alpha^*$.

The proof can be found in Appendix A. Note that $\bar{\ell_1}=\bar{\ell_2}$ in our scenario. We know (cf. Section 2) that

$$\mathsf{P}\Big\{\big||\boldsymbol{m_2}|-\bar{\ell_2}\big|\ge(\sqrt{3}-1)\cdot\bar{\ell_2}\Big\}\le\Big((\sqrt{3}-1)^2\cdot2\cdot(n-1)\Big)^{-1}<(n-1)^{-1},$$

and thus, the condition "$|\boldsymbol{m_2}|=\Theta(\bar{\ell_1})$" is not met only with probability $O(1/n)$. Whether or not this condition is met, trivially $\Delta_1\ge-|\boldsymbol{m_1}|$, and consequently, $\mathsf{E}\big[\mathsf{E}\big[\Delta_1\cdot\mathbb{1}_{\{-\alpha\Delta_1\le\Delta_2\}}\big]\big]\ge-\bar{\ell_1}$. Applying this rough bound only in the case of $\big||\boldsymbol{m_2}|-\bar{\ell_1}\big|>(\sqrt{3}-1)\cdot\bar{\ell_1}$ and $(\Delta_1,\Delta_2)\in\mathbb{R}_{<0}\times\mathbb{R}_{\ge0}\cup\mathbb{R}_{\ge0}\times\mathbb{R}_{<0}$, the preceding lemma reads: if $\mathsf{P}\{\Delta_1\ge0\}=\Omega(1)$ then for $\alpha\ge\alpha^*$

$$\mathsf{E}\big[\mathsf{E}\big[\Delta_1\cdot\mathbb{1}_{\{-\alpha\Delta_1\le\Delta_2\}}\big]\big] \;\ge\; \mathsf{E}\big[\Delta_1\cdot\mathbb{1}_{\{\Delta_1\ge0\}}\big]\cdot\mathsf{P}\{\Delta_2\ge0\}-\frac{\bar{\ell_1}}{n-1}.$$

Next we will see that this additive error term vanishes in situations that arise due to the 1/5-rule.

Lemma 3. *If* $\mathsf{P}\{\Delta_1 \geq 0\}$ *and* $\mathsf{P}\{\Delta_2 \geq 0\}$ *are* $\Omega(1)$*, respectively, there exists a constant* α^* *such that for* $\alpha \geq \alpha^*$ *and* n *large enough*

$$\mathsf{E}\big[\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{f' \leq f\}}\big]\big] \quad \geq \quad \mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geq 0\}}\big] \cdot \mathsf{P}\{\Delta_2 \geq 0\}/\,2.$$

Proof. Recall that $f' \leq f \wedge -\alpha\Delta_1 > \Delta_2$ implies $\Delta_1 > 0 > \Delta_2$. Consequently, all (Δ_1, Δ_2)-tuples zeroed out by $\mathbb{1}_{\{-\alpha\Delta_1 \leq \Delta_2\}}$, but kept by $\mathbb{1}_{\{f' \leq f\}}$ are in $\mathbb{R}_{>0} \times \mathbb{R}_{<0}$. Analogously, $f' > f \wedge -\alpha\Delta_1 \leq \Delta_2$ implies $\Delta_1 < 0 < \Delta_2$ so that all (Δ_1, Δ_2)-tuples kept by $\mathbb{1}_{\{-\alpha\Delta_1 \leq \Delta_2\}}$, but zeroed out by $\mathbb{1}_{\{f' \leq f\}}$ are in $\mathbb{R}_{<0} \times \mathbb{R}_{>0}$. Hence,

$$\begin{array}{rcl} \mathsf{E}\big[\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{f' \leq f\}}\big]\big] & \geq & \mathsf{E}\big[\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leq \Delta_2\}}\big]\big] \\ \big(\text{and } \mathsf{E}\big[\mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{f' \leq f\}}\big]\big] & \leq & \mathsf{E}\big[\mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leq \Delta_2\}}\big]\big]\big). \end{array}$$

As $\mathsf{P}\{\Delta_1 \geq 0\} = \Omega(1)$ implies $\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geq 0\}}\big] = \Omega(\bar{\ell_1}/\sqrt{n})$ (cf. the results restated in Section 2), the error term $\bar{\ell_1}/(n-1)$ is by an $O(1/\sqrt{n})$-factor smaller than $\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geq 0\}}\big] \cdot \mathsf{P}\{\Delta_2 \geq 0\} = \Omega(\bar{\ell_1}/\sqrt{n}) \cdot \Omega(1)$. Finally, for n large enough $1 - O(1/\sqrt{n}) \geq 1/2$. □

Recall: we expect $\alpha'' = \alpha$ iff $\xi \cdot \mathsf{E}\big[\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{f' \leq f\}}\big]\big] = \alpha \cdot \mathsf{E}\big[\mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{f' \leq f\}}\big]\big]$ or, equivalently, iff $\mathsf{E}\big[\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{f' \leq f\}}\big]\big]/d_1 = \mathsf{E}\big[\mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{f' \leq f\}}\big]\big]/d_2$. Thus there exists a distinct α_0 such that there is no drift w. r. t. the ratio d_1/d_2, i. e., this ratio becomes steady-state. Then for $\alpha < \alpha_0$, α is more likely to increase than to decrease, and for $\alpha > \alpha_0$, α is more likely to decrease than to increase.

