UNIVERSITÄT DORTMUND

REIHE COMPUTATIONAL INTELLIGENCE

SONDERFORSCHUNGSBEREICH 531

Design und Management komplexer technischer Prozesse und Systeme mit Methoden der Computational Intelligence

Analysis of a simple (1+1) ES for the Class of Positive Definite Quadratic Forms with Bounded Condition Number

Jens Jägersküpper

Nr. CI-199/05

Interner Bericht ISSN 1433-3325 November 2005

Sekretariat des SFB 531 · Universität Dortmund · Fachbereich Informatik/XI
44221 Dortmund · Germany

Diese Arbeit ist im Sonderforschungsbereich 531, „Computational Intelligence“, der Universität Dortmund entstanden und wurde auf seine Veranlassung unter Verwendung der ihm von der Deutschen Forschungsgemeinschaft zur Verfügung gestellten Mittel gedruckt.

Analysis of a simple (1+1) ES for the Class of Positive Definite Quadratic Forms with Bounded Condition Number

Jens Jägersküpper [1]

Universität Dortmund, Informatik 2, 44221 Dortmund, Germany

JJ@Ls2.cs.uni-dortmund.de

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 1/5-rule for mutation adaptation, and we are interested in how the runtime, which we define as the 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 (PDQFs) that induce ellipsoidal fitness landscapes which are "close to being spherically symmetric." Namely, all functions $\boldsymbol{x} \mapsto \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}$ are covered, where $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ is positive definite such that its condition number, which equals the ratio of the largest of the n eigenvalues of $\boldsymbol{Q}$ to the smallest one, is $O(1)$.

We show that indeed the order of the runtime does not change compared to Sphere. Namely, we prove that *any* (1+1) ES using isotropic mutations needs $\Omega(n)$ function evaluations to halve the approximation error in expectation and yet with an overwhelming probability. On the other hand, also with an overwhelming probability $O(n)$ function evaluations suffice to halve the approximation error when a (1+1) ES uses Gaussian mutations adapted by a 1/5-rule.

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

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 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 its 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. Wegener (2001) and Droste et al. (2002). Recently, first results for non-artificial but well-known problems have been obtained, e. g. for the maximum matching problem by Giel and Wegener (2003),

[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.

for the minimum spanning-tree by Neumann and Wegener (2004), and for the partition problem by Witt (2005).

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. Rudolph (1997), Greenwood and Zhu (2001), Bienvenue and Francois (2003)). A lot of results have been obtained by analyzing a simplifying model of the stochastic process induced by the EA, for instance by letting the number of dimensions approach infinity. Unfortunately, such results rely on experimental validation as a justification for the simplifications/inaccuracies introduced by the modeling. In particular Beyer has obtained numerous results that focus on local performance measures (*progress rate*, *fitness gain*; cf. Beyer (2001)), 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 1/5-rule was presented (Jägersküpper, 2003). The function/fitness landscape considered therein is the well-know Sphere-function, given by $\textsc{Sphere}(\boldsymbol{x}) := \sum_{i=1}^{n} x_i^2 = \boldsymbol{x}^\top \boldsymbol{I} \boldsymbol{x}$, and the multi-step behavior that the (1+1) ES bears when using the 1/5-rule for the adaptation of the mutation strength is rigorously analyzed. As mentioned in the abstract, the present 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 by Beyer (1994).

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

The Algorithm

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

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

Since a worse mutant (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 $|\boldsymbol{x}|$ denotes 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 hypersphere 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 *1/5-(success-)rule.* The idea behind this adaptation mechanism is that in a step of the (1+1) ES the mutant should be accepted with probability 1/5. Hereinafter, a mutation that results in $f(\boldsymbol{c}') \leq f(\boldsymbol{c})$ is called *successful,* and hence, when talking about a mutation, *success probabil-*

