Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Theoretical Computer Science 361 (2006) 38–56

Theoretical Computer Science

www.elsevier.com/locate/tcs

How the (1+1) ES using isotropic mutations minimizes positive definite quadratic forms[☆]

Jens Jägersküpper[*, 1]

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

Abstract

The (1+1) evolution strategy (ES), a simple, mutation-based evolutionary algorithm for continuous optimization problems, is analyzed. In particular, we consider the most common type of mutations, namely Gaussian mutations, and the $\frac{1}{5}$-rule for mutation adaptation, and we are interested in how the runtime/number of function evaluations to obtain a predefined reduction of the approximation error depends on the dimension of the search space.

The most discussed function in the area of ES is the so-called SPHERE-function given by SPHERE: $\mathbb{R}^n \to \mathbb{R}$ with $\boldsymbol{x} \mapsto \boldsymbol{x}^\top \boldsymbol{I}\boldsymbol{x}$ (where $\boldsymbol{I} \in \mathbb{R}^{n\times n}$ is the identity matrix), which also has already been the subject of a runtime analysis. This analysis is extended to arbitrary positive definite quadratic forms that induce ellipsoidal fitness landscapes which are "close to being spherically symmetric", showing that the order of the runtime does not change compared to SPHERE. Furthermore, certain positive definite quadratic forms $f: \mathbb{R}^n \to \mathbb{R}$ with $\boldsymbol{x} \mapsto \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}$, where $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$, inducing ellipsoidal fitness landscapes that are "far away from being spherically symmetric" are exemplarily investigated, namely

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

with $\xi = \text{poly}(n)$ such that $1/\xi \to 0$ as $n \to \infty$. It is proved that the optimization very quickly stabilizes and that, subsequently, the runtime to halve the approximation error is $\Theta(\xi \cdot n)$ compared to $\Theta(n)$ for SPHERE.

MSC: 68W40; 90C56; 90C59

Keywords: Continuous optimization; Evolutionary direct search method; Probabilistic analysis of runtime

1. Introduction

Methods for solving continuous optimization problems (search space $\mathbb{R}^n$) are usually classified into first-order, second-order, and zeroth-order methods depending on whether they utilize the gradient (the first derivative) of the

[☆] Expanded and extended version of a talk presented at the 8th International Workshop on Foundations of Genetic Algorithms (FOGA 2005) in Aizu-Wakamatsu, Japan.

[*] Tel.: +49 231 755 2435; fax: +49 231 755 2047.

E-mail address: jens.jaegerskuepper@udo.edu

[1] Supported by the German Research Foundation (DFG) as part of the collaborative research center "Computational Intelligence" (SFB 531).

0304-3975/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2006.04.004

objective function, the gradient and the Hessian (the second derivative), or neither of the two. [2] Zeroth-order methods are also called *derivative-free* or *direct search methods*. Newton's method is a classical second-order method. First-order methods are commonly (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*). Although it is obvious, we should note here that, in general, we cannot expect a zeroth-order method to out-perform first-order methods or even second-order methods.

In cases 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 theoretician's 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.

Such a theory has been developed successfully since the mid-1990s for discrete search spaces, essentially $\{0, 1\}^n$; cf. [16,4]. Recently, first results for non-artificial but well-known problems have been obtained, e.g. for the maximum matching problem [5], for the minimum spanning-tree [12], and for the partition problem [17].

The situation for continuous evolutionary optimization is different. Here, the vast majority of the results are based on empiricism, i.e., experiments are performed and their outcomes are interpreted. Also, convergence properties of EAs have been studied to a considerable extent (e.g. [14,6,3]). 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. [2]), 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 algorithmic runtime, given by the number of function evaluations, of the (1+1) ES using the $\frac{1}{5}$-rule was presented [11]. 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 $\frac{1}{5}$-rule for the adaptation of the mutation strength is rigorously analyzed. As mentioned in the abstract, the present article will extend this result to a broader class of functions, where we are going to apply differential geometry in the analysis of fitness landscapes, which was already suggested in [1].

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 which 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, [11] gives a proof that the $\frac{1}{5}$-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 $\Omega(n)$.

[2] Note that here "continuous" relates to the search space rather than to *f*, and that, unlike in mathematical 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 search space.

1.1. The Algorithm

We will concentrate on the (1+1) evolution strategy ((1+1) ES), which dates back to the mid-1960s (cf. [13,15]). 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 $f(\boldsymbol{c}') \leqslant f(\boldsymbol{c})$ 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 (w. r. t. 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 generating a *Gaussian mutation* vector $\widetilde{\boldsymbol{m}} \in \mathbb{R}^n$ each component of which is independently standard normal distributed first; 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. Let $|\boldsymbol{x}|$ denote the Euclidean length of a vector $\boldsymbol{x} \in \mathbb{R}^n$, i.e. its L^2-norm. The crucial property of a Gaussian mutation is that $\widetilde{\boldsymbol{m}}$, and with it $\boldsymbol{m}$, is isotropically distributed, i.e., $\boldsymbol{m}/|\boldsymbol{m}|$ is uniformly distributed upon the unit hyper-sphere and the length of the mutation, namely the random variable $|\boldsymbol{m}|$, is independent of the direction $\boldsymbol{m}/|\boldsymbol{m}|$.

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 point, 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 (precisely for this reason, the optimization scenario is also called *black-box optimization*). Based on experiments and rough calculations for two function scenarios (namely SPHERE and a corridor function), Rechenberg proposed the $\frac{1}{5}$-(*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 $\frac{1}{5}$. Hereinafter, a mutation that results in $f(\boldsymbol{c}') \leqslant f(\boldsymbol{c})$ 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 $\frac{1}{5}$, we would observe that on average one fifth of the mutations are successful. Thus, the $\frac{1}{5}$-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. Naturally, various implementations of the $\frac{1}{5}$-rule can be found in the literature, yet in fact, one result of [11] 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. Also the proofs presented here remain valid for such implementations of the $\frac{1}{5}$-rule; the parameters n, 2, and $\frac{1}{2}$ are chosen merely for notational convenience. We can even substitute any positive constant strictly smaller than $\frac{1}{2}$ for the "$\frac{1}{5}$."

The state of the art in mutation adaptation, however, seems to be the *covariance matrix adaptation* (*CMA*) [7] 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. Obviously, an algorithmic analysis of CMA is a much more complex task—apparently, too complex at present.

1.2. The function scenario

In this section we will have a closer look at the fitness landscape under consideration. 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 positive definite quadratic forms (PDQFs).

At first glance, one might guess that mixed terms (e.g. $3x_1x_2$) may crucially affect the fitness landscape induced by a PDQF $\boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}$. However, this is not the case: first note that we can assume $\boldsymbol{Q}$ to be symmetric (by balancing Q_{ij} with Q_{ji} for $i \neq j$ since they affect only the term $(Q_{ij}+Q_{ji})\, x_{ij}\, x_{ji}$ in the quadratic function to be black-box-optimized). Furthermore, any symmetric matrix can be diagonalized since it has n eigenvectors. Namely, eigen-decomposition yields $\boldsymbol{Q} = \boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^{-1}$ for a diagonal matrix $\boldsymbol{D}$ and an orthogonal matrix[3] $\boldsymbol{R}$.

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 $(\boldsymbol{R}\boldsymbol{x})^\top(\boldsymbol{R}\boldsymbol{x}) = \boldsymbol{x}^\top\boldsymbol{R}^\top\boldsymbol{R}\boldsymbol{x} = \boldsymbol{x}^\top\boldsymbol{R}^{-1}\boldsymbol{R}\boldsymbol{x} = \boldsymbol{x}^\top\boldsymbol{I}\boldsymbol{x} = \boldsymbol{x}^\top\boldsymbol{x}$). In short, we can assume the basis to coincide with $\boldsymbol{Q}$'s principal axes. Consequently, we can assume in the following that $\boldsymbol{Q}$ is a diagonal matrix each entry of which is positive ($\boldsymbol{Q}$'s canocial form). In other words, when talking about PDQFs we are 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 can even assume $\xi_1 \geqslant \cdots \geqslant \xi_n$. In fact, $\xi_1, \ldots, \xi_n$ are the n eigenvalues of $\boldsymbol{Q}$ (which need not necessarily be distinct). Then $\boldsymbol{Q}$'s condition number equals ξ_1/ξ_n.

For a given f-value of ϕ, the corresponding *level set* is defined as $\{\boldsymbol{x} \mid f(\boldsymbol{x}) = \phi\} \subseteq \mathbb{R}^n$ and the *lower level set* is given by $\{\boldsymbol{x} \mid f(\boldsymbol{x}) < \phi\} \subseteq \mathbb{R}^n$. For instance, the level set defined by SPHERE $= \phi^2$ forms the hyper-sphere with radius ϕ centered at the origin, and the corresponding lower level set forms the corresponding open hyper-ball. Furthermore, for a non-empty set $M \subseteq \mathbb{R}^n \setminus \{\boldsymbol{0}\}$ we let $\sup_{\boldsymbol{x},\boldsymbol{y}\in M}\{|\boldsymbol{x}|/|\boldsymbol{y}|\}$ denote the *bandwidth* of the set. Note that 1 is the smallest possible bandwidth, then all vectors in M are of the same length. The level sets of SPHERE have bandwidth 1, for instance.

The level set E_{ϕ^2} defined by $\sum_{i=1}^n \xi_i \cdot {x_i}^2 = \phi^2 > 0$ forms a hyper-surface, namely a hyper-ellipsoid, and since $\xi_1 \geqslant \cdots \geqslant \xi_n$, $\min\{|\boldsymbol{x}| \mid \boldsymbol{x} \in E_{\phi^2}\} = \phi/\sqrt{\xi_1}$ and $\max\{|\boldsymbol{x}| \mid \boldsymbol{x} \in E_{\phi^2}\} = \phi/\sqrt{\xi_n}$ so that the level sets of a PDQF have bandwidth $\sqrt{\xi_1/\xi_n}$. Note the relationship between this bandwidth and $\boldsymbol{Q}$'s condition number, namely, the condition number equals the square of the bandwidth. We call the fitness landscape induced by a PDQF *close to being spherically symmetric* if the bandwidth (and with it the condition number) is $\mathrm{O}(1)$, i.e., if the n eigenvalues are in $[a, \kappa \cdot a]$ for some $a > 0$ (which may depend on n) and a constant $\kappa \geqslant 1$. We may also use the notion *PDQF of/with bounded bandwidth* in such cases.

As mentioned in the abstract, besides of PDQFs with bounded bandwidth, we will exemplarily consider the following class of (sequences of) quadratic forms, where $n \in 2\mathbb{N}$ and $\xi = \mathrm{poly}(n)$ such that $1/\xi \to 0$ as $n \to \infty$:

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

Since $n/2$ of the eigenvalues equal 1, respectively, and the other $n/2$ eigenvalues equal ξ, respectively, the corresponding ellipsoidal fitness landscape has level sets of bandwidth $\sqrt{\xi} = \omega(1)$, i.e., the condition number (which equals ξ) is unbounded.

In the next section, some of the results presented in [11], which will be used here, will be shortly restated. In Section 3, the complete class of fitness landscapes induced by PDQFs of bounded bandwidth are investigated, whereas in Section 4 the fitness landscapes of unbounded bandwidth induced by the function class f_n defined above is considered. 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 [11] that we will use are recapitulated; for more details cf. [10].

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

[3] An orthogonal matrix $\boldsymbol{R}$ corresponds to an orthonormal transformation, i.e. a (possibly improper) rotation; then $\boldsymbol{R}^{-1}$ is the corresponding "anti-rotation."