Since $\mathsf{E}\big[\mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{f' \leq f\}}\big]\big] \leq \mathsf{E}\big[\mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leq \Delta_2\}}\big]\big] \leq \mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{\Delta_2 \geq 0\}}\big]$ and $\xi = \omega(1)$, we have $\xi \cdot \mathsf{P}\{\Delta_2 \geq 0\}/2 \geq \alpha^*$ for n large enough if $\mathsf{P}\{\Delta_2 \geq 0\} = \Omega(1)$, and hence, $\alpha_0 \leq \alpha^* = O(1)$ under the conditions of Lemma 3. Besides, the 1/5-rule just ensures these conditions as long as $d_1 = O(d_2)$. For the same reasons, there exists $\alpha_\downarrow > \alpha_0$ such that $\xi \cdot \mathsf{E}\big[\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{f' \leq f\}}\big]\big] \geq \mathbf{2} \cdot \alpha \cdot \mathsf{E}\big[\mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{f' \leq f\}}\big]\big]$ (for n large enough) and $\alpha_\downarrow = O(1)$ again under the conditions of Lemma 3. Thus, when $\alpha \geq \alpha_\downarrow$ there is a drift towards smaller α; more formally:

Lemma 4. *Let the scaling factor* s *be fixed. If* $\mathsf{P}\{\Delta_1 \geq 0\}$ *and* $\mathsf{P}\{\Delta_2 \geq 0\}$ *are* $\Omega(1)$*, respectively, there exists a constant* $\alpha_\downarrow$ *such that for* n *large enough, if in the* i^{th} *step* $\alpha^{[i]} \geq \alpha_\downarrow$ *(yet* $\alpha^{[i]} = O(\xi)$*), then w. o. p. after at most* $n^{0.3}$ *steps the search is located at a point for which* $\alpha < \alpha^{[i]}$*, and furthermore, w. o. p.* $\alpha \leq \alpha^{[i]} + O(\alpha^{[i]}/n^{0.6})$ *in all intermediate steps.*

The proof can be found in Appendix B. Since the 1/5-rule keeps the scaling factor unchanged for n steps, we can virtually partition each such observation phase in $n/n^{0.3} = n^{0.7}$ sub-phases to each of which this lemma applies. Since $O(\alpha^{[i]}/n^{0.6}) \leq \alpha^{[i]}$ for n large enough, the preceding lemma tells us that, when starting at a point with $\alpha^{[0]} = O(1)$, i. e. $d_1^{[0]}/d_2^{[0]} = O(1/\xi)$, then α will be upper bounded by $2 \cdot \max\{\alpha^{[0]}, \alpha_\downarrow\} = O(1)$ w. o. p. for any polynomial number of steps. Incorporating these new insights into the argumentation for the 1/5-rule known from the analysis of Sphere finally enables us to replace the objectionable condition "$d_1/d_2 = O(1/\xi)$ in the complete optimization process" in Theorem 1 by "$d_1/d_2 = O(1/\xi)$ for the initial search point" – yielding the main result on the rutime of the (1+1) ES on the quadratic forms considered:

Theorem 2. *If the initialization satisfies $s = \Theta((\sqrt{f(\boldsymbol{c})}/n)/\xi)$ and $d_1/d_2 = O(1/\xi)$, then w. o. p. the number of steps/f-evaluations to reduce the initial approximation error/f-value to a 2^{-t}-fraction, $t = poly(n)$, is $\Theta(t \cdot \xi \cdot n)$.*

Naturally, one might ask what happens if the optimization starts at a point for which d_1 is not $O(d_2/\xi)$. A closer look at the argumentation in the proof of the preceding lemma reveals that the same argumentation results in the proof of the existence of another constant $\alpha_{\Downarrow} > \alpha_{\downarrow}$ such that the drift towards smaller α is that strong when $\alpha \geq \alpha_{\Downarrow}$ that w. o. p. α drops by a constant fraction within at most n steps:

Lemma 5. *Let the scaling factor s be fixed. If $\mathsf{P}\{\Delta_1 \geq 0\}$, $1/2 - \mathsf{P}\{\Delta_1 \geq 0\}$, $\mathsf{P}\{\Delta_2 \geq 0\}$ are $\Omega(1)$, respectively, then there exists a constant $\alpha_{\Downarrow}$ such that for n large enough: if in the i^{th} step $\alpha^{[i]} \geq \alpha_{\Downarrow}$ (yet $\alpha^{[i]} = O(\xi)$, i. e. $d_1 = O(d_2)$), then w. o. p. after at most n steps the search is located at a point with $\alpha \leq \alpha^{[i]} - \Omega(\alpha^{[i]})$.*

See Appendix C for the proof. Finally, this lemma shows that α drops very quickly if the lemma's conditions are met. Again utilizing the results for SPHERE, it is simple to check that these conditions are met when d_1 is $O(d_2)$ (and $\Omega(d_2/\xi)$, of course). If d_1 is not $O(d_2)$, for instance if we start at a point of steepest descent (w. r. t. all points of a fixed f-value), i. e. $d_2 = 0$ so that $f = \xi d_1^2$, then a simple argumentation using rough bounds on Δ_1 and Δ_2 yields that d_1/d_2 drops even faster than in situations covered by the preceding lemma – which is hardly surprising since the (expected) spatial gain of a step in S_1 (on the ξ-weighted SPHERE$_{n/2}$) is negative whereas the one in S_2 is positive.

