Probabilistic Analysis of Evolution Strategies Using Isotropic Mutations

Dissertation

zur Erlangung des Grades eines
Doktors der Naturwissenschaften
der Universität Dortmund
am Fachbereich Informatik

von

Jens Jägersküpper

Dortmund
2006

Tag der mündlichen Prüfung:	19. Dezember 2006
Dekan:	Professor Dr. Peter Buchholz
Gutachter:	Professor Dr. Ingo Wegener, Juniorprofessor Dr. Thomas Jansen

Summary

This dissertation deals with optimization in high-dimensional Euclidean space $\mathbb{R}^n$. Namely, a particular type of direct-search methods known as Evolution Strategies (ESs) are investigated. Evolution Strategies mimic natural evolution, in particular mutation, in order to "evolve" an approximate solution.

As this dissertation focuses on theoretical investigation of ESs in the way randomized approximation algorithms are analyzed in theoretical computer science (rather than by means of convergence theory or dynamical-system theory), very basic and simple ESs are considered. Namely, the only search operator that is applied are so-called isotropic mutations. That is, a new candidate solution is obtained by adding a random vector to the current candidate solution the distribution of which is spherically symmetric.

General lower bounds on the number of steps/isotropic mutations which are necessary to reduce the approximation error in the search space are proved, where the focus is on how the number of optimization steps depends on (and scales with) the dimensionality of the search space. These lower bounds hold independently of the function to be optimized and for large classes of ESs. Moreover, for several concrete optimization scenarios where certain ESs optimize a unimodal function, upper bounds on the number of optimization steps are proved.

Acknowledgment

I would like to thank my advisor Ingo Wegener, not only for posing the initial question underlying this dissertation—namely whether it might be possible to obtain theoretical results for evolutionary optimization in continuous search spaces like the ones that had been obtained before for discrete search spaces—but for productive discussions, for hints, for his steady support and encouragement.

Furthermore, I would like to thank my colleagues at the chair of *Efficient Algorithms and Complexity Theory* at the computer science department of the Dortmund University for the stimulating working atmosphere, and in particular Carsten Witt and Stefan Droste for productive discussions on the topics of this dissertation.

The financial support by the German Research Foundation (DFG) through the collaborative research center "Design and Management of Complex Technical Processes and Systems by Means of Computational Intelligence Methods" (SFB 531) is kindly acknowledged.

For many things, yet in particular for their patience, I thank Annette and Nele—my family.

Symbols and Abbreviations

i. i. d.	independently identically distributed
a. s.	almost sure, i. e. with probability one
w. o. p.	with overwhelming probability (page 15)
PDQF	positive definite quadratic form
$\mathbb{1}$	indicator variable (page 14)
$X \succ Y$	the random variable X stochastically dominates the random variable Y (page 13)
$X \sim Y$	$X \succ Y$ and $X \prec Y$, i. e., the random variables X and Y are equidistributed
$\boldsymbol{x}$	bold small letters usually denote vectors/search points
$\boldsymbol{Q}$	bold capital letters usually denote matrices; $\boldsymbol{I}$ denotes the identity matrix
$\lvert\boldsymbol{x}\rvert$	Euclidean norm of the vector $\boldsymbol{x} \in \mathbb{R}^n$, i. e., $\sqrt{{x_1}^2 + \cdots + {x_n}^2}$
$\mathsf{P}\{\mathcal{E}\}$	probability of the event $\mathcal{E}$
$\mathsf{E}[X]$	expectation of the random variable X
$\mathsf{Var}[X]$	variance of the random variable X
$X^{\langle i:j \rangle}$	ith order statistic (of j) of the random variable X
X^+, X^-	$X \cdot \mathbb{1}_{\{X \geq 0\}}$ resp. $X \cdot \mathbb{1}_{\{X \leq 0\}}$ (where X is a random variable)
Γ	the (complete) Gamma function
G	the random variable defined in Equation (3.2) on page 21
$\Delta_{x^*,\ell}$	the random variable defined in Equation (4.1) on page 32
$\widetilde{G}$, $\widetilde{\Delta}$, etc.	random variables that relate to a so-called Gaussian mutation
$\mathcal{X}$	an individual, where an individual is more than just a search point
$O, \Omega, \Theta, o, \omega$	asymptotic notations (page 15)
poly(n)	$O(n^c)$ for some constant c
$\asymp$	asymptotically equal (page 15)
e	Euler's constant 2.7182… (base of the natural logarithm, i. e., $\ln \mathrm{e} = 1$)
$\mathbb{R}$	the reals
$\mathbb{R}_{>0}$	the positive reals
$\mathbb{N}$	the set $\{1, 2, 3, \ldots\}$ of natural numbers
$\mathbb{N}_0$	$\mathbb{N} \cup \{0\}$
Ψ	the value of $\int_{-1}^{1} (1 - x^2)^{(n-3)/2} \, \mathrm{d}x$, cf. Inequality (3.6) on page 24