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

Recall the following asymptotics when $g(n), h(n) > 0$ for n large enough:

- $g(n) = \mathrm{O}(h(n))$ if there exists a positive constant κ such that $g(n) \leqslant \kappa \cdot h(n)$ for n large enough,
- $g(n) = \Omega(h(n))$ if $h(n) = \mathrm{O}(g(n))$,
- $g(n) = \Theta(h(n))$ if $g(n)$ is $\mathrm{O}(h(n))$ as well as $\Omega(h(n))$,
- $g(n) = \mathrm{poly}(n)$ if $g(n) = \mathrm{O}(n^{\kappa})$ for some constant κ,
- $g(n) = \mathrm{o}(h(n))$ if $g(n)/h(n) \to 0$ as $n \to \infty$,
- $g(n) = \omega(h(n))$ if $h(n) = \mathrm{o}(g(n))$.

As we are interested in how the runtime (defined as the number of *f*-evaluations) depends on n, the dimensionality of the search space, all asymptotics are w. r. t. this parameter (unless stated differently).

Let $\boldsymbol{c} \in \mathbb{R}^n \setminus \{\boldsymbol{0}\}$ denote a search point and $\boldsymbol{m}$ a scaled Gaussian mutation. Furthermore, we let $\varDelta := |\boldsymbol{c}| - |\boldsymbol{c} + \boldsymbol{m}|$ denote the spatial gain of a mutation towards the origin, the optimum for SPHERE. Since $\text{SPHERE}(\boldsymbol{c}) = |\boldsymbol{c}|^2$, we have $\text{SPHERE}(\boldsymbol{c} + \boldsymbol{m}) < \text{SPHERE}(\boldsymbol{c}) \iff \varDelta > 0$, i.e., there is progress in the objective space iff there is progress towards the (unique) optimum in the search space. The analysis of the (1+1) ES for SPHERE has shown that

$$\begin{array}{l} \mathsf{P}\{\varDelta \geqslant 0 \mid |\boldsymbol{m}| = \ell\} \geqslant \varepsilon, \\ \text{for a constant } \varepsilon \in (0, \frac{1}{2}) \text{ for } n \text{ large enough} \end{array} \iff \ell = \mathrm{O}(|\boldsymbol{c}|/\sqrt{n}),$$

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 $\mathrm{O}(1/\sqrt{n})$-fraction of the distance between $\boldsymbol{c}$ and this point. On the other hand,

$$\begin{array}{l} \mathsf{P}\{\varDelta \geqslant 0 \mid |\boldsymbol{m}| = \ell\} \leqslant \frac{1}{2} - \varepsilon, \\ \text{for a constant } \varepsilon \in (0, \frac{1}{2}) \text{ for } n \text{ large enough} \end{array} \iff \ell = \Omega(|\boldsymbol{c}|/\sqrt{n}),$$

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 $\frac{1}{2}$ 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 correlates with the constant in the O-notation resp. in the Ω-notation.)

Since $|\widetilde{\boldsymbol{m}}|$, the length of a Gaussian mutation, is χ-distributed with n degrees of freedom, the expected length of the mutation vector $\boldsymbol{m}$ equals $s \cdot \mathsf{E}[|\widetilde{\boldsymbol{m}}|] = s \cdot \sqrt{n} \cdot (1 - \Theta(1/n))$. Moreover, with $\bar{\ell} := \mathsf{E}[|\boldsymbol{m}|]$ we have $\mathsf{P}\{||\boldsymbol{m}| - \bar{\ell}\,| \geqslant \delta \cdot \bar{\ell}\,\} \leqslant \delta^{-2}/(2n-1)$ for $\delta > 0$, in other words, there is only small deviation in the length of a Gaussian mutation; e.g., with probability $1 - \mathrm{O}(1/n)$ the mutation vector's actual length differs from its expected length by no more than ±1%. This implies that—when scaled Gaussian mutations are used—the following three events/conditions are equivalent:

- $s = \Theta(|\boldsymbol{c}|/n)$,
- $\bar{\ell} = \Theta(|\boldsymbol{c}|/\sqrt{n})$,
- $\exists$ constant $\varepsilon > 0$ such that $\mathsf{P}\{\varDelta \geqslant 0\} \in [\varepsilon, \frac{1}{2} - \varepsilon]$ for n large enough,
  i.e., $\mathsf{P}\{\varDelta \geqslant 0\}$ is $\Omega(1)$ as well as $\frac{1}{2} - \Omega(1)$.

This equivalence will be of great help in the upcoming reasonings.

Concerning the (expected) spatial gain towards the optimum, recall that for SPHERE a mutation is accepted by elitist selection iff $\varDelta \geqslant 0$, i.e., negative gains are zeroed out so that the expected spatial gain of a step is $\mathsf{E}[\varDelta \cdot \mathbb{1}_{\{\varDelta \geqslant 0\}}]$. For scaled Gaussian mutations, we know that $\mathsf{E}[\varDelta \cdot \mathbb{1}_{\{\varDelta \geqslant 0\}}]$ is $\mathrm{O}(\bar{\ell}/\sqrt{n})$. Moreover, we know that $\mathsf{E}[\varDelta \cdot \mathbb{1}_{\{\varDelta \geqslant 0\}}]$ is $\mathrm{O}(|\boldsymbol{c}|/n)$ for *any* isotropic mutation, i.e., not only for an arbitrarily scaled Gaussian mutation, but for *any* distribution of $|\boldsymbol{m}|$.

On the other hand, for scaled Gaussian mutations $\mathsf{E}[\varDelta \cdot \mathbb{1}_{\{\varDelta \geqslant 0\}} \mid s = \Theta(|\boldsymbol{c}|/n)]$ is $\Omega(\bar{\ell}/\sqrt{n})$, i.e. $\Omega(|\boldsymbol{c}|/n)$. In other words, the distance from the optimum is expected to decrease by a $\Theta(1/n)$-fraction if s is chosen appropriately. Furthermore, in this situation for any constant $\kappa > 0$ the distance decreases (at least) by an κ/n-fraction with probability $\Omega(1)$.

Concerning the mutation adaptation by the $\frac{1}{5}$-rule for SPHERE, note that during an observation phase (in which the scaling factor s is kept unchanged) the success probabilities are non-increasing since the distance from the optimum is non-increasing. Hence, if $\mathsf{P}\{\varDelta \geqslant 0\}$ is smaller than, say, 0.1 in the first step of a phase, then the expected number of successful steps (of the n steps) in this phase is smaller than $0.1n$ and, by Chernoff bounds, w. o. p. less than $0.2n$ steps are observed so that s is halved. Analogously, if $\mathsf{P}\{\varDelta \geqslant 0\}$ is larger than, say, 0.3 in the last step of a phase then the expected number of successful steps in this phase is larger than $0.3n$ and, again by Chernoff bounds, w. o. p. more than $0.2n$ steps are observed so that s is doubled. This can be used to show that w. o. p. the $\frac{1}{5}$-rule is able to keep the scaling factor optimal up to constant factors, i.e. $s = \Theta(|\boldsymbol{c}|/n)$, for an arbitrary polynomial number of steps, implying that in each of these steps $\mathsf{P}\{\Delta \geqslant 0\}$ is $\Omega(1)$ as well as $\frac{1}{2} - \Omega(1)$.

3. Fitness landscapes that are close to being spherically symmetric (bounded bandwidth/condition number)

In this section, we are going to formally prove that "slightly deforming" Sphere does not affect the order of the algorithmic runtime of a (1+1) ES using isotropic mutations. More important than this result itself, however, the line of reasoning will be made clear so that we can concentrate on the crucial difference that "an unbounded deformation" of Sphere makes in the subsequent section.

As we have already noted in the introduction of the fitness landscape, the level set E_{ϕ^2} forms a hyper-ellipsoid. When we want to utilize the results for Sphere, we need to know what the maximum and the minimum curvature at points in E_{ϕ^2} are. Since $\xi_1 \geqslant \cdots \geqslant \xi_n$, it is sufficient to consider the plane curve defined by the intersection of E_{ϕ^2} with the x_1-x_n-plane. Let I denote this intersection, which forms a plane curve. All points in I satisfy $\xi_1 x_1^2 + \xi_n x_n^2 = \phi^2$, i.e. $x_n = \sqrt{(\phi^2 - \xi_1 \cdot x_1^2)/\xi_n}$ as a function of $x_1 \in [-\phi/\sqrt{\xi_1}, \phi/\sqrt{\xi_1}\,]$. Since the curvature at a point in I (as a function of x_1) equals

$$\frac{\mathrm{d}^2 x_n/(\mathrm{d}x_1)^2}{(1 + (\mathrm{d}x_n/\mathrm{d}x_1)^2)^{3/2}} = \frac{\xi_1 \cdot \xi_n \cdot \phi^2}{(\xi_n \cdot \phi^2 + (\xi_1 - \xi_n) \cdot \xi_1 \cdot x_1^2)^{3/2}},$$

the maximum curvature of the plane curve I equals $\xi_1/(\sqrt{\xi_n} \cdot \phi)$ at the point $(0, \ldots, 0, \phi/\sqrt{\xi_n})$, which has maximum distance from the optimum/the origin w. r. t. all points in E_{ϕ^2}. Analogously, the minimum curvature equals $\xi_n/(\sqrt{\xi_1} \cdot \phi)$ at the point $(\phi/\sqrt{\xi_1}, 0, \ldots, 0)$, which has minimum distance from the optimum w. r. t. all points in E_{ϕ^2}.

In particular, this result on the curvature tells us that for *any* $\boldsymbol{c}$ in E_{ϕ^2}, there is a hyper-sphere $S^+ \ni \boldsymbol{c}$ with radius $\phi \cdot \sqrt{\xi_1}/\xi_n$ such that the lower level set $E_{<\phi^2}$ lies completely inside S^+ (i.e. $S^+ \cap E_{<\phi^2} = \emptyset$ and $E_{<\phi^2}$ is a subset of the open hyper-ball B^+ whose missing boundary is S^+), and that there is another hyper-sphere $S^- \ni \boldsymbol{c}$ with radius $\phi \cdot \sqrt{\xi_n}/\xi_1$ such that the open ball B^- whose missing boundary is S^- is a subset of the lower level set $E_{<\phi^2}$. Note that, for PDQFs with level sets of bounded bandwidth, the radii of S^+ and S^- are of the same order, namely $\Theta(|\boldsymbol{c}|)$. This will be crucial in the following.

Now consider a mutation $\boldsymbol{c}' := \boldsymbol{c} + \boldsymbol{m}$. Then $\boldsymbol{c}'$ is as good as $\boldsymbol{c}$ iff $\boldsymbol{c}' \in E_{\phi^2}$ and better than $\boldsymbol{c}$ iff $\boldsymbol{c}' \in E_{<\phi^2}$. Hence, the mutation is accepted iff $\boldsymbol{c}' \in E_{\leqslant\phi^2} := E_{\phi^2} \cup E_{<\phi^2}$. As we have just seen, $\boldsymbol{c}' \in E_{\leqslant\phi^2} \Rightarrow \boldsymbol{c}' \in B^+ \cup S^+$, and therefore we obtain

$$\begin{aligned}\mathsf{E}\left[\Delta \cdot \mathbb{1}_{\{f(\boldsymbol{c}') \leqslant f(\boldsymbol{c})\}}\right] &= \mathsf{E}\left[\Delta \cdot \mathbb{1}_{\{\boldsymbol{c}' \in E_{\leqslant\phi^2}\}}\right] \leqslant \mathsf{E}\left[\Delta \cdot \mathbb{1}_{\{\boldsymbol{c}' \text{ is at least as close to the center of } S^+ \text{ as } c\}}\right] \\ &= \mathsf{E}\left[\Delta \cdot \mathbb{1}_{\{\Delta \geqslant 0\}} \mid \textsc{Sphere}(\boldsymbol{c}) = \phi^2 \xi_1/\xi_n^2\right]\end{aligned}$$

for the expected spatial gain—independent of the distribution of $|\boldsymbol{m}|$, i.e., in particular, for any given scaling factor s for a Gaussian mutation.