5 Conclusion

Based on the results on how the (1+1) ES minimizes the well-known SPHERE-function, we have extended these results to a broader class of functions consisting of certain positive definite quadratic forms. The main insight of the results presented is that Gaussian mutations adapted by the 1/5-rule result in the optimization process to stabilize such that the trajectory of the evolving search point takes course very close to the gentlest descent of the ellipsoidal fitness landscape. However, more insight into how EAs for continuous optimization work is gained, contributing to building an algorithmic EA-theory for continuous search spaces.

Naturally, the results carry over to functions that are translations of the considered functions. Furthermore, the argumentation presented here yields that for arbitrary positive definite quadratic forms – which we may assume to be of the form $f_n(\boldsymbol{x}) = \sum_{i=1}^{n} \xi_i \cdot {x_i}^2$ with $\xi_n \geq \cdots \geq \xi_1 > 0$ as we have seen – the number of steps to halve the function value is $O(n \cdot \xi_n/\xi_1)$. This is due to the maximum curvature being upper bounded by $(\xi_n/\xi_1)/\sqrt{f}$ so that the radius of the hypersphere S is at least $\sqrt{f} \cdot \xi_1/\xi_n$. As a direct consequence, we obtain a $\Theta(n)$-bound for functions where all the ξ_is are of the same order, i. e. $\xi_n = \Theta(\xi_1)$. This is the reason why ξ was chosen to be $\omega(1)$.

Acknowledgments

I would like to thank two of the three reviewers for their helpful comments, in particular for pointing out that applying differential geometry in the analysis of fitness landscapes was already proposed by Beyer (1994).

A Proof of Lemma 2

> "If $\mathsf{P}\{\Delta_1 \geq 0\} = \Omega(1)$ and $|\boldsymbol{m_2}| = \Theta(\bar{\ell_1})$, there exists a constant α^* such that for n large enough inequality (1) on page 272 holds for all $\alpha \geq \alpha^*$."

Let us assume for a moment that the distribution of $|\boldsymbol{m_2}|$ were concentrated at a certain ℓ_2, and let "$\mathsf{D}\{\cdot\}$" denote the density of an event. Then

$$\begin{aligned} & \mathsf{E}\big[\mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leq \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_2 \geq 0\}}\big] \cdot \mathbb{1}_{\{\Delta_1 < 0\}}\big] \\ = & \int_0^{\ell_2} x \cdot \mathsf{D}\{\Delta_2 = x\} \cdot \mathsf{P}\{-x/\alpha \leq \Delta_1 < 0\} \,\mathrm{d}x \qquad \text{and} \end{aligned}$$

$$\begin{aligned} & \mathsf{E}\big[\mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leq \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_2 < 0\}}\big] \cdot \mathbb{1}_{\{\Delta_1 \geq 0\}}\big] \\ = & \int_{-\ell_2}^{0} y \cdot \mathsf{D}\{\Delta_2 = y\} \cdot \mathsf{P}\{\Delta_1 \geq -y/\alpha\} \,\mathrm{d}y \\ = & \int_0^{\ell_2} -x \cdot \mathsf{D}\{\Delta_2 = -x\} \cdot \mathsf{P}\{\Delta_1 \geq x/\alpha\} \,\mathrm{d}x\,. \end{aligned}$$

We know from the analysis of SPHERE that for $x \in [0, \ell_2)$

$$\mathsf{D}\{\Delta_2 = x\} \;<\; \frac{\Psi_n}{\ell_2} \cdot (1 - (x/\ell_2)^2)^{(n-3)/2} \;<\; \mathsf{D}\{\Delta_2 = -x\}$$

(with $\Psi_n := \pi^{-1/2} \cdot \Gamma(n/2) \,/\, \Gamma(n/2 - 1/2) = \Theta(\sqrt{n})$, where "$\Gamma$" denotes the well-known Gamma function).

Thus, the LHS of (1) on page 272 is smaller than

$$\begin{aligned} & \int_0^{\ell_2} x \cdot \frac{\Psi_n}{\ell_2} \cdot (1 - (x/\ell_2)^2)^{(n-3)/2} \cdot \mathsf{P}\{-x/\alpha \leq \Delta_1 < 0\} \,\mathrm{d}x \\ & - \int_0^{\ell_2} x \cdot \frac{\Psi_n}{\ell_2} \cdot (1 - (x/\ell_2)^2)^{(n-3)/2} \cdot \mathsf{P}\{\Delta_1 \geq x/\alpha\} \,\mathrm{d}x \\ = & \int_0^{\ell_2} x \,\frac{\Psi_n}{\ell_2}\, (1 - (x/\ell_2)^2)^{(n-3)/2} \big(\mathsf{P}\{-x/\alpha \leq \Delta_1 < 0\} - \mathsf{P}\{\Delta_1 \geq x/\alpha\}\big) \,\mathrm{d}x. \end{aligned}$$

Let $\Phi\colon [0, \ell_2] \to [-1, 1]$ be defined by $\Phi(y) := \mathsf{P}\{-y \leq \Delta_1 < 0\} - \mathsf{P}\{\Delta_1 \geq y\}$.