ity denotes the probability that the mutant $\boldsymbol{c}' = \boldsymbol{c} + \boldsymbol{m}$ is at least as good as $\boldsymbol{c}$. Obviously, when elitist selection is used, the success probability of a step equals the probability that the mutation is accepted in this step. If every step was successful with probability 1/5, we would observe that on the average one fifth of the mutations are successful. Thus, the 1/5-rule works as follows: The optimization process is observed for n steps without changing s; if more than one fifth of the steps in this observation phase have been successful, s is doubled, otherwise s is halved. Naturally, various implementations of the 1/5-rule can be found in the literature, yet in fact, one result of Jägersküpper (2003) is that the order of the runtime is indeed not affected as long as the observation lasts $\Theta(n)$ steps and the scaling factor s is multiplied by a constant greater than 1 resp. by a positive constant smaller than 1. Also the proofs presented here remain valid for such implementations of the 1/5-rule; the parameters n, 2, and 1/2 are chosen merely for notational convenience. We can even substitute any positive constant strictly smaller than 1/2 for the "1/5."

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

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 eigen vectors. Namely, eigendecomposition yields $\boldsymbol{Q} = \boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^{-1}$ for a diagonal matrix $\boldsymbol{D}$ and an orthogonal matrix[3] $\boldsymbol{R}$.

[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."

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 $\xi_1 \geq \cdots \geq \xi_n$. 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 hypersphere 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 hypersurface, namely a hyper-ellipsoid, and since $\xi_1 \geq \cdots \geq \xi_n$, $\min\{|\boldsymbol{x}| \mid \boldsymbol{x} \in E_{\phi^2}\} = \phi/\sqrt{\xi_1}$ 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 $O(1)$, i. e., if the n eigenvalues are in $[a, \kappa \cdot a]$ for some $a > 0$ (which may depend on n) and a constant $\kappa \geq 1$. We may also use the notion **PDQF of/with bounded bandwidth** in such cases.

In the next section some of the results presented by Jägersküpper (2003), 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. We end with some concluding remarks in Section 4.

2 Preliminaries

In this section some notions and notations are introduced. Furthermore, the results obtained for the Sphere-scenario in (Jägersküpper, 2003) that we will

use are recapitulated; for more details cf. (Jägersküpper, 2002).

Definition 1 *A probability $p(n)$ is* ***exponentially small*** *in n if $p(n) \leq \exp(-g(n))$ 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 $1 - \mathsf{P}\{A(n)\}$ is exponentially small in n.*
A statement $Z(n)$ holds ***for*** *n* ***large enough*** *if $(\exists\, n_0 \in \mathbb{N})(\forall\, n \geq n_0)\, Z(n)$.*

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

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

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 $\boldsymbol{c} \in \mathbb{R}^n \setminus \{\mathbf{0}\}$ denote a search point and $\boldsymbol{m}$ a scaled Gaussian mutation. Furthermore, we let $\Delta := |\boldsymbol{c}| - |\boldsymbol{c} + \boldsymbol{m}|$ denote the spatial gain of a mutation towards the origin, the optimum for Sphere. Since $\textsc{Sphere}(\boldsymbol{c}) = |\boldsymbol{c}|^2$, we have $\textsc{Sphere}(\boldsymbol{c} + \boldsymbol{m}) < \textsc{Sphere}(\boldsymbol{c}) \iff \Delta > 0$, 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

$$\mathsf{P}\{\Delta \geq 0 \mid |\boldsymbol{m}| = \ell\} \geq \varepsilon \quad \text{for a constant } \varepsilon \in (0, \tfrac{1}{2}) \text{ for } n \text{ large enough} \qquad \Longleftrightarrow \quad \ell = O(|\boldsymbol{c}|\,/\sqrt{n}),$$

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

$$\mathsf{P}\{\Delta \geq 0 \mid |\boldsymbol{m}| = \ell\} \leq 1/2 - \varepsilon \quad \text{for a constant } \varepsilon \in (0, \tfrac{1}{2}) \text{ for } n \text{ large enough} \qquad \Longleftrightarrow \quad \ell = \Omega(|\boldsymbol{c}|\,/\sqrt{n}),$$

in other words, the mutant obtained by an isotropic mutation of $\boldsymbol{c}$ is closer to a predefined point (here again the origin) with a constant probability strictly smaller than 1/2 iff the length of the mutation vector is at least an