As noted in the preliminaries, the results for Sphere have shown that in such a situation the expected spatial gain is O(radius of S^+/n), i.e. $\mathrm{O}((\phi/n)\sqrt{\xi_1}/\xi_n)$, independent of how the distribution of $|\boldsymbol{m}|$ is chosen.[4] However, we are interested in how fast the f-value reduces during a run of the (1+1) ES rather than the distance from the optimum point. Naturally, we obtain an upper bound if we assume that the spatial gain is realized completely along the component with the heaviest weight ξ_1. Hence, for an f-value of ϕ^2 we assume that the search were located at $\boldsymbol{c} = (\phi/\sqrt{\xi_1}, 0, \ldots, 0)$ and that the mutant were located at $\boldsymbol{c}' = (\phi/\sqrt{\xi_1} - \varepsilon \cdot (\phi/n)\sqrt{\xi_1}/\xi_n, 0, \ldots, 0)$ for some positive $\varepsilon = \mathrm{O}(1)$.

[4] In fact, the expected gain is maximum if the RV $|\boldsymbol{m}|$ is concentrated on a certain value that is Θ(radius of $S^+/\sqrt{n}$).

Then

$$f(\boldsymbol{c}') = \xi_1 \cdot \left(\frac{\phi}{\sqrt{\xi_1}} - \frac{\phi \cdot \varepsilon \cdot \sqrt{\xi_1}}{n \cdot \xi_n}\right)^2 = \xi_1 \cdot \phi^2 \cdot \left(\frac{1}{\xi_1} - \frac{2 \cdot \varepsilon}{n \cdot \xi_n} + \frac{\varepsilon^2 \cdot \xi_1}{n^2 \cdot \xi_n^2}\right) \geqslant \xi_1 \cdot \phi^2 \cdot \left(\frac{1}{\xi_1} - \frac{2 \cdot \varepsilon}{n \cdot \xi_n}\right)$$
$$= \phi^2 \cdot \left(1 - \frac{2 \cdot \varepsilon \cdot \xi_1}{n \cdot \xi_n}\right) = f(\boldsymbol{c}) \cdot \left(1 - \mathrm{O}\left(\frac{\xi_1/\xi_n}{n}\right)\right).$$

Obviously, this upper bound is useful only when $\xi_1/\xi_n = \mathrm{o}(n)$. One reason for this is that the maximum radius of curvature, which we have just used for the upper bound, is $\phi \cdot \sqrt{\xi_1}/\xi_n$, whereas the maximum radius of E_{ϕ^2} is only $\phi/\sqrt{\xi_n}$, i.e., the radius of S^+ is by a factor of $\sqrt{\xi_1/\xi_n}$ larger. However, for PDQFs of bounded bandwidth we have (by definition) $\xi_1 \leqslant \kappa \cdot \xi_n$ for a positive constant κ, i.e. $\xi_1/\xi_n = \mathrm{O}(1)$, so that the upper bound on a step's maximum expected *f*-gain of $\mathrm{O}((f(\boldsymbol{c})/n)(\xi_1/\xi_n))$ becomes $\mathrm{O}(f(\boldsymbol{c})/n)$—which is the same order as for SPHERE. Consequently, we obtain the same asymptotic lower bound on the runtime.

Theorem 2. *Let a* $(1+1)$ *ES using isotropic mutations minimize a PDQF of bounded bandwidth in* $\mathbb{R}^n$, *i.e., the corresponding condition number is* $\mathrm{O}(1)$. *Then, independently of the mutation adaptation, the number of steps to reduce the approximation error to a* 2^{-b}*-fraction,* $1 \leqslant b = poly(n)$, *is* $\Omega(b \cdot n)$ *in expectation and yet w. o. p.*

Proof. Assume that the optimization starts at $\boldsymbol{c} \in \mathbb{R}^n$, and recall that the *f*-value is non-increasing during the optimization (due to elitist selection). Then even when $|\boldsymbol{m}|$ is chosen optimally, the expected *f*-gain of a step is $\mathrm{O}(f(\boldsymbol{c})/n)$ as we have just seen. Hence, there is a constant $\kappa > 0$ such that the total expected *f*-gain in $k := \kappa \cdot n$ steps is greater than $f(\boldsymbol{c})/5$ but smaller than $f(\boldsymbol{c})/4$. By Markov's inequality, with a probability of at least $\frac{1}{2}$, the total gain in these k steps is smaller than $f(\boldsymbol{c})/2$. In other words, with a probability of at least $\frac{1}{2}$ more than k steps are necessary to halve the approximation error, and consequently, the expected number of steps to halve the approximation error is larger than $k \cdot \frac{1}{2} = \Omega(n)$. By iterating this argument using the linearity of expectation, we obtain a bound of $\Omega(b \cdot n)$ on the expected number of steps to halve the approximation error b times.

The next step is to apply Hoeffding's bound to the total gain which a sequence of steps yields. Unfortunately, the RVs corresponding to the single-step gains are not independent (which is not an issue above because of the linearity of expectation). Recall the assumption that $|\boldsymbol{m}|$ were chosen optimally in each and every step; then the optimal choice for $|\boldsymbol{m}|$ in the second step depends on the gain realized in the first step, for instance. However, also part of our best case assumption is that $\boldsymbol{c}$ is, respectively, located at a point (in the respective level set) where the curvature is minimum (so that the radius of the sphere that we use in the estimate, namely S^+, is maximum, which again results in maximum expected gain). As the *f*-value is non-increasing, we thus obtain an upper bound on the total gain of k subsequent steps by adding up the gain of k independent instances of the first step. Therefore, let $X_1, \ldots, X_k$ denote independent instances of the RV corresponding to the *f*-gain in the first step, and let $X := X_1 + \cdots + X_k$. If $0 \leqslant X_i \leqslant z > 0$, then [8] tells us that $\mathsf{P}\{X \geqslant \mathsf{E}[X] + v\} \leqslant \exp\{-2(v/z)^2/k\}$ for $v > 0$. With $v := \mathsf{E}[X]$ this inequality becomes $\mathsf{P}\{X \geqslant 2\mathsf{E}[X]\} \leqslant \exp\{-2(\mathsf{E}[X]/z)^2/k\} =: p$, and hence, the probability that k steps suffice to halve the approximation error is not only bounded by $\frac{1}{2}$ (as we have seen above) but also by p. If we can show that $(\mathsf{E}[X]/z)^2 = \Omega(n^{1+\varepsilon})$ for some constant $\varepsilon > 0$, then p is exponentially small so that the arguments used above (for the bound on the expected number of steps) yields that $b \cdot k = \Omega(b \cdot n)$ steps are necessary (to halve the approximation error b times) not only in expectation but also w. o. p.

As we know from SPHERE that w. o. p. $\varDelta = \mathrm{O}(|\boldsymbol{c}|/n^{1-\delta})$ for any positive constant δ, substituting "$n^{1-\delta}$" for "n" in the estimation of $f(\boldsymbol{c}')$, which precedes Theorem 2, yields that a step's *f*-gain is w. o. p. $\mathrm{O}((f(\boldsymbol{c})/n^{1-\delta})$ $(\xi_1/\xi_n))$, i.e. $\mathrm{O}(f(\boldsymbol{c})/n^{1-\delta})$, for any constant $\delta > 0$. Thus, when considering a polynomial number of steps, w. o. p. in all these steps the *f*-gain is $\mathrm{O}(f(\boldsymbol{c})/n^{1-\delta}))$, respectively. We obtain

$$(\mathsf{E}[X]/z)^2 = \left(\frac{\Omega(f(\boldsymbol{c}))}{\mathrm{O}(f(\boldsymbol{c}) \cdot n^{\delta-1})}\right)^2 = \Omega(n^{2-2\delta}),$$

which implies (as we have already seen above) that p is in fact exponentially small—and with it the probability to halve the approximation error within k steps. □

In the preceding lower-bound proof we assume optimal adaption of the length of the mutation vectors. Consequently, the concrete adaptation mechanism is irrelevant, and moreover, the arguments for halving the approximation error can simply be iterated to obtain a lower bound on the runtime necessary to reduce the approximation error to a certain fraction. For an upper bound on the runtime, however, precisely these two aspects are the crucial points in an analysis.

Theorem 3. *Let a* (1+1) *ES using Gaussian mutations adapted by a* $\frac{1}{5}$*-rule minimize a PDQF with bounded bandwidth in* $\mathbb{R}^n$, i.e., *the corresponding condition number is* $\mathrm{O}(1)$. *If the initialization is such that the success probability of the mutation in the first step is* $\Omega(1)$ *as well as* $\frac{1}{2} - \Omega(1)$, *then w. o. p. the* $\frac{1}{5}$*-rule maintains this property for an arbitrary polynomial number of steps.*

Proof. The crucial property that will help us with the analysis is the bounded bandwidth. It implies that, for a given $f(\boldsymbol{c})$-value of ϕ^2, either s is $\Theta(|\boldsymbol{c}|/n)$ or it is not, independent of where the current search point $\boldsymbol{c}$ is located in the ellipsoidal level set E_{ϕ^2}. Thus, we can switch back and forth between the assumptions that $\boldsymbol{c}$ is located at minimum or at maximum distance from the optimizer (w. r. t. the given f-value). Equivalently (cf. Section 2), either s is such that the probability of generating a better mutant is $\Omega(1)$ as well as $\frac{1}{2} - \Omega(1)$, or it is not—wherever $\boldsymbol{c}$ is located in E_{ϕ^2}.

For a fixed scaling factor s, we let $p_{\boldsymbol{c}} := \mathsf{P}\{f(\boldsymbol{c}') \leqslant f(\boldsymbol{c})\}$ denote the success probability (of the mutation in this step) as well as

$$p_{\boldsymbol{c}}^{\max} := \max_{\boldsymbol{x} \in E_{f(\boldsymbol{c})}} \mathsf{P}\{f(\boldsymbol{x}') \leqslant f(\boldsymbol{x})\} \quad \text{and} \quad p_{\boldsymbol{c}}^{\min} := \min_{\boldsymbol{x} \in E_{f(\boldsymbol{c})}} \mathsf{P}\{f(\boldsymbol{x}') \leqslant f(\boldsymbol{x})\},$$

we may drop the subscript "$\boldsymbol{c}$" in unambiguous situations. Thus, $p \in [\varepsilon, \frac{1}{2} - \varepsilon]$ for a constant $\varepsilon > 0$ implies $\varepsilon' \leqslant p^{\min} \leqslant p \leqslant p^{\max} \leqslant \frac{1}{2} - \varepsilon'$ for a constant $\varepsilon' > 0$ (because of the boundedness).

During a phase in a run of the (1+1) ES the scaling factor is kept unchanged, and since elitist selection is used, i.e. the f-value is non-increasing, $p^{\max}$ as well as $p^{\min}$ are non-increasing during a phase, although p may increase from one step to another within a phase. This enables us to apply the same reasoning to $p^{\max}$ and to $p^{\min}$ which was applied to the success probability in the analysis of the minimization of SPHERE. This reasoning will be recapitulated in short in the following.