Hence,

$$\int_0^{\ell_2} x \cdot \frac{\Psi_n}{\ell_2} \cdot (1-(x/\ell_2)^2)^{(n-3)/2} \cdot \Phi(x/\alpha)\,\mathrm{d}x \quad \leq \quad 0$$

implies the inequality (1). Note that, obviously, $\mathsf{P}\{-0 \leq \Delta_1 < 0\} = 0$ and, by assumption, $\mathsf{P}\{\Delta_1 \geq 0\} = \Omega(1)$. Since, $\mathsf{P}\{\Delta_1 \geq y\}$ decreases monotonically, whereas $\mathsf{P}\{-y \leq \Delta_1 < 0\}$ increases monotonically when y grows, $\Phi(y)$ is monotone increasing for $0 \leq y \leq \min\{\ell_1, \ell_2\}$ and equals $\mathsf{P}\{\Delta_1 < 0\}$ for $y \geq \ell_1$. Furthermore, if ε denotes an arbitrary constant with $0 < \varepsilon < \mathsf{P}\{\Delta_1 \geq 0\}$, then $\mathsf{P}\{\Delta_1 \geq y\} = \varepsilon$ implies $y = \Theta(\bar{\ell_1}/\sqrt{n})$. Analogously, if $0 < \varepsilon < \mathsf{P}\{\Delta_1 < 0\}$, then $\mathsf{P}\{-y \leq \Delta_1 < 0\} = \varepsilon$ implies $y = \Theta(\bar{\ell_1}/\sqrt{n})$. Thus, there exists $\breve{y} = \kappa \cdot \bar{\ell_1}/\sqrt{n-1}$ with $\kappa = \Theta(1)$ such that $\mathsf{P}\{\Delta_1 \geq \breve{y}\} = \mathsf{P}\{-\breve{y} \leq \Delta_1 < 0\}$, i. e., $\Phi(\breve{y}) = 0$, and hence, the inequality to be shown reads

$$\begin{aligned} & -\frac{\Psi_n}{\ell_2} \int_0^{\alpha \cdot \breve{y}} x \cdot (1-(x/\ell_2)^2)^{(n-3)/2} \cdot \Phi(x/\alpha)\,\mathrm{d}x \\ \geq \quad & \frac{\Psi_n}{\ell_2} \int_{\alpha \cdot \breve{y}}^{\ell_2} x \cdot (1-(x/\ell_2)^2)^{(n-3)/2} \cdot \Phi(x/\alpha)\,\mathrm{d}x. \end{aligned} \tag{2}$$

For the RHS we have, using $(1-a/(n-1))^{(n-1)/2} \leq \mathrm{e}^{-a/2}$ for $n-1 > a > 0$,

$$\begin{aligned} & \int_{\alpha \cdot \breve{y}}^{\ell_2} x \cdot (1-(x/\ell_2)^2)^{(n-3)/2} \cdot \Phi(x/\alpha)\,\mathrm{d}x \\ \leq \quad & \int_{\alpha \cdot \breve{y}}^{\ell_2} x \cdot (1-(x/\ell_2)^2)^{(n-3)/2} \cdot 1\,\mathrm{d}x \\ = \quad & \left[\frac{-\ell_2^2}{2} \cdot \frac{(1-(x/\ell_2)^2)^{(n-1)/2}}{(n-1)/2}\right]_{\alpha \cdot \breve{y}}^{\ell_2} \\ = \quad & 0 - \left(\frac{-\ell_2^2}{n-1} \cdot (1-(\alpha \cdot \breve{y}/\ell_2)^2)^{(n-1)/2}\right) \\ = \quad & \frac{\ell_2^2}{n-1} \cdot (1-(\alpha \cdot \breve{y}/\ell_2)^2)^{(n-1)/2} \\ \leq \quad & \frac{\ell_2^2}{n-1} \cdot \left(1-(\alpha \cdot \kappa \cdot \bar{\ell_1}/\ell_2)^2/(n-1)\right)^{(n-1)/2} \\ \leq \quad & \frac{\ell_2^2}{n-1} \cdot \mathrm{e}^{-(\alpha \cdot \kappa \cdot \bar{\ell_1}/\ell_2)^2/2} \qquad \text{if } n-1 > \left(\alpha \cdot \kappa \cdot \frac{\bar{\ell_1}}{\ell_2}\right)^2. \end{aligned}$$

For the LHS of (2) note that, by the same arguments, there exists $\ddot{y} = \tau \cdot \bar{\ell_1}/\sqrt{n-1}$ with $\tau = \Theta(1)$ such that $\mathsf{P}\{\Delta_1 \geq \ddot{y}\} = 2 \cdot \mathsf{P}\{-\ddot{y} \leq \Delta_1 < 0\}$, and thus, for $0 \leq y \leq \ddot{y}$ we have $\mathsf{P}\{\Delta_1 \geq y\} \geq 2 \cdot \mathsf{P}\{-y \leq \Delta_1 < 0\}$, i. e., $-\Phi(y) \geq p := \mathsf{P}\{\Delta_1 \geq \ddot{y}\}/2 = \Omega(1)$. Hence,