$\Omega(1/\sqrt{n})$-fraction of the distance between $\boldsymbol{c}$ and this point. (The actual constant ε 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 $\mathsf{P}\left\{\left||\boldsymbol{m}| - \bar{\ell}\right| \geq \delta \cdot \bar{\ell}\right\} \leq \delta^{-2}/(2n-1)$ for $\delta > 0$, in other words, there is only small deviation in the length of a Gaussian mutation; e. g., with probability $1 - O(1/n)$ the mutation vector's actual length differs from its expected length by no more than $\pm 1\%$. 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 $\mathsf{P}\{\Delta \geq 0\} \in [\varepsilon, 1/2 - \varepsilon]$ for n large enough, i. e., $\mathsf{P}\{\Delta \geq 0\}$ is $\Omega(1)$ as well as $1/2 - \Omega(1)$

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 $\Delta \geq 0$, i. e., negative gains are zeroed out so that the expected spatial gain of a step is $\mathsf{E}\left[\Delta \cdot \mathbb{1}_{\{\Delta \geq 0\}}\right]$. For scaled Gaussian mutations, we know that $\mathsf{E}\left[\Delta \cdot \mathbb{1}_{\{\Delta \geq 0\}}\right]$ is $O(\bar{\ell}/\sqrt{n})$. Moreover, we know that $\mathsf{E}\left[\Delta \cdot \mathbb{1}_{\{\Delta \geq 0\}}\right]$ is $O(|\boldsymbol{c}|/n)$ 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 $\mathsf{E}\left[\Delta \cdot \mathbb{1}_{\{\Delta \geq 0\}} \mid s = \Theta(|\boldsymbol{c}|/n)\right]$ 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 an $\Theta(1/n)$-fraction if s is chosen/adapted appropriately. Furthermore, in this situation for any constant $\kappa > 0$ the distance decreases (at least) by an κ/n-fraction with probability $\Omega(1)$.

Concerning the mutation adaptation by the 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 $\mathsf{P}\{\Delta \geq 0\}$ 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 $\mathsf{P}\{\Delta \geq 0\}$ 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 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 $\mathsf{P}\{\Delta \geq 0\}$ is $\Omega(1)$ as well as $1/2 - \Omega(1)$.

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.

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 $\xi_1 \geq \cdots \geq \xi_n$, 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. $x_n = \sqrt{(\phi^2 - \xi_1 \cdot x_1^2)/\xi_2}$ 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

$$\frac{\frac{\mathrm{d}^2 x_n}{(\mathrm{d}x_1)^2}}{\left(1 + \left(\frac{\mathrm{d}x_2}{\mathrm{d}x_1}\right)^2\right)^{3/2}} = \frac{\xi_1 \cdot \xi_n \cdot \phi^2}{(\xi_n \cdot \phi^2 + (\xi_1 - \xi_n) \cdot \xi_1 \cdot x_1^2)^{3/2}},$$

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 hypersphere $S^+ \ni \boldsymbol{c}$ 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 $\boldsymbol{c}' \in E_{\leq\phi^2} := E_{\phi^2} \cup E_{<\phi^2}$. As we have just seen, $\boldsymbol{c}' \in E_{\leq\phi^2} \Rightarrow \boldsymbol{c}' \in B^+ \cup S^+$, and therefore

we obtain

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

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 $O(\text{radius of } S^+/n)$, i. e. $O((\phi/n)\sqrt{\xi_1}/\xi_n)$, 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 $\varepsilon = O(1)$. Then