We are going to show that (w. o. p. for an arbitrary polynomial number of steps) $p^{\min} = \Omega(1)$, i.e., it does not drop below a constant positive threshold, and that $p^{\max} = \frac{1}{2} - \Omega(1)$ on the other hand.

Let $p_{(i)}$ denote the success probability in the first step of the ith phase. Assume that the mutation strength s is large such that $\varepsilon \geqslant p_{(i)}^{\max} = \Omega(1)$ for a constant ε, which we will choose appropriately small later, and n large enough. Since $p^{\max}$ is non-increasing and $p \leqslant p^{\max}$ during a phase, in each step of this phase $p \leqslant \varepsilon$, and hence, we expect at most an ε-fraction of the steps in this phase to be successful. By Chernoff bounds, w. o. p. less than a 2ε-fraction of the steps are successful so that the scaling factor s is halved (we choose $2\varepsilon \leqslant \frac{1}{5}$), resulting in a larger success probability—when comparing $p_{(i+1)}$ with the success probability in the last step of the ith phase. The crucial question is, however, whether $p_{(i+1)}^{\max}$ is at least $p_{(i)}^{\max}$. If this is the case, then $p^{\min}$ in the last step of the ith phase is the (lower) threshold for the success probability we are aiming at (since $p^{\max} = \Omega(1) \Rightarrow p^{\min} = \Omega(1)$ because of the boundedness). Here is the point where the choice of ε comes into play. The (upper bound on the) (expected) number of successful steps in the phase is proportional to ε, and since only successful steps can result in a gain, by choosing a smaller ε we can make the phase's total gain smaller. All in all, we can choose ε small enough such that the increase of the success probability due to the halving of s (over)balances the (potential) decrease due to the phase's (potential) spatial gain towards the optimum. It remains to show that our choice satisfies $\varepsilon = \Omega(1)$. To this end we can use the lower bound on the runtime we have already shown. Namely, the proof of Theorem 2 in Section 3 tells us that the spatial gain of a phase (of $\mathrm{O}(n)$ steps) is such that after the phase the distance is at least a constant fraction of the initial one. This implies that the success probability at the end of the phase is also at least a constant fraction of the initial one, i.e., if it is $\Omega(1)$ in the first step, then it is $\Omega(1)$ also in the last step of the phase. This observation finishes the $\Omega(1)$-threshold on the steps' success probabilities.

Fortunately, the upper threshold of $\frac{1}{2} - \Omega(1)$ on the steps' success probabilities is easier to show. Assume that the mutation strength s is small such that in the last step of the jth phase the success probability is large, say, $p^{\min} \in [0.3, 0.4]$. Since $p \geqslant p^{\min} \geqslant 0.3$ and during a phase (in which s is kept unchanged) $p^{\min}$ is non-increasing, we expect at least 30%

of the steps in the jth phase to be successful. By Chernoff bounds, w. o. p. more than 20% successful steps are observed so that s is doubled, resulting in a larger mutation strength and, as a consequence, in a smaller $p^{\min}$ in the first step of the $(j+1)$th phase, compared to the last step of the jth phase, yet also compared to $p^{\min}_{(j)}$, the success probability in the first step of jth phase, because $p^{\min}$ is non-increasing during a phase. Then $p^{\max}_{(j)}$ is the upper threshold we are aiming at. To see that $p^{\max}_{(j)}$ is at most $\frac{1}{2} - \Omega(1)$, recall that due to the boundedness $p^{\min} = \frac{1}{2} - \Omega(1) \Rightarrow p^{\max} = \frac{1}{2} - \Omega(1)$, and that due to the upper bound on the gain of a phase, we have $p^{\min}_{(j)} = \frac{1}{2} - \Omega(1)$ if in the last step of the jth phase $p^{\min} = \frac{1}{2} - \Omega(1)$ (because the distance at the end of the phase is at least a constant fraction of the distance at the beginning).

All together we have shown that w. o. p. in each of an arbitrary polynomial number of steps the success probability is $\Omega(1)$ as well as $\frac{1}{2} - \Omega(1)$. □

Interestingly—and fortunately—, in the preceding proof of that the $\frac{1}{5}$-rule works, we merely need that the gain of a phase is not too large. However, having proved that the $\frac{1}{5}$-rule works, we can now show that the gain of a phase is large enough to obtain an upper bound on the runtime that asymptotically matches the more general (w. r. t. the adaptation) lower bound obtained in Theorem 2 in Section 3.

Theorem 4. *Let a* $(1+1)$ *ES using Gaussian mutations adapted by a* $\frac{1}{5}$*-rule minimize a PDQF with bounded bandwidth in* $\mathbb{R}^n$, i.e., *the corresponding condition number is* $\mathrm{O}(1)$. *If the initialization is such that* $s = \Theta(|\boldsymbol{c}|/n)$, *then the number of steps to reduce the approximation error to a* 2^{-b}*-fraction*, $1 \leqslant b = poly(n)$, *is* $\mathrm{O}(b \cdot n)$ *w. o. p.*

Proof. First note that the assumption on the initialization implies that $p_{(1)}$ is $\Omega(1)$ as well as $\frac{1}{2} - \Omega(1)$ and that Theorem 3 in Section 3 tells us that this also holds (at least w. o. p.) for an arbitrary polynomial number of steps. Hence, $s = \Theta(|\boldsymbol{c}|/n)$ in all these steps.

Analogous to the arguments preceding Theorem 2 in Section 3, we have $f(\boldsymbol{c}') \leqslant f(\boldsymbol{c}) \Leftrightarrow \boldsymbol{c}' \in E_{\leqslant \phi^2} \Leftarrow \boldsymbol{c}' \in B^- \cup S^-$, and hence, we obtain

$$\begin{aligned} \mathsf{E}\left[\varDelta \cdot \mathbb{1}_{\{f(\boldsymbol{c}') \leqslant f(\boldsymbol{c})\}}\right] = \mathsf{E}\left[\varDelta \cdot \mathbb{1}_{\{\boldsymbol{c}' \in E_{\leqslant \phi^2}\}}\right] \geqslant \mathsf{E}\left[\varDelta \cdot \mathbb{1}_{\{\boldsymbol{c}' \text{ is at least as close to the center of } S^- \text{ as } c\}}\right] \\ = \mathsf{E}\left[\varDelta \cdot \mathbb{1}_{\{\varDelta \geqslant 0\}} \mid \textsc{Sphere}(\boldsymbol{c}) = \phi^2 \xi_n / \xi_1^2\right] \end{aligned}$$

for the expected spatial gain of a step— for any distribution of $|\boldsymbol{m}|$, i.e., in particular for scaled Gaussian mutations.

As noted in the preliminaries, the results for SPHERE have shown that the spatial gain is $\Omega(\text{radius of } S^-/n)$, i.e. $\Omega((\phi/n)\sqrt{\xi_n}/\xi_1)$ which is $\Omega(\phi/n)$ because of the boundedness, in expectation as well as with probability $\Omega(1)$, if the scaling factor s is such that S^- is hit with a probability that is $\Omega(1)$ as well as $\frac{1}{2} - \Omega(1)$, which is actually the case as we have seen. Moreover, even when such a spatial gain is realized completely along the component with the lightest weight ξ_n, it corresponds to an f-gain of an $\Omega(1/n)$-fraction. Thus, each step reduces the approximation error by an $\Omega(1/n)$-fraction with probability $\Omega(1)$. By Chernoff bounds, in a phase of $\Theta(n)$ steps, the number of steps each of which does actually reduce the f-value by an $\Omega(1/n)$-fraction is $\Omega(n)$ w. o. p. Consequently, w. o. p. the approximation error/the f-value is reduced by a constant fraction within a phase. In particular, w. o. p. a constant number of phases, i.e. $\mathrm{O}(n)$ steps, suffice to halve the approximation error, so that finally in $\mathrm{O}(b)$ phases, i.e. $\mathrm{O}(b \cdot n)$ steps, the approximation error is reduced to a 2^{-b}-fraction w. o. p. □

Now that we know how and why the $\frac{1}{5}$-rule works on PDQFs with bounded bandwidth, we are ready to consider deformations that result in ellipsoidal level sets with unbounded bandwidth. In this section it has not been necessary to care about the actual location of the search point in its respective level set. And precisely the answer to this question, where the trajectory of the search points is located in the fitness landscape, whether in a region of high or of low curvature, will be the crucial point in the analysis of how the (1+1) ES using Gaussian mutations adapted by a $\frac{1}{5}$-rule minimizes such an "ill-conditioned" function.

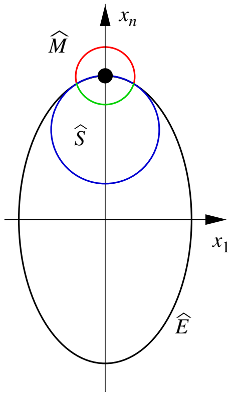
4. Fitness landscapes that are far away from being spherically symmetric (unbounded bandwidth/condition number)

We focus on the (1+1) ES using Gaussian mutations adapted by a $\frac{1}{5}$-rule in this section, and as mentioned in the abstract, we will exemplarily consider the following class of (sequences of) PDQFs, where $n \in 2\mathbb{N}$ and $\xi = \omega(1)$:

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

Since $f_n(\boldsymbol{x}) = \xi \cdot \mathrm{SPHERE}_{n/2}(\boldsymbol{y}) + \mathrm{SPHERE}_{n/2}(\boldsymbol{z})$ where $\boldsymbol{y} := (x_1, \ldots, x_{n/2})$ and $\boldsymbol{z} := (x_{n/2+1}, \ldots, x_n)$, the aim is to minimize the sum of two separate SPHERE functions, one in $S_1 = \mathbb{R}^{n/2}$ and one in $S_2 = \mathbb{R}^{n/2}$, having weight ξ resp. 1. For short: $f(\boldsymbol{x}) = \xi \cdot |\boldsymbol{y}|^2 + |\boldsymbol{z}|^2$. 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 $\xi \cdot |\boldsymbol{y} + \boldsymbol{m_1}|^2 + |\boldsymbol{z} + \boldsymbol{m_2}|^2$.

Hereinafter, all results will be obtained w. r. t. the scenario described in the preceding paragraph.



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 $\hat{\ } : \mathbb{R}^n \to \mathbb{R}^2$. 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, i.e. with probability 1, be re-rotated).

4.1. Gain in a single step

In this section we have a closer look at the properties of a single Gaussian mutation in the ellipsoidal fitness landscape we consider. Since $\xi = \omega(1)$, $\xi > 1$ for n large enough, and therefore, we assume $\xi > 1$ hereinafter. 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 $f = \xi \cdot {d_1}^2 + {d_2}^2$ (for the current search point) and $f' = \xi \cdot {d_1'}^2 + {d_2'}^2$ (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 $\frac{1}{5}$. 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 $\nabla \hat{f}(d_1, d_2) = (\xi\, 2\, d_1, 2\, d_2)^\top$, 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 $d_1/d_2 = \mathrm{O}(1/\xi)$, a mutation succeeds with a probability that is $\Omega(1)$ as well as $\frac{1}{2} - \Omega(1)$ 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 $d_1/d_2 = \mathrm{O}(1/\xi)$ 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 SPHERE$(\boldsymbol{x}) = \Theta(f/\xi^2)$ (when using the same scaling factor).