Contents

1 Introduction **1**
- 1.1 Overview . . . 6
- 1.2 The Evolution Strategies under Consideration . . . 8
  - 1.2.1 (1+λ) Evolution Strategy . . . 8
  - 1.2.2 (1,λ) Evolution Strategy . . . 9
  - 1.2.3 (μ+1) Evolution Strategy . . . 9
  - 1.2.4 Additional Notes, Notions, and Notations . . . 10
- 1.3 Underlying Publications . . . 11

2 Preliminaries **13**

3 Isotropic Mutations **17**
- 3.1 Isotropic Probability Distributions . . . 17
- 3.2 Gaussian Mutations . . . 19
- 3.3 Spatial Gain of an Isotropic Mutation . . . 20
  - 3.3.1 Spatial Gain of a Unit Isotropic Mutation . . . 20
  - 3.3.2 Spatial Gain of a Gaussian Mutation . . . 26
- 3.4 Additional Notes . . . 28

4 General Lower Bounds **31**
- 4.1 Spatial Gain Towards a Fixed Search Point . . . 31
- 4.2 Lower Bound on the Expected Number of Steps of (1+1) ESs . . . 35
- 4.3 Lower Bound for (1‡λ) ESs which Holds with Overwhelming Probability . . . 39
- 4.4 Lower Bound for (μ+1) ESs which Holds with Overwhelming Probability . . . 43
- 4.5 Overcoming Gaps with Elitist Selection . . . 47
  - 4.5.1 Linearly Separated Gaps . . . 48
  - 4.5.2 Spherically Separated Gaps . . . 50
  - 4.5.3 Exemplary Application to Concrete Functions . . . 52
  - 4.5.4 Additional Notes on Overcoming Gaps . . . 53
- 4.6 Remarks on the Lower-Bound Results . . . 54

5 Bounds for Concrete Scenarios **57**
- 5.1 Gaussian Mutations and 1/5-Rule . . . 57
  - 5.1.1 . . . for the (1‡λ) Evolution Strategy . . . 58
  - 5.1.2 . . . for the (μ+1) Evolution Strategy . . . 59
  - 5.1.3 . . . and the Spatial Gain . . . 59

5.2 SPHERE-like Functions 61
5.2.1 . . . and the (1+1) ES with 1/5-Rule 62
5.2.2 . . . and the (1+λ) ES with 1/5-Rule 68
5.2.3 . . . and a Modified 1/5-Rule for the (1+λ) ES 70
5.2.4 . . . and the (1,λ) ES with 1/5-Rule 73
5.2.5 . . . and the (μ+1) ES with 1/5-Rule 79
5.3 The (1+1) ES on Positive Definite Quadratic Forms 84
5.3.1 . . . with Bounded Condition Number 85
5.3.2 . . . with Unbounded Condition Number 92
5.3.3 Remarks 104

6 Conclusion and Outlook **105**

1 Introduction

Finding an optimum of a given function $f : S \to \mathbb{R}$ is one of the fundamental problems—in theory as well as in practice. The search space S can be discrete or continuous, like $\mathbb{N}$ or $\mathbb{R}$. If S has more than one dimension, it may also be a mixture, like it is