$$\begin{aligned}
f(\boldsymbol{c}') &= \xi_1 \cdot \left(\frac{\phi}{\sqrt{\xi_1}} - \frac{\phi\cdot\varepsilon\cdot\sqrt{\xi_1}}{n\cdot\xi_n}\right)^2 \\
&= \xi_1\cdot\phi^2\cdot\left(\frac{1}{\xi_1} - \frac{2\cdot\varepsilon}{n\cdot\xi_n} + \frac{\varepsilon^2\cdot\xi_1}{n^2\cdot\xi_n^2}\right) \\
&\ge \xi_1\cdot\phi^2\cdot\left(\frac{1}{\xi_1} - \frac{2\cdot\varepsilon}{n\cdot\xi_n}\right) \\
&= \phi^2\cdot\left(1 - \frac{2\cdot\varepsilon\cdot\xi_1}{n\cdot\xi_n}\right) \;=\; f(\boldsymbol{c})\cdot\left(1 - O\left(\frac{\xi_1/\xi_n}{n}\right)\right).
\end{aligned}$$

Obviously, this upper bound is useful only when $\xi_1/\xi_n = o(n)$. 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. [5] However, for PDQFs of bounded bandwidth we have (by definition) $\xi_1 \le \kappa\cdot\xi_n$ for a positive constant κ, i. e. $\xi_1/\xi_n = O(1)$, so that the upper bound on a step's maximum expected f-gain of $O((f(\boldsymbol{c})/n)(\xi_1/\xi_n))$ becomes $O(f(\boldsymbol{c})/n)$ – which is the same order as for Sphere. Consequently, we obtain the same asymptotic lower bound on the runtime.

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

[5] Notice that this factor equals the bandwidth of (the level sets of) the PDQF.

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 $O(1)$. Then – independently of the mutation adaptation – the number of steps to reduce the approximation error to a 2^{-b}-fraction, $1 \leq b = poly(n)$, 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 $O(f(\boldsymbol{c})/n)$ 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 $1/2$, the total gain in these k steps is smaller than $f(\boldsymbol{c})/2$. In other words, with a probability of at least $1/2$ 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 $k \cdot 1/2 = \Omega(n)$. 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 $0 \leq X_i \leq z > 0$, then Hoeffding (1963) tells us that $\mathsf{P}\{X \geq \mathsf{E}[X] + v\} \leq \exp\{-2(v/z)^2/n\}$ for $v > 0$. With $v := \mathsf{E}[X]$ this inequality becomes $\mathsf{P}\{X \geq 2\mathsf{E}[X]\} \leq \exp\{-2(\mathsf{E}[X]/z)^2/n\} =: p$, and hence, the probability that k steps suffice to halve the approximation error is not only bounded by $1/2$ (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. $\Delta = O(|\boldsymbol{c}|/n^{1-\delta})$ for any positive constant δ, substituting "$n^{1-\delta}$" for "n" in the estimation of $f(\boldsymbol{c}')$, which pre-

cedes Theorem 2 on the preceding page, yields that a step's f-gain is w. o. p. $O(f(\boldsymbol{c})/n^{1-\delta})(\xi_1/\xi_n))$, i. e. $O(f(\boldsymbol{c})/n^{1-\delta}))$, for any constant $\delta > 0$. Thus, when considering a polynomial number of steps, w. o. p. in all these steps the f-gain is $O(f(\boldsymbol{c})/n^{1-\delta}))$, respectively. We obtain

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

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 scaling factor. 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 1/5-rule minimize a PDQF with bounded bandwidth in $\mathbb{R}^n$, i. e., the corresponding condition number is $O(1)$. If the initialization is such that the success probability of the mutation in the first step is $\Omega(1)$ as well as $1/2-\Omega(1)$, then w. o. p. the 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. page 8), either s is such that the probability of generating a better mutant is $\Omega(1)$ as well as $1/2 - \Omega(1)$, or it is not – wherever $\boldsymbol{c}$ is located in E_{ϕ^2}.