We start our analysis at a point $\boldsymbol{c}$ with $\hat{\boldsymbol{c}} = (0, \phi)$, i.e. $d_1 = 0$ and $d_2 = \phi$, so that $f = \phi^2$. Consequently, $\hat{\boldsymbol{c}}$ 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 $\hat{E}$ of the n-ellipsoid $E_{\phi^2} = \{\boldsymbol{x} \mid f(\boldsymbol{x}) = \phi^2\} \subset \mathbb{R}^n$, is maximum at $\hat{\boldsymbol{c}}$. By a simple application of differential geometry as in Section 3, we get that the curvature of this 2D-curve at $\hat{\boldsymbol{c}}$ equals ξ/ϕ. Consequently, the radius of the osculating circle ($\hat{S}$ in the figure of Section 4) equals ϕ/ξ. As this circle $\hat{S}$ 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 $f' \leqslant f$, i.e.,

$$\begin{aligned}
&\mathsf{P}\{f' \leqslant f\}\\
&\quad = \mathsf{P}\{\boldsymbol{c} + \boldsymbol{m} \text{ lies inside } E\}\\
&\quad \geqslant \mathsf{P}\{\boldsymbol{c} + \boldsymbol{m} \text{ lies inside } S\}\\
&\quad = \mathsf{P}\{|\boldsymbol{x} + \boldsymbol{m}| \leqslant |\boldsymbol{x}| \text{ for some } \boldsymbol{x} \text{ with } |\boldsymbol{x}| = \text{radius of } \hat{S} = \phi/\xi\}\\
&\quad = \mathsf{P}\{\text{SPHERE}(\boldsymbol{x} + \boldsymbol{m}) \leqslant \text{SPHERE}(\boldsymbol{x}) \mid \text{SPHERE}(\boldsymbol{x}) = (\phi/\xi)^2\}.
\end{aligned}$$

In fact, our arguments yields that the above (in)equalities hold for any fixed length ℓ of an isotropic 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 $\ell \geqslant 2\phi$, 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 $\hat{M}$ is a circle (cf. the figure in Section 4) with radius ℓ centered at $\hat{\boldsymbol{c}}$. (Note that, though $\boldsymbol{c}' = \boldsymbol{c} + \boldsymbol{m}$ (where $|\boldsymbol{m}| = \ell$) is uniformly distributed upon M, $\hat{\boldsymbol{c}'}$ is *not* uniformly distributed upon $\hat{M}$.) Now consider the curvature at a point in $\hat{E} \cap \hat{M} = \{z_1, z_2\}$ (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 $\ell = \mathrm{O}(\phi/\xi)$. As the curvature at any point of $\hat{E}$ that lies inside $\hat{M}$ is greater than κ_ℓ (since $\xi > 1$), $\hat{\boldsymbol{c}}$ 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 $\ell = \mathrm{O}(\phi/\xi)$. Thus, there is also a circle with radius r_ℓ passing through $\hat{\boldsymbol{c}}$ such that $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ lie inside this circle. Consequently, 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,

$$\mathsf{P}\{f' \leqslant f \mid |\boldsymbol{m}| = \ell\} \leqslant \mathsf{P}\{\text{SPHERE}(\boldsymbol{x} + \boldsymbol{m}) \leqslant \text{SPHERE}(\boldsymbol{x}) \mid \text{SPHERE}(\boldsymbol{x}) = (b \cdot \phi/\xi)^2,\ |\boldsymbol{m}| = \ell\},$$

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

Recall that we assumed $\hat{\boldsymbol{c}} = (0, \phi) \in \mathbb{R}^2$, i.e. $d_1 = 0$ and $d_2 = \phi$, in the above reasoning. The estimates we have made to bound the probability that a mutation hits inside the n-ellipsoid E, however, remain valid as long as $d_1/d_2 = \mathrm{O}(1/\xi)$ as we will see: since ξ/ϕ is the maximum curvature of $\hat{E}$, there is always a circle $\hat{S}$ with radius ϕ/ξ lying inside $\hat{E}$ such that $\hat{S} \cap \hat{E} = \{\hat{\boldsymbol{c}}\}$. And since $\hat{S}$ 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 $d_1/d_2 = \mathrm{O}(1/\xi)$ and $\ell = \mathrm{O}(\phi/\xi)$, the curvature κ_ℓ remains $\Theta(\xi/\phi)$.

Hence, when $f(\boldsymbol{c}) = \phi^2$ such that $\boldsymbol{c}$ satisfies $d_1/d_2 = \mathrm{O}(1/\xi)$, we are in a situation resembling (w. r. t. the success probability of a Gaussian mutation with a fixed scale) the minimization of SPHERE at a point having distance $\Theta(\phi/\xi)$

from the optimum point. Concerning the $\frac{1}{5}$-rule, we then know (cf. Section 2) that

$$\begin{aligned} & \mathsf{P}\{f' \leqslant f\} \text{ is } \Omega(1) \text{ as well as } \tfrac{1}{2} - \Omega(1) \\ \Longleftrightarrow\ & s = \Theta((\phi/\xi)/n) \\ \Longleftrightarrow\ & \bar{\ell} = \Theta((\phi/\xi)/\sqrt{n}). \end{aligned}$$

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 hyper-spherical 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, namely $d_1 = 0$ as well as $d_1' = 0$ so that $d_2' = d_2 - \Theta((\phi/\xi)/n)$, we obtain that with probability $\Omega(1)$

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

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. $f'' \leqslant f$, this implies for the expected f-gain of a step

$$\mathsf{E}\left[f - f'' \,\middle|\, s = \Theta((\sqrt{f}/n)/\xi)\right] = \Omega(f/(\xi n)).$$

Due to the pessimistic assumptions, this lower bound on the f-gain is valid only for $s = \Theta((\sqrt{f}/n)/\xi)$, yet it holds independently of the ratio d_1/d_2, i.e. independently of where $\boldsymbol{c}$ is located in E_{ϕ^2}. A spatial gain of $\Theta(f/(\xi n))$ could result in a much larger f-gain, though. If $d_1/d_2 = \mathrm{O}(1/\xi)$, however, the f-gain is also $\mathrm{O}(f/(\xi n))$ as we will see.

Therefore, let $d_1 = b \cdot \phi/\xi$ with $b = \mathrm{O}(1)$ and still $f = \xi \cdot {d_1}^2 + {d_2}^2 = \phi^2$. Owing to the reasoning 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 hyper-spherical cap $C \subset M$ lying inside E. Owing to the results for SPHERE, we know that $\mathsf{E}[\mathrm{dist}(\boldsymbol{c}', I) \cdot \mathbb{1}_{\{\boldsymbol{c}' \in C\}}] = \mathrm{O}(r/n)$ even for an isotropic mutation of optimum length (resulting in minimum expected distance of the selected search point $\boldsymbol{c}''$ from the center of S). In other words, we know that if a mutation hits inside E, its expected distance from E is $\mathrm{O}(r/n) = \mathrm{O}((\phi/\xi)/n)$ 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

$$\begin{aligned} &\mathsf{E}\big[f'' \mid d_1/d_2 = \mathrm{O}(1/\xi)\big] \\ &\quad= \mathsf{E}\Big[\xi \cdot {d_1''}^2 + {d_2''}^2 \mid d_1/d_2 = \mathrm{O}(1/\xi)\Big] \\ &\quad\geqslant \xi \cdot \big(d_1 - \mathrm{O}((\phi/\xi)/n)\big)^2 + {d_2}^2 \\ &\quad= \xi \cdot \big(b\,\phi/\xi - \mathrm{O}((\phi/\xi)/n)\big)^2 + {d_2}^2 \\ &\quad\geqslant \xi \cdot \big((b\,\phi/\xi)^2 - 2b(\phi/\xi) \cdot \mathrm{O}((\phi/\xi)/n)\big) + {d_2}^2 \\ &\quad= \xi \cdot {d_1}^2 - \mathrm{O}(\phi^2/(\xi\, n)) + {d_2}^2 \\ &\quad= \phi^2 - \mathrm{O}(\phi^2/(\xi\, n)) \\ &\quad= f - \mathrm{O}(f/(\xi\, n)). \end{aligned}$$

This upper bound on the expected f-gain of a step holds for $d_1/d_2 = \mathrm{O}(1/\xi)$ only, yet for any length of an isotropic mutation, which is converse to the lower bound. However, altogether we have proved the following lemma on the spatial gain of a step when the search (point) is located in the region consisting of all search points for which $d_1/d_2 = \mathrm{O}(1/\xi)$. (Recall the initial guess that the search stabilizes in this region.)

Lemma 5. *If the current search point is such that* $d_1/d_2 = \mathrm{O}(1/\xi)$, *then* $\mathsf{P}\{f' \leqslant f\}$ *is* $\Omega(1)$ *as well as* $\frac{1}{2} - \Omega(1)$ *if and only if s, the scaling factor of the Gaussian mutation, is* $\Theta((\sqrt{f}/n)/\xi)$.

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

4.2. Multi-step behavior

The preceding lemma on the single-step behavior enables us to obtain theorems on the runtime of the (1+1) ES for the "unbounded" scenario considered here in the same way as we did in Section 3 for PDQFs with bounded bandwidth. Namely, if $d_1/d_2 = \mathrm{O}(1/\xi)$ during a phase of n steps (an observation phase of the $\frac{1}{5}$-rule) and $s = \Theta((\sqrt{f}/n)/\xi)$, i.e. $\mathsf{P}\{f' \leqslant f\}$ is $\Omega(1)$ as well as $\frac{1}{2} - \Omega(1)$, 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. Then w. o. p. after $\Theta(\xi)$ consecutive phases the approximation error is halved—*if during all these phases* $d_1/d_2 = \mathrm{O}(1/\xi)$. Since, up to now, the arguments follow the ones in Section 3, in particular the reasoning on the $\frac{1}{5}$-rule can be adopted, and we obtain the following result:

Theorem 6. *If* $d_1/d_2 = \mathrm{O}(1/\xi)$ *in the complete optimization process and the initialization is such that* $s = \Theta((\sqrt{f(\boldsymbol{c})}/n)/\xi)$, *then w. o. p. the number of steps/f-evaluations to reduce the initial f-value/approximation error to a* 2^{-b}*-fraction,* $1 \leqslant b = poly(n)$, *is* $\Theta(b \cdot \xi \cdot n)$.

Obviously, the assumption "$d_1/d_2 = \mathrm{O}(1/\xi)$ 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 $\mathrm{O}(1/\xi)$ (once this is the case). This crucial question will be tackled by a rigorous analysis in the remainder of this article.