$$
\begin{aligned}
& - \int_0^{\alpha\cdot\check{y}} x\cdot(1-(x/\ell_2)^2)^{(n-3)/2}\cdot\Phi(x/\alpha)\,\mathrm{d}x \\
\geq\ & \int_0^{\alpha\cdot\ddot{y}} x\cdot(1-(x/\ell_2)^2)^{(n-3)/2}\cdot p\,\mathrm{d}x \\
=\ & p\cdot\left[\frac{-\ell_2^2}{2}\cdot\frac{(1-(x/\ell_2)^2)^{(n-1)/2}}{(n-1)/2}\right]_0^{\alpha\cdot\ddot{y}} \\
=\ & p\cdot\frac{-\ell_2^2}{n-1}\cdot\left(\left(1-(\alpha\cdot\ddot{y}/\ell_2)^2\right)^{(n-1)/2}-1\right) \\
=\ & p\cdot\frac{\ell_2^2}{n-1}\cdot\left(1-\left(1-\frac{(\alpha\cdot\tau\cdot\bar{\ell}_1/\ell_2)^2}{n-1}\right)^{(n-1)/2}\right) \\
\geq\ & p\cdot\frac{\ell_2^2}{n-1}\cdot\left(1-\mathrm{e}^{-(\alpha\cdot\tau\cdot\bar{\ell}_1/\ell_2)^2/2}\right) \qquad \text{if } n-1>\left(\alpha\cdot\tau\cdot\frac{\bar{\ell}_1}{\ell_2}\right)^2.
\end{aligned}
$$

All in all, we have broken it down into the inequality

$$p\cdot\frac{\ell_2^2}{n-1}\cdot\left(1-\mathrm{e}^{-(\alpha\cdot\tau\cdot\bar{\ell}_1/\ell_2)^2/2}\right) \quad\geq\quad \frac{\ell_2^2}{n-1}\cdot\mathrm{e}^{-(\alpha\cdot\kappa\cdot\bar{\ell}_1/\ell_2)^2/2}.$$

Since p, τ, and κ are $\Theta(1)$, it is finally obvious that $\alpha = O(1)$ can be chosen large enough for this inequality to hold for n large enough if $\bar{\ell}_1/\ell_2 = \Theta(1)$, i. e. $\ell_2 = \Theta(\bar{\ell}_1)$.

B Proof of Lemma 4

> "Let the scaling factor s be fixed. If $\mathsf{P}\{\Delta_1 \geq 0\}$ and $\mathsf{P}\{\Delta_2 \geq 0\}$ are $\Omega(1)$, respectively, there exists a constant $\alpha_\downarrow$ such that for n large enough, if in the i^{th} step $\alpha^{[i]} \geq \alpha_\downarrow$ (yet $\alpha^{[i]} = O(\xi)$), then w. o. p. after at most $n^{0.3}$ steps the search is located at a point for which $\alpha < \alpha^{[i]}$, and furthermore, w. o. p. $\alpha \leq \alpha^{[i]} + O(\alpha^{[i]}/n^{0.6})$ in all intermediate steps."

We begin by proving the second claim. Let us assume that, starting with the i^{th} step, $\alpha \geq \alpha^{[i]}$ for $k \leq n^{0.3}$ steps. Recall that, due to elitist selection, the f-value is non-increasing. As $d_2 > d_2^{[i]}$ and $f \leq f^{[i]}$ implies $d_1 < d_1^{[i]}$, which again implies $\alpha/\xi = d_1/d_2 < d_1^{[i]}/d_2^{[i]} = \alpha^{[i]}/\xi$, we have just proved that (surely) $d_2 \leq d_2^{[i]}$ in these k steps, respectively. Since, irrespective of the adaptation of the length of an isotropic mutation, in a step w. o. p. $\Delta_2 = O(d_2/n^{0.9})$, in all $k \leq n^{0.3}$ steps w. o. p. $d_2 \geq d_2^{[i]} - k\cdot O(d_2^{[i]}/n^{0.9}) \geq d_2^{[i]} - O(d_2^{[i]}/n^{0.6})$, i. e., $d_2 = d_2^{[i]}(1-\psi)$ for some $\psi = O(n^{-0.6})$, respectively. Concerning an upper bound on d_1, we have

$$f \;=\; \xi d_1^2 + d_2^2 \;=\; \xi d_1^2 + \left(d_2^{[i]} - \psi d_2^{[i]}\right)^2 \;\leq\; f^{[i]} \;=\; \xi {d_1^{[i]}}^2 + {d_2^{[i]}}^2,$$

and hence

$$\begin{aligned}
& \xi d_1^2 \leq \xi {d_1^{[i]}}^2 + (2\psi - \psi^2){d_2^{[i]}}^2 \\
\Leftrightarrow \quad & d_1^2 \leq {d_1^{[i]}}^2 + (2\psi - \psi^2)\frac{{d_2^{[i]}}^2}{\xi} = {d_1^{[i]}}^2 + (2\psi - \psi^2)\frac{{d_1^{[i]}}^2}{\alpha^{[i]}} \\
& = {d_1^{[i]}}^2 \left(1 + \frac{\psi(2-\psi)}{\alpha^{[i]}}\right)
\end{aligned}$$

Since $\psi(2-\psi)/\alpha^{[i]}$ is $O(\psi)$, i. e. $O(n^{-0.6})$, we finally get that in all k steps

$$\frac{\alpha}{\xi} = \frac{d_1}{d_2} \leq \frac{d_1^{[i]}}{d_2^{[i]}} \cdot \frac{\sqrt{1+O(n^{-0.6})}}{1-O(n^{-0.6})} = \frac{\alpha^{[i]}}{\xi} \cdot (1+O(n^{-0.6})).$$