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

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

we may drop the subscript "$\boldsymbol{c}$" in unambiguous situations. Thus, $p \in [\varepsilon, 1/2-\varepsilon]$ for a constant $\varepsilon > 0$ implies $\varepsilon' \le p^{\min} \le p \le p^{\max} \le 1/2 - \varepsilon'$ 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 $p^{\max}$ resp. $p^{\min}$ 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 $p^{\max} = 1/2 - \Omega(1)$ on the other hand.

Let $p_{(i)}$ denote the success probability in the first step of the i^{th} phase. Assume that the mutation strength s is large such that $\varepsilon \geq p^{\max}_{(i)} = \Omega(1)$ for a constant ε, which we will choose appropriately small later, and n large enough. Since $p^{\max}$ is non-increasing and $p \leq p^{\max}$ during a phase, in each step of this phase $p \leq \varepsilon$, 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 $2\varepsilon \leq 1/5$), resulting in a larger success probability – when comparing $p_{(i+1)}$ with the success probability in the last step of the i^{th} phase. The crucial question is, however, whether $p^{\max}_{(i+1)}$ is at least $p^{\max}_{(i)}$. If this is the case, then $p^{\min}$ in the last step of the i^{th} 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 on page 11 tells us that the spatial gain of a phase (of $O(n)$ 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 $1/2 - \Omega(1)$ on the steps' success probabilities is easer to show. Assume that the mutation strength s is small such that in the last step of the j^{th} phase the success probability is large, say, $p^{\min} \in [0.3, 0.4]$. Since $p \geq p^{\min} \geq 0.3$ 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 j^{th} 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 $(j{+}1)^{\text{th}}$ phase – compared to the last step of the j^{th} phase, yet also compared to $p^{\min}_{(j)}$, the success probability in the first step of j^{th} 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 $1/2-\Omega(1)$, recall that due to the boundedness $p^{\min} = 1/2-\Omega(1) \Rightarrow p^{\max} = 1/2-\Omega(1)$, and that due to the upper bound on the gain of a phase, we have $p^{\min}_{(j)} = 1/2-\Omega(1)$ if in the last step of the j^{th} phase $p^{\min} = 1/2-\Omega(1)$ (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 $1/2-\Omega(1)$. □

Interestingly – and fortunately –, in the preceeding proof of that the 1/5-rule works, we merely need that the gain of a phase is not too large. However, having proved that the 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 on page 11.

Theorem 4 *Let a (1+1) ES using Gaussian mutations adapted by a 1/5-rule minimize a PDQF with bounded bandwidth in $\mathbb{R}^n$, i. e., the corresponding condition number is $O(1)$. 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, $1 \le b = \mathit{poly}(n)$, is $O(b \cdot n)$ w. o. p.*

PROOF. First note that the assumption on the initialization implies that $p_{(1)}$ is $\Omega(1)$ as well as $1/2-\Omega(1)$ and that Theorem 3 on page 12 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.

Analogously to the arguments preceding Theorem 2 on page 11, we have $f(\boldsymbol{c}') \le f(\boldsymbol{c}) \Leftrightarrow \boldsymbol{c}' \in E_{\le\phi^2} \Leftarrow \boldsymbol{c}' \in B^- \cup S^-$, and hence, we obtain

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

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 $1/2 - \Omega(1)$, 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. $O(n)$ steps, suffice to halve the approximation error, so that finally in $O(b)$ phases, i. e. $O(b \cdot n)$ steps, the approximation error is reduced to a 2^{-b}-fraction w. o. p. □

4 Conclusion

Based on the results on how the (1+1) ES minimizes the well-known SPHERE-function when using isotropic mutations, we have extended these results to a broader class of functions, namely to all positive definite quadratic forms with bounded bandwidth/condition number. The lower bound holds for *any* (1+1) ES as long as isotropic mutations are used. The upper bound, however, applies to Gaussian mutations adapted by a 1/5-rule.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Witt, C. [2005]. Worst-case and average-case approximations by simple randomized search heuristics. *Proceedings of the 22nd Annual Symposium on*

Theoretical Aspects of Computer Science (STACS), volume 3404 of *LNCS*. Springer, 44–56.