Let $\varDelta_1 := d_1 - d_1'$ and $\varDelta_2 := d_2 - d_2'$ 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 $\varDelta_1/d_1 > \varDelta_2/d_2$. Unfortunately, $\varDelta_1$ and $\varDelta_2$ correlate because $\boldsymbol{m_1}$ and $\boldsymbol{m_2}$ are adapted using the same scaling factor s. Moreover, we must take selection into account since only certain combinations of $\varDelta_1$ and $\varDelta_2$ will be accepted. To see which combinations are actually accepted, note that

$$f' = \xi (d_1 - \varDelta_1)^2 + (d_2 - \varDelta_2)^2 = \xi d_1^2 - \xi 2 d_1 \varDelta_1 + \xi \varDelta_1^2 + d_2^2 - 2 d_2 \varDelta_2 + \varDelta_2^2,$$

and hence,

$$f' \leqslant f \iff f' - f \leqslant 0 \iff -\xi 2 d_1 \varDelta_1 + \xi \varDelta_1^2 - 2 d_2 \varDelta_2 + \varDelta_2^2 \leqslant 0.$$

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

$$\begin{aligned} & -2\alpha d_2 \varDelta_1 + \xi \varDelta_1^2 - 2 d_2 \varDelta_2 + \varDelta_2^2 \leqslant 0 \\ \iff & -\alpha \varDelta_1 + \frac{\xi \varDelta_1^2}{2 d_2} \leqslant \varDelta_2 - \frac{\varDelta_2^2}{2 d_2} \\ \iff & -\alpha \varDelta_1 \left(1 - \frac{\varDelta_1}{2 d_1}\right) \leqslant \varDelta_2 \left(1 - \frac{\varDelta_2}{2 d_2}\right) \quad \text{(using } d_2 = \xi \cdot d_1/\alpha\text{).} \end{aligned}$$

Thus, when using elitist selection, the mutant is accepted iff the last inequality holds. Whenever a mutation satisfying $-\alpha \varDelta_1 > \varDelta_2$ is accepted, then necessarily

$$1 - \frac{\varDelta_1}{2 d_1} < 1 - \frac{\varDelta_2}{2 d_2} \iff \frac{\varDelta_1}{d_1} > \frac{\varDelta_2}{d_2} \iff \varDelta_1 > \frac{d_1}{d_2} \varDelta_2 \iff \varDelta_1 > \frac{\alpha}{\xi} \varDelta_2,$$

implying that $\varDelta_1 > 0 > \varDelta_2$. Consequently, such a step surely results in $d_1''/d_2'' < d_1/d_2$, i.e. $\alpha'' < \alpha$, and hence, in the following we may concentrate on the accepted mutations for which $-\alpha \varDelta_1 \leqslant \varDelta_2$.

So, let us assume for a moment that the mutant replaces/becomes the current individual iff $-\alpha\Delta_1 \leqslant \Delta_2$. As Δ_{3-i}, $i \in \{1, 2\}$, is random, $\mathsf{E}[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}}]$ is a random variable. For instance, the RV $\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}}]$ takes the value $\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant x\}}]$ whenever the RV Δ_2 happens to take the value x. We are interested in $\mathsf{E}[\mathsf{E}[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}}]] = \mathsf{E}[d_i - d_i'']$, the expected reduction of the distance from the optimum in S_i in a step. In particular, $\mathsf{E}[d_1'']/\mathsf{E}[d_2''] \leqslant d_1/d_2$ (for $d_1, d_2 > 0$) iff the expected relative gain in S_1 is at least as large as the one in S_2, i.e., iff

$$\begin{aligned} & \mathsf{E}\big[\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}}\big]\big]/d_1 \geqslant \mathsf{E}\big[\mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}}\big]\big]/d_2 \\ \iff & \mathsf{E}\big[\mathsf{E}\big[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}}\big]\big] \cdot \xi \geqslant \mathsf{E}\big[\mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}}\big]\big] \cdot \alpha. \end{aligned}$$

In order to prove that this inequality holds for $\alpha \geqslant \alpha^*$ for some constant α^*, we aim at a lower bound on $\mathsf{E}[\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}}]]$ and at an upper bound on $\mathsf{E}[\mathsf{E}[\Delta_2 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}}]]$ in the following. Note that

$$\begin{aligned} \mathsf{E}\big[\mathsf{E}\big[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}}\big]\big] = \; & \mathsf{E}\big[\mathsf{E}\big[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_i < 0\}}\big] \cdot \mathbb{1}_{\{\Delta_{3-i} < 0\}}\big] \\ & + \mathsf{E}\big[\mathsf{E}\big[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_i < 0\}}\big] \cdot \mathbb{1}_{\{\Delta_{3-i} \geqslant 0\}}\big] \\ & + \mathsf{E}\big[\mathsf{E}\big[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_i \geqslant 0\}}\big] \cdot \mathbb{1}_{\{\Delta_{3-i} < 0\}}\big] \\ & + \mathsf{E}\big[\mathsf{E}\big[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_i \geqslant 0\}}\big] \cdot \mathbb{1}_{\{\Delta_{3-i} \geqslant 0\}}\big] \end{aligned}$$

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

$$\mathsf{E}\big[\mathsf{E}\big[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_i \geqslant 0\}}\big] \cdot \mathbb{1}_{\{\Delta_{3-i} \geqslant 0\}}\big] = \mathsf{E}\big[\mathsf{E}\big[\Delta_i \cdot \mathbb{1}_{\{\Delta_i \geqslant 0\}}\big] \cdot \mathbb{1}_{\{\Delta_{3-i} \geqslant 0\}}\big] = \mathsf{E}\big[\Delta_i \cdot \mathbb{1}_{\{\Delta_i \geqslant 0\}}\big] \cdot \mathsf{P}\{\Delta_{3-i} \geqslant 0\}.$$

Thus, for the expected gain of a step in S_i

$$\begin{aligned} \mathsf{E}\big[\mathsf{E}\big[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}}\big]\big] = \; & \mathsf{E}\big[\Delta_i \cdot \mathbb{1}_{\{\Delta_i \geqslant 0\}}\big] \cdot \mathsf{P}\{\Delta_{3-i} \geqslant 0\} \\ & + \mathsf{E}\big[\mathsf{E}\big[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_i \geqslant 0\}}\big] \cdot \mathbb{1}_{\{\Delta_{3-i} < 0\}}\big] \\ & + \mathsf{E}\big[\mathsf{E}\big[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_i < 0\}}\big] \cdot \mathbb{1}_{\{\Delta_{3-i} \geqslant 0\}}\big]. \end{aligned}$$

Since we aim at a lower bound on $\mathsf{E}[\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}}]]$, we may ignore $\mathsf{E}[\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_1 \geqslant 0\}}] \cdot \mathbb{1}_{\{\Delta_2 < 0\}}]$ (since it is non-negative anyway), and moreover, we may pessimistically assume (when $\alpha > 0$) that $\Delta_1 = -x/\alpha$ whenever Δ_2 happens to equal $x \geqslant 0$, implying

$$\mathsf{E}[\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_1 < 0\}}] \cdot \mathbb{1}_{\{\Delta_2 \geqslant 0\}}] \geqslant -\mathsf{E}[\mathsf{E}[\Delta_2 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_2 \geqslant 0\}}] \cdot \mathbb{1}_{\{\Delta_1 < 0\}}]/\alpha.$$

Since furthermore

$$\mathsf{E}\big[\mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_2 \geqslant 0\}}\big] \cdot \mathbb{1}_{\{\Delta_1 < 0\}}\big] \leqslant \mathsf{E}\big[\mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{\Delta_2 \geqslant 0\}}\big] \cdot \mathbb{1}_{\{\Delta_1 < 0\}}\big] = \mathsf{E}\big[\Delta_2 \cdot \mathbb{1}_{\{\Delta_2 \geqslant 0\}}\big] \cdot \mathsf{P}\{\Delta_1 < 0\},$$

we obtain the following lower bound for the expected gain of a step in S_1 (on the ξ-weighted SPHERE$_{n/2}$):

$$\mathsf{E}[\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}}]] \geqslant \mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geqslant 0\}}] \cdot \mathsf{P}\{\Delta_2 \geqslant 0\} - \mathsf{E}[\Delta_2 \cdot \mathbb{1}_{\{\Delta_2 \geqslant 0\}}] \cdot \mathsf{P}\{\Delta_1 < 0\}/\alpha. \tag{1}$$

For the expected gain of a step in S_2 (on the 1-weighted SPHERE$_{n/2}$), however, we will use the trivial upper bound

$$\mathsf{E}[\mathsf{E}[\Delta_2 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}}]] \leqslant \mathsf{E}[\Delta_2 \cdot \mathbb{1}_{\{\Delta_2 \geqslant 0\}}]. \tag{2}$$

With the help of these two bounds we can now proof that the relative gain of a step in S_1 becomes larger than the one in S_2 when d_1/d_2 exceeds α^*/ξ for some α^* that is indeed O(1).

Lemma 7. *If* $\mathsf{P}\{\Delta_1 \geqslant 0\}$ *and* $\mathsf{P}\{\Delta_2 \geqslant 0\}$ *are* $\Omega(1)$, *there exists a constant* α^* *such that for* $d_1/d_2 \geqslant \alpha^*/\xi$ *yet* $d_1/d_2 = \mathrm{o}(1)$

$$\mathsf{E}[\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{f' \leqslant f\}}]]/d_1 \geqslant \kappa \cdot \mathsf{E}[\mathsf{E}[\Delta_2 \cdot \mathbb{1}_{\{f' \leqslant f\}}]]/d_2$$

for any constant κ *for n large enough.*

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

$$\begin{aligned}
&\mathsf{E}[\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{f' \leqslant f\}}]] \geqslant \mathsf{E}[\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}}]] \quad \text{and} \\
&\mathsf{E}[\mathsf{E}[\Delta_2 \cdot \mathbb{1}_{\{f' \leqslant f\}}]] \leqslant \mathsf{E}[\mathsf{E}[\Delta_2 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}}]].
\end{aligned}$$

As $d_1 \cdot \xi = d_2 \cdot \alpha$ by definition, we have to show that, if $\mathsf{P}\{\Delta_1 \geqslant 0\}$ and $\mathsf{P}\{\Delta_2 \geqslant 0\}$ are $\Omega(1)$, there exists a constant α^* such that for $\alpha \geqslant \alpha^*$ yet $\alpha = \mathrm{o}(\xi)$ and n large enough

$$\xi \cdot \mathsf{E}[\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{f' \leqslant f\}}]] \geqslant \kappa \cdot \alpha \cdot \mathsf{E}[\mathsf{E}[\Delta_2 \cdot \mathbb{1}_{\{f' \leqslant f\}}]].$$

Using the lower/upper bound on the expected gain of a step in S_1 resp. S_2, namely the inequalities (1) and (2), it is sufficient to show that

$$\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geqslant 0\}}] \cdot \mathsf{P}\{\Delta_2 \geqslant 0\} - \mathsf{E}[\Delta_2 \cdot \mathbb{1}_{\{\Delta_2 \geqslant 0\}}]\big/\alpha \geqslant \mathsf{E}[\Delta_2 \cdot \mathbb{1}_{\{\Delta_2 \geqslant 0\}}] \cdot \kappa \cdot \alpha/\xi$$

in such situations. Since $\mathsf{P}\{\Delta_1 \geqslant 0\}$ and $\mathsf{P}\{\Delta_2 \geqslant 0\}$ are $\Omega(1)$ (by assumption), $\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geqslant 0\}}]$ and $\mathsf{E}[\Delta_2 \cdot \mathbb{1}_{\{\Delta_2 \geqslant 0\}}]$ are of the same order, namely $\Theta(\bar{\ell}/\sqrt{n})$. Thus, we can choose a constant α^* such that the LHS of the preceding inequality (and with it $\mathsf{E}[\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{f' \leqslant f\}}]]$) is at least $\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geqslant 0\}}] \cdot \mathsf{P}\{\Delta_2 \geqslant 0\}/2$ for $\alpha \geqslant \alpha^*$ (and n large enough). Thus, for $\alpha \geqslant \alpha^*$ the LHS is $\Omega(\bar{\ell}/\sqrt{n})$, whereas the RHS is $o(\bar{\ell}/\sqrt{n})$ since $\kappa \cdot \alpha/\xi = \mathrm{o}(1)$ by assumption. This directly implies that the inequality holds for n large enough. □