Now we are ready for the proof of the lemma's first claim. Therefore, assume that $\alpha \geq \alpha^{[i]} \geq \alpha_{\downarrow}$ for $n^{0.3}+1$ steps. We are going to show that the probability of observing such a sequence of steps is exponentially small. Note that, since w. o. p. $d_2 \geq d_2^{[i]}(1-\psi)$ as we have seen, our assumption implies that also w. o. p. $d_1 \geq d_1^{[i]}(1-\psi)$, i. e., w. o. p. $d_1 = d_1^{[i]} - O(d_1^{[i]}/n^{0.6})$ in all $n^{0.3}$ steps. Let $X_j^{[k]}$, $j \in \{1,2\}$, denote the RV $\Delta_j \cdot \mathbb{1}_{\{f' \leq f\}}$ in the $(i-1+k)^{\text{th}}$ step (so that $\mathsf{E}[X_j] = \mathsf{E}\big[\mathsf{E}\big[\Delta_j \cdot \mathbb{1}_{\{f' \leq f\}}\big]\big]$). Then, according to the arguments preceding the lemma, for $1 \leq k \leq n^{0.3}$, $\mathsf{E}\Big[X_1^{[k]}\Big]/d_1^{[k]} \geq 2 \cdot \mathsf{E}\Big[X_2^{[k]}\Big]/d_2^{[k]}$, i. e.,

$$\xi \cdot \mathsf{E}\Big[X_1^{[k]}\Big] \geq 2 \cdot \alpha^{[k]} \cdot \mathsf{E}\Big[X_2^{[k]}\Big] \geq 2 \cdot \alpha^{[i]} \cdot \mathsf{E}\Big[X_2^{[k]}\Big].$$

Let $S_j^{[k]} := X_j^{[1]} + \cdots + X_j^{[k]}$ denote the total gain of k steps w. r. t. to d_j. By linearity of expectation, $\mathsf{E}\Big[S_1^{[k]}\Big]/d_1^{[i]} \geq 2 \cdot \mathsf{E}\Big[S_2^{[k]}\Big]/d_2^{[i]}$ for $1 \leq k \leq n^{0.3}$; however, the goal is to show that $\mathsf{P}\Big\{S_1^{[k]}/d_1^{[i]} \leq S_2^{[k]}/d_2^{[i]} \text{ for } 1 \leq k \leq n^{0.3}\Big\}$ is exponentially small.

Therefore, we will assume the worst case (w. r. t. to the analysis, i. e. the best case w. r. t. the chance of observing such a sequence) that $\mathsf{E}\Big[X_1^{[k]}\Big]/d_1^{[i]} = 2 \cdot \mathsf{E}\Big[X_2^{[k]}\Big]/d_2^{[i]}$ in each step. To see that this is in fact the worst case consider a search point $\boldsymbol{x}$ for which $\alpha \geq \alpha^{[i]}$, i. e. $d_1/d_2 > d_1^{[i]}/d_2^{[i]}$, so that $\xi \cdot \mathsf{E}[X_1] > 2 \cdot \alpha \cdot \mathsf{E}[X_2]$. Now consider a search point $\widetilde{\boldsymbol{x}}$ with $f(\widetilde{\boldsymbol{x}}) = f(\boldsymbol{x})$ but $\widetilde{\alpha} < \alpha$, i. e., $\widetilde{d_1} < d_1$ and $\widetilde{d_2} > d_2$. Owing to the results on Sphere we know that, for an isotropic mutation of an arbitrary fixed length ℓ_j, for any fixed $g \in (-\ell_j, \ell_j)$, $\mathsf{P}\{\Delta_j \geq g\}$ strictly increases with d_j (when $d_j > \ell_j$). Consequently, (independently of the distribution of $|\boldsymbol{m}|$) $\widetilde{\Delta_1}$ is stochastically dominated by Δ_1, whereas $\widetilde{\Delta_2}$ stochastically dominates Δ_2. This implies that X_1 dominates $\widetilde{X_1}$, whereas X_2 is dominated by $\widetilde{X_2}$ (in particular, we have $\mathsf{E}[X_1] < \mathsf{E}\Big[\widetilde{X_2}\Big]$ and $\mathsf{E}[X_2] > \mathsf{E}\Big[\widetilde{X_2}\Big]$).

As we have just seen, we may pessimistically assume that in each step the search is located at a point for which $\xi \cdot \mathsf{E}[X_1] = 2 \cdot \alpha \cdot \mathsf{E}[X_2]$. Hence, $\mathsf{E}\Big[S_1^{[k]}\Big]/d_1^{[i]} =$

$2 \cdot \mathsf{E}\left[S_2^{[k]}\right]/d_2^{[i]}$. Let $S_j := S_j^{[n^{0.3}]}$. Since $1.2/0.8 = 1.5 < 2$, it is sufficient to show that w. o. p. $S_1 \geq 0.8 \cdot \mathsf{E}[S_1]$ and w. o. p. $S_2 \leq 1.2 \cdot \mathsf{E}[S_2]$. The Hoeffding bounds (1963) (cf. Section 2.6.2 of (Hofri, 1987)) state that, for $X_j^{[k]} \in [a_j, b_j]$ and $t_j > 0$,

$$\begin{aligned} \mathsf{P}\{S_1 - \mathsf{E}[S_1] \leq -n^{0.3} \cdot t_1\} &\leq \exp\left(\frac{-2 \cdot n^{0.3} \cdot t_1^2}{(b_1 - a_1)^2}\right) \quad \text{and} \\ \mathsf{P}\{S_2 - \mathsf{E}[S_2] \geq n^{0.3} \cdot t_2\} &\leq \exp\left(\frac{-2 \cdot n^{0.3} \cdot t_2^2}{(b_2 - a_2)^2}\right). \end{aligned}$$

For $t_j = 0.2 \cdot \mathsf{E}[S_j]/n^{0.3}$, both exponents equal

$$-0.08 \cdot n^{-0.3} \cdot \mathsf{E}[S_j]^2/(b_j - a_j)^2 \quad = \quad -\Omega(n^{-0.3}) \cdot \left(\frac{\mathsf{E}[S_j]}{b_j - a_j}\right)^2,$$

respectively. Therefore, our goal is to show that $\mathsf{E}[S_j]/(b_j - a_j) = \Omega(n^{0.2})$.

First we concentrate on $\mathsf{E}[S_1]$. Since S_1 is the sum of $n^{0.3}$ RVs $X_1^{[k]}$, it suffices to show that $\mathsf{E}\left[X_1^{[k]}\right]/(b_1 - a_1) = \Omega(n^{-0.1})$ for $1 \leq k \leq n^{0.3}$. In the following we assume that $d_1 = d_1^{[i]} \pm O(d_1^{[i]}/n^{0.6})$ and $d_2 \in \left[d_2^{[i]} - O(d_2^{[i]}/n^{0.6}), d_2^{[i]}\right]$ since we have seen (in the preceding proof of the second claim) that this happens w. o. p. Owing to the results for Sphere, we know that $\mathsf{P}\{\Delta_j \geq 0\} = \Omega(1)$ implies that the scaling factor s is $O(d_j/n)$, which results in $\bar{\ell}_j = O(d_j/\sqrt{n})$, and that, under these conditions, w. o. p. $|\Delta_j| = O(\bar{\ell}_j/n^{0.4})$. Recall that $\mathsf{E}\left[\Delta_1 \cdot \mathbb{1}_{\{f' \leq f\}}\right]$ is at least $\mathsf{E}\left[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geq 0\}}\right] \cdot \mathsf{P}\{\Delta_2 \geq 0\}/2$. Since $\mathsf{P}\{\Delta_2 \geq 0\} = \Omega(1)$ in i^{th} step and $d_2 \geq d_2^{[i]}(1 - O(n^{-0.6}))$ in all $n^{0.3}$ steps, in each of these steps $\mathsf{P}\{\Delta_2 \geq 0\} = \Omega(1)$. Hence, $\mathsf{E}[X_1] = \Omega(\mathsf{E}\left[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geq 0\}}\right])$ in each of the $n^{0.3}$ steps. Owing to the results for Sphere, we know that (since $\bar{\ell}_1 = O(d_1/\sqrt{n})$ as we have seen) $\mathsf{E}\left[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geq 0\}}\right] = \Theta(\bar{\ell}_1/\sqrt{n})$ so that $\mathsf{E}[X_1] = \Omega(\bar{\ell}_1/\sqrt{n})$. Thus, $\mathsf{E}[S_1] = n^{0.3} \cdot \Omega(\bar{\ell}_1/\sqrt{n}) = \Omega(\bar{\ell}_1/n^{0.2})$ and $b_1 - a_1 = O(\bar{\ell}_1/n^{0.4})$, i. e., $\mathsf{E}[S_1]/(b_1 - a_1) = \Omega(n^{0.2})$.

Concerning a lower bound on $\mathsf{E}[S_2]$, recall that $\mathsf{E}[S_1]/d_1^{[i]} = 2 \cdot \mathsf{E}[S_2]/d_2^{[i]}$, i. e., $\mathsf{E}[S_2] = \mathsf{E}[S_1] \cdot d_2^{[i]}/(2 \cdot d_1^{[i]}) = \Omega(\bar{\ell}_1/n^{0.2}) \cdot \Omega(\xi/\alpha^{[i]})$. As $\bar{\ell}_1 = \bar{\ell}_2$ and (by assumption) $\alpha^{[i]} = O(\xi)$, we have $\mathsf{E}[S_2] = \Omega(\bar{\ell}_2/n^{0.2})$. Since $b_2 - a_2 = O(\bar{\ell}_2/n^{0.4})$ (see above), $\mathsf{E}[S_2]/(b_2 - a_2) = \Omega(\bar{\ell}_2/n^{0.2})/O(\bar{\ell}_2/n^{0.4})$ is also $\Omega(n^{0.2})$.

All in all, our initial assumption that $\alpha \geq \alpha^{[i]} \geq \alpha_{\downarrow}$ for $n^{0.3} + 1$ steps implies that w. o. p. for the first $n^{0.3}$ steps $S_1/S_2 > \alpha^{[i]}/\xi$, i. e., that w. o. p. after at most $n^{0.3}$ steps α drops below $\alpha^{[i]}$ – showing that the sequence of steps we assumed to be observed happens only with an exponentially small probability.