Now, the preceding Lemma tells us that when the current search point is located at a point for which $\alpha \geqslant \alpha^*$, then the expected relative gain (of the next step) towards the optimum in S_1 (on the ξ-weighted SPHERE$_{n/2}$) is, for instance, twice as large as the one in S_2 (for n large enough). Having in mind that the variations of those gains are small, it becomes apparent that α is more likely to decrease than to increase in such a step. Formally, we obtain that the probability that α does not decrease in a small number of such steps is exponentially small:

Lemma 8. *Let the scaling factor s be fixed. If in the ith step* $\alpha^{[i]} \geqslant \alpha^*$ *yet* $\alpha^{[i]} = \mathrm{o}(\xi)$ *and* $\mathsf{P}\{\Delta_1 \geqslant 0\}$ *as well as* $\mathsf{P}\{\Delta_2 \geqslant 0\}$ *are* $\Omega(1)$, *then (for n large enough) 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.* $\alpha \leqslant \alpha^{[i]} + \mathrm{O}(\alpha^{[i]}/n^{0.6})$ *in all intermediate steps.*

Proof. We begin by proving the second claim. Let us assume that, starting with the ith step, $\alpha \geqslant \alpha^{[i]}$ for $k \leqslant n^{0.3}$ steps. Recall that, due to elitist selection, the f-value is non-increasing. Since $d_2 > d_2^{[i]} \wedge f \leqslant f^{[i]}$ 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) $d_2 \leqslant d_2^{[i]}$ during these k steps. Since (for any choice of the length of an isotropic mutation) in a step w. o. p. $\Delta_2 = \mathrm{O}(d_2/n^{0.9})$, in all $k \leqslant n^{0.3}$ steps w. o. p. $d_2 \geqslant d_2^{[i]} - k \cdot \mathrm{O}(d_2^{[i]}/n^{0.9}) \geqslant d_2^{[i]} - \mathrm{O}(d_2^{[i]}/n^{0.6})$, i.e., $d_2 = d_2^{[i]}(1-\psi)$ for some $\psi = \mathrm{O}(n^{-0.6})$, respectively.

Concerning an upper bound on d_1, we have

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

and hence,

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

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

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

Now we are ready for the proof of the lemma's first claim. Therefore, assume that $\alpha\geqslant\alpha^{[i]}\geqslant\alpha^*$ 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. $d_2\geqslant d_2^{[i]}(1-\psi)$ as we have seen, this assumption implies that also w. o. p. $d_1\geqslant d_1^{[i]}(1-\psi)$, i.e., w. o. p. $d_1=d_1^{[i]}-\mathrm{O}(d_1^{[i]}/n^{0.6})$ in all $n^{0.3}$ steps. Let $X_j^{[k]}$, $j\in\{1,2\}$, denote the RV $\Delta_j\cdot\mathbb{1}_{\{f'\leqslant f\}}$ in the $(i-1+k)$th step (so that $\mathsf{E}[X_j]=\mathsf{E}[\mathsf{E}[\Delta_j\cdot\mathbb{1}_{\{f'\leqslant f\}}]]$). Then, by choosing $\kappa=2$ in Lemma 7, for $1\leqslant k\leqslant n^{0.3}$, $\mathsf{E}[X_1^{[k]}]/d_1^{[k]}\geqslant 2\cdot\mathsf{E}[X_2^{[k]}]/d_2^{[k]}$, i.e.

$$\xi\cdot\mathsf{E}\left[X_1^{[k]}\right]\geqslant 2\cdot\alpha^{[k]}\cdot\mathsf{E}\left[X_2^{[k]}\right]\geqslant 2\cdot\alpha^{[i]}\cdot\mathsf{E}\left[X_2^{[k]}\right].$$

Let $G_j^{[k]}:=X_j^{[1]}+\cdots+X_j^{[k]}$ denote the total gain of the k steps w. r. t. d_j. By linearity of expectation, $\mathsf{E}[G_1^{[k]}]/d_1^{[i]}\geqslant 2\cdot\mathsf{E}[G_2^{[k]}]/d_2^{[i]}$ for $1\leqslant k\leqslant n^{0.3}$; however, the goal is to show that $\mathsf{P}\{G_1^{[k]}/d_1^{[i]}\leqslant G_2^{[k]}/d_2^{[i]}$ for $1\leqslant k\leqslant n^{0.3}\}$ 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 $\mathsf{E}[X_1^{[k]}]/d_1^{[i]}=2\cdot\mathsf{E}[X_2^{[k]}]/d_2^{[i]}$ in each step. To see that this is in fact the worst case, consider a search point $\boldsymbol{x}$ for which $\alpha\geqslant\alpha^{[i]}$, i.e. $d_1/d_2>d_1^{[i]}/d_2^{[i]}$, such 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 for SPHERE, we know that, for an isotropic mutation of an arbitrary fixed length ℓ_j, for any potential fixed gain $g\in(-\ell_j,\ell_j)$, $\mathsf{P}\{\Delta_j\geqslant g\}$ strictly increases with d_j (for $d_j>\ell_j$). 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 $\mathsf{E}[X_1]<\mathsf{E}[\widetilde{X_1}]$ and $\mathsf{E}[X_2]>\mathsf{E}[\widetilde{X_2}]$).

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, $\mathsf{E}[G_1^{[k]}]/d_1^{[i]}=2\cdot\mathsf{E}[G_2^{[k]}]/d_2^{[i]}$. Let $G_j:=G_j^{[n^{0.3}]}$. Since $1.2/0.8=1.5<2$, it is sufficient to show that w. o. p. $G_1\geqslant 0.8\cdot\mathsf{E}[G_1]$ and w. o. p. $G_2\leqslant 1.2\cdot\mathsf{E}[G_2]$. The Hoeffding bounds [8] (cf. Section 2.6.2 of [9]) state that, for $X_j^{[k]}\in[a_j,b_j]$ and $t_j>0$,

$$\mathsf{P}\{G_1-\mathsf{E}[G_1]\leqslant -n^{0.3}\cdot t_1\}\leqslant\exp\left(\frac{-2\cdot n^{0.3}\cdot {t_1}^2}{(b_1-a_1)^2}\right)\quad\text{and}$$
$$\mathsf{P}\{G_2-\mathsf{E}[G_2]\geqslant n^{0.3}\cdot t_2\}\leqslant\exp\left(\frac{-2\cdot n^{0.3}\cdot {t_2}^2}{(b_2-a_2)^2}\right).$$

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

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

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

First we concentrate on $\mathsf{E}[G_1]$. Since G_1 is the sum of $n^{0.3}$ RVs $X_1^{[k]}$, it suffices to show that $\mathsf{E}[X_1^{[k]}]/(b_1-a_1)=\Omega(n^{-0.1})$ for $1\leqslant k\leqslant n^{0.3}$. In the following, we assume that $d_1=d_1^{[i]}\pm\mathrm{O}(d_1^{[i]}/n^{0.6})$ and $d_2\in\left[d_2^{[i]}-\mathrm{O}(d_2^{[i]}/n^{0.6}),d_2^{[i]}\right]$ since we have seen (in the proof of the lemma's second claim) that this happens w. o. p. Owing to the results for SPHERE, we know that $\mathsf{P}\{\Delta_j\geqslant 0\}=\Omega(1)$ implies that the scaling factor s is $\mathrm{O}(d_j/n)$, which results in $\bar{\ell}_j=\mathrm{O}(d_j/\sqrt{n})$, and that, under these conditions, w. o. p. $|\Delta_j|=\mathrm{O}(\bar{\ell}_j/n^{0.4})$. Recall that $\mathsf{E}[\Delta_1\cdot\mathbb{1}_{\{f'\leqslant f\}}]$ is at least $\mathsf{E}[\Delta_1\cdot\mathbb{1}_{\{\Delta_1\geqslant 0\}}]\cdot\mathsf{P}\{\Delta_2\geqslant 0\}/2$. Since $\mathsf{P}\{\Delta_2\geqslant 0\}=\Omega(1)$ in ith step and $d_2\geqslant d_2^{[i]}(1-\mathrm{O}(n^{-0.6}))$ in all $n^{0.3}$ steps, in each of these steps $\mathsf{P}\{\Delta_2\geqslant 0\}=\Omega(1)$. Hence, $\mathsf{E}[X_1]=\Omega(\mathsf{E}[\Delta_1\cdot\mathbb{1}_{\{\Delta_1\geqslant 0\}}])$ in each of the $n^{0.3}$ steps. Owing to the results for SPHERE, we know (since

$\bar{\ell}_1 = \mathrm{O}(d_1/\sqrt{n})$ as we have seen) that $\mathrm{E}[\varDelta_1 \cdot \mathbb{1}_{\{\varDelta_1 \geqslant 0\}}] = \Theta(\bar{\ell}_1/\sqrt{n})$ so that $\mathrm{E}[X_1] = \Omega(\bar{\ell}_1/\sqrt{n})$. As a consequence, $\mathrm{E}[G_1] = n^{0.3} \cdot \Omega(\bar{\ell}_1/\sqrt{n}) = \Omega(\bar{\ell}_1/n^{0.2})$ and $b_1 - a_1 = \mathrm{O}(\bar{\ell}_1/n^{0.4})$, implying $\mathrm{E}[G_1]/(b_1 - a_1) = \Omega(n^{0.2})$.

Concerning a lower bound on $\mathrm{E}[G_2]$, recall that $\mathrm{E}[G_1]/d_1^{[i]} = 2 \cdot \mathrm{E}[G_2]/d_2^{[i]}$. Thus, $\mathrm{E}[G_2] = \mathrm{E}[G_1] \cdot d_2^{[i]}/(2 \cdot d_1^{[i]}) = \Omega(\bar{\ell}_1/n^{0.2}) \cdot \Omega(\xi/\alpha^{[i]})$. As $\bar{\ell}_1 = \bar{\ell}_2$ and (by assumption) $\alpha^{[i]} = \mathrm{O}(\xi)$, we have $\mathrm{E}[G_2] = \Omega(\bar{\ell}_2/n^{0.2})$. Since $b_2 - a_2 = \mathrm{O}(\bar{\ell}_2/n^{0.4})$ (see above), $\mathrm{E}[G_2]/(b_2 - a_2) = \Omega(\bar{\ell}_2/n^{0.2})/\mathrm{O}(\bar{\ell}_2/n^{0.4})$ is also $\Omega(n^{0.2})$.