C Proof of Lemma 5

> "Let the scaling factor s be fixed. If $\mathsf{P}\{\Delta_1 \geq 0\}$, $1/2 - \mathsf{P}\{\Delta_1 \geq 0\}$, $\mathsf{P}\{\Delta_2 \geq 0\}$ are $\Omega(1)$, respectively, then there exists a constant $\alpha_{\Downarrow}$ such that for n large enough: if in the i^{th} step $\alpha^{[i]} \geq \alpha_{\Downarrow}$ (yet $\alpha^{[i]} = O(\xi)$, i. e. $d_1 = O(d_2)$), then w. o. p. after at most n steps the search is located at a point with $\alpha \leq \alpha^{[i]} - \Omega(\alpha^{[i]})$."

By the same arguments used before, under the given assumptions there exists $\alpha' = O(1)$ such that for n large enough $\xi \cdot \mathsf{E}\big[\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{f' \leq f\}}\big]\big] \geq \mathbf{3} \cdot \alpha \cdot \mathsf{E}\big[\mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{f' \leq f\}}\big]\big]$. Let $\alpha_{\Downarrow} := 2 \cdot \alpha'$. Assume that $\alpha^{[i]} \geq \alpha_{\Downarrow}$ and $\alpha \geq \alpha_{\Downarrow}/2 = \alpha'$ for n steps (if α drops below $\alpha_{\Downarrow}/2$ within one of these n steps, there is nothing to show). Following the same argumentation used in the proof of the preceding lemma (except for S_j now being the sum of n (instead of $n^{0.3}$) RVs), we get that w. o. p. $S_1/S_2 > 2 \cdot \alpha^{[i]}/\xi$, and hence, after these n steps w. o. p.

$$\frac{d_1}{d_2} = \frac{d_1^{[i]} - S_1}{d_2^{[i]} - S_2} < \frac{d_1^{[i]} - S_1}{d_2^{[i]} - S_1 \cdot \xi/(2 \cdot \alpha^{[i]})} = \frac{d_1^{[i]} \qquad - S_1}{d_1^{[i]} \cdot \xi/\alpha^{[i]} - S_1 \cdot \xi/(2 \cdot \alpha^{[i]})}$$
$$= \frac{d_1^{[i]} - S_1}{d_1^{[i]} - S_1/2} \cdot \frac{\alpha^{[i]}}{\xi} = \left(1 - \frac{S_1/2}{d_1^{[i]} - S_1/2}\right) \cdot \frac{d_1^{[i]}}{d_2^{[i]}}.$$

Thus, we must finally show that $S_1 = \Omega(d_1^{[i]})$. Recall that S_1 is the sum of n RVs $X_1^{[k]}$ ($\Delta_1 \cdot \mathbb{1}_{\{f' \leq f\}}$ in the $(i-1+k)^{\text{th}}$ step, respectively). In the following we consider the ith step. Our argumentation just bases on the fact that $\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{f' \leq f\}}\big] \geq \mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geq 0\}}\big] \cdot \mathsf{P}\{\Delta_2 \geq 0\}/2$ as we have seen, and since $\mathsf{P}\{\Delta_2 \geq 0\} = \Omega(1)$ by assumption, $\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{f' \leq f\}}\big] = \Omega(\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geq 0\}}\big])$. Furthermore, since $\mathsf{P}\{\Delta_1 \geq 0\}$ as well as $1/2 - \mathsf{P}\{\Delta_1 \geq 0\}$ are $\Omega(1)$ by assumption, we know that $\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geq 0\}}\big] = \Theta(d_1/n)$ (cf. Section 2). Thus, the assumptions ensure $\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{f' \leq f\}}\big] = \Omega(d_1/n)$, and hence, $\mathsf{E}[S_1] = n \cdot \Omega(d_1/n) = \Omega(d_1)$. Applying Hoeffding's bound just as in the proof of the preceding lemma, we immediately get that S_1 is $\Omega(\mathsf{E}[S_1])$, i. e. $\Omega(d_1^{[i]})$, w. o. p.

References

Beyer, H.-G. [1994]. Towards a theory of evolution strategies: Progress rates and quality gain for $(1 \stackrel{+}{,} \lambda)$-strategies on (nearly) arbitrary fitness functions. *Proc. of Parallel Problem Solving from Nature 3 (PPSN)*, LNCS 866. Springer, 58–67.

Beyer, H.-G. [2001a]. On the performance of $(1, \lambda)$-evolution strategies for the ridge function class. *IEEE Transactions on Evolutionary Computation* 5(3): 218–235.

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

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

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

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

Greenwood, G. W. and Zhu, Q. J. [2001]. Convergence in evolutionary programs with self-adaptation. *Evolutionary Computation* 9(2): 147–157.

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

Hoeffding, W. [1963]. Probability inequalities for sums of bounded random variables. *American Statistical Association Journal* 58(301): 13–30.

Hofri, M. [1987]. *Probabilistic Analysis of Algorithms.* Springer.

Jägersküpper, J. [2002]. Analysis of a simple evolutionary algorithm for the minimization in euclidean spaces. Technical Report CI-140/02, Univ. Dortmund, SFB 531. `http://sfbci.uni-dortmund.de`→Publications→Tech-Reports.

Jägersküpper, J. [2003]. Analysis of a simple evolutionary algorithm for minimization in Euclidean spaces. *Proc. of the 30th Int'l Colloquium on Automata, Languages and Programming (ICALP '03)*, LNCS 2719. Springer, 1068–1079.

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

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

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

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

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

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