All in all, our initial assumption that $\alpha \geqslant \alpha^{[i]} \geqslant \alpha^*$ for $n^{0.3} + 1$ steps implies that w. o. p. for the first $n^{0.3}$ steps $G_1/G_2 > \alpha^{[i]}/\xi$, i.e., that w. o. p. after at most $n^{0.3}$ steps α drops below $\alpha^{[i]}$. Thus, the sequence of steps we assumed to be observed happens only with an exponentially small probability. □

Since the $\frac{1}{5}$-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 $\mathrm{O}(\alpha^{[i]}/n^{0.6}) \leqslant \alpha^{[i]}$ for n large enough, the preceding lemma shows: when starting at a point with $\alpha^{[0]} = \mathrm{O}(1)$, i.e. $d_1^{[0]}/d_2^{[0]} = \mathrm{O}(1/\xi)$, then α remains smaller than $2 \cdot \max\{\alpha^{[0]}, \alpha^*\} = \mathrm{O}(1)$ w. o. p. for any polynomial number of steps. Incorporating these new insights into the reasoning for the $\frac{1}{5}$-rule known from the analysis of SPHERE finally enables us to replace the objectionable condition "$d_1/d_2 = \mathrm{O}(1/\xi)$ in the complete optimization process" in Theorem 6 in Section 4.2 by "$d_1/d_2 = \mathrm{O}(1/\xi)$ for the initial search point"—yielding the main result on the runtime of the (1+1) ES on the quadratic forms considered:

Theorem 9. *If the initialization is such that $s = \Theta(\sqrt{f(\boldsymbol{c})}/(n \cdot \xi))$ and $d_1/d_2 = \mathrm{O}(1/\xi)$, then w. o. p. the number of steps/f-evaluations to reduce the initial approximation error/f-value to a 2^{-b}-fraction, $1 \leqslant b = poly(n)$, is $\Theta(b \cdot \xi \cdot n)$.*

Knowing that α does never (w. o. p. for any polynomial number of steps) exceed $2 \cdot \max\{\alpha^{[0]}, \alpha^*\}$ is sufficient to obtain this theorem. If in the first step $\alpha^{[0]}$ is considerably larger than α^*, however, we would like to know that there is a drift towards smaller α, namely towards α^*. And in fact, a closer look at the arguments in the proof of Lemma 8 reveals that the same arguments show that the drift towards smaller α is that strong when $\alpha \geqslant 2 \cdot \alpha^*$ that w. o. p. α drops by a constant fraction within at most n steps:

Lemma 10. *Let the scaling factor s be fixed. If $\mathrm{P}\{\varDelta_1 \geqslant 0\}$, $\frac{1}{2} - \mathrm{P}\{\varDelta_1 \geqslant 0\}$, $\mathrm{P}\{\varDelta_2 \geqslant 0\}$ are $\Omega(1)$, respectively, then for n large enough: if in the ith step $\alpha^{[i]} \geqslant 2 \cdot \alpha^*$ yet $\alpha^{[i]} = \mathrm{o}(\xi)$, then w. o. p. after at most n steps the search is located at a point with $\alpha \leqslant \alpha^{[i]} - \Omega(\alpha^{[i]})$.*

Proof. Choosing $\kappa = 3$ in Lemma 7, we obtain $\xi \cdot \mathrm{E}[\mathrm{E}[\varDelta_1 \cdot \mathbb{1}_{\{f' \leqslant f\}}]] \geqslant 3 \cdot \alpha \cdot \mathrm{E}[\mathrm{E}[\varDelta_2 \cdot \mathbb{1}_{\{f' \leqslant f\}}]]$ for n large enough. Assume that $\alpha^{[i]} \geqslant 2\alpha^*$ and $\alpha \geqslant \alpha^*$ for n steps (if α drops below α^* within these n steps, there is nothing to show since α has been at least halved). Following the same arguments used in the proof of Lemma 8 in Section 4.2 (except for G_j now being the sum of n instead of $n^{0.3}$ RVs), we obtain that w. o. p. $G_1/G_2 > 2 \cdot \alpha^{[i]}/\xi$, and hence, after these n steps w. o. p.

$$
\begin{aligned}
\frac{d_1}{d_2} = \frac{d_1^{[i]} - G_1}{d_2^{[i]} - G_2} &< \frac{d_1^{[i]} - G_1}{d_2^{[i]} - G_1 \cdot \xi/(2 \cdot \alpha^{[i]})} \\
&= \frac{d_1^{[i]} \qquad - G_1}{d_1^{[i]} \cdot \xi/\alpha^{[i]} - G_1 \cdot \xi/(2 \cdot \alpha^{[i]})} \\
&= \frac{d_1^{[i]} - G_1}{d_1^{[i]} - G_1/2} \cdot \frac{\alpha^{[i]}}{\xi} \\
&= \left(1 - \frac{G_1/2}{d_1^{[i]} - G_1/2}\right) \cdot \frac{d_1^{[i]}}{d_2^{[i]}}.
\end{aligned}
$$

Thus, we must finally show that $G_1 = \Omega(d_1^{[i]})$. Therefore, recall that G_1 is the sum of n RVs $X_1^{[k]}$ (namely $\varDelta_1 \cdot \mathbb{1}_{\{f' \leqslant f\}}$ in the $(i-1+k)$th step, respectively). In the following we consider a single step.

As shown in the proof of Lemma 8, $\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{f' \leqslant f\}}] = \Omega(\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geqslant 0\}}])$ under the given assumptions, and since $\mathsf{P}\{\Delta_1 \geqslant 0\}$ is $\Omega(1)$ as well as $\frac{1}{2} - \Omega(1)$ by assumption, we know (cf. Section 2) that $\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geqslant 0\}}] = \Theta(d_1/n)$. All in all, the Lemma's assumptions ensure that $\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{f' \leqslant f\}}] = \Omega(d_1/n)$ in a step.

Hence, $\mathsf{E}[G_1] = n \cdot \Omega(d_1/n) = \Omega(d_1)$, and by applying Hoeffding's bound just like in the proof of Lemma 8, we finally obtain that G_1 is $\Omega(\mathsf{E}[G_1])$, i.e. $\Omega(d_1^{[i]})$, w. o. p. □

This lemma shows that α drops very quickly—if the lemma's conditions are met. Utilizing the results for SPHERE just as in Section 3, it is simple to check that the condition "$\mathsf{P}\{\Delta_1 \geqslant 0\}$ and $\frac{1}{2} - \mathsf{P}\{\Delta_1 \geqslant 0\}$ are $\Omega(1)$" is in fact ensured by the $\frac{1}{5}$-rule for $d_1/d_2 \geqslant \alpha^*/\xi$ (recall that the case $d_1/d_2 = \mathrm{O}(1/\xi)$ is covered by the arguments and proofs of Section 3; cf. the beginning of this section). The two conditions "$\alpha = \mathrm{o}(\xi)$" and "$\mathsf{P}\{\Delta_2 \geqslant 0\} = \Omega(1)$", however, originate from Lemma 7 where they enable a short and simple proof.

Naturally, for $\alpha > \alpha^*$ the drift towards smaller α increases when α increases, and the statement of the preceding lemma is true without these two conditions. So why does the proof rely on them? In the very beginning of the reasoning we decided to focus on small α, namely on α that are $\mathrm{O}(1)$. As a consequence, we decided to disregard "$\Delta_2 < 0$": it appears neither in the lower bound on the expected gain in S_1 (namely inequality (1) in Section 4.2) nor in the upper bound on the expected gain in S_2 (namely inequality (2) in Section 4.2); neither in an indicator variable, nor in a probability. Yet in fact, for a fixed positive *f*-value and a fixed positive scaling factor, $\mathsf{P}\{\Delta_2 < 0\} \to 1$ as $\alpha \to \infty$, since the mutation of a search point with $d_2 = 0$ results in $d_2' = |\boldsymbol{m_2}| \overset{\text{a.s.}}{>} 0$. Formally, we would show that $\mathsf{E}[\Delta_2 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}}]$ becomes negative when α exceeds a certain α^{**}, and for the lower bound on a step's expected gain in S_1, we would show that the term $\mathsf{E}[\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leqslant \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_1 \geqslant 0\}}] \cdot \mathbb{1}_{\{\Delta_2 < 0\}}]$, which we decided to ignore, is actually $\Omega(\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geqslant 0\}}])$ for large α. However, since it is rather evident that the drift towards smaller α becomes larger and larger as α grows, we refrain from a full formal treatment.

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. Namely, on the one hand, all positive definite quadratic forms with bounded bandwidth/condition number are covered, and on the other hand, we tackled the algorithmic analysis of the (1+1) ES using Gaussian mutations adapted by a $\frac{1}{5}$-rule for a certain subclass of positive definite quadratic forms with unbounded bandwidth, which are also sometimes called "ill-conditioned."

The main insight of these results is that Gaussian mutations adapted by the $\frac{1}{5}$-rule make the optimization process stabilize such that the trajectory of the evolving search point takes course very close to the gentlest descent of the ellipsoidal fitness landscape, i.e., in the region of (almost) maximum curvature, which leads to a poor performance. 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 (w. r. t. the search space) of a considered quadratic function *f*, namely to functions $g(\boldsymbol{x}) = f(\boldsymbol{x} - \boldsymbol{x}^*)$ for a fixed translation vector $\boldsymbol{x}^* \in \mathbb{R}^n$. Rather than considering the distance from the origin (e.g. "$|\boldsymbol{c}|$"), we merely must consider the distance from the optimum point $\boldsymbol{x}^*$ (e.g. "$|\boldsymbol{c} - \boldsymbol{x}^*|$") in all arguments/conditions. The implications for functions that are translations w. r. t. the objective space, namely $g(\boldsymbol{x}) = f(\boldsymbol{x}) + \kappa$ for some constant $\kappa \in \mathbb{R}$, are also straight forward. Since the minimum value equals κ in that case, however, we can no longer use the current function value as the measure of the approximation error. Either we use $g(\boldsymbol{x}) - \kappa$, or we restrict ourselves to the approximation error w. r. t. the search space, i.e., to the distance from the optimum search point.

References

[1] H.-G. Beyer, Towards a theory of evolution strategies: progress rates and quality gain for $(1 \overset{+}{,} \lambda)$-strategies on (nearly) arbitrary fitness functions, in: Proceedings of Parallel Problem Solving from Nature 3 (PPSN), Lecture Notes in Computer Science, Vol. 866, Springer, Berlin, 1994, pp. 58–67.

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

[3] A. Bienvenue, O. Francois, Global convergence for evolution strategies in spherical problems: some simple proofs and difficulties, Theoretical Computer Science 306 (2001) 269–289.

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

[5] O. Giel, I. Wegener, Evolutionary algorithms and the maximum matching problem, in: Proceedings of the 20th International Symposium on Theoretical Aspects of Computer Science (STACS), Lecture Notes in Computer Science, Vol. 2607, Springer, Berlin, 2003, pp. 415–426.

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

[7] N. Hansen, A. Ostermeier, Adapting arbitrary normal mutation distributions in evolution strategies: the covariance matrix adaptation, in: Proceedings of the IEEE International Conference on Evolutionary Computation (ICEC), 1996, pp. 312–317.

[8] W. Hoeffding, Probability inequalities for sums of bounded random variables, Amer. Statist. Assoc. J. 58 (301) (1963) 13–30.

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

[10] J. Jägersküpper, Analysis of a simple evolutionary algorithm for the minimization in Euclidean spaces, Technical Report CI-140/02, University of Dortmund, SFB 531, ⟨http://sfbci.uni-dortmund.de⟩→Publications→Reihe CI.

[11] J. Jägersküpper, Analysis of a simple evolutionary algorithm for minimization in Euclidean spaces, in: Proceedings of the 30th International Colloquium on Automata, Languages and Programming (ICALP), Lecture Notes in Computer Science, Vol. 2719, Springer, Berlin, 2003, pp. 1068–1079.

[12] F. Neumann, I. Wegener, Randomized local search, evolutionary algorithms, and the minimum spanning tree problem, in: Proceedings of the Genetic and Evolutionary Computation Conference (GECCO), Lecture Notes in Computer Science, Vol. 3102, Springer, Berlin, 2004, pp. 713–724.

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

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

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

[16] I. Wegener, Theoretical aspects of evolutionary algorithms, in: Proceedings of the 28th International Colloquium on Automata, Languages and Programming (ICALP), Lecture Notes in Computer Science, Vol. 2076, Springer, Berlin, 2001, pp. 64–78.

[17] C. Witt, Worst-case and average-case approximations by simple randomized search heuristics, in: Proceedings of the 22nd Annual Symposium on Theoretical Aspects of Computer Science (STACS), Lecture Notes in Computer Science, Vol. 3404, Springer, Berlin, 2005, pp. 44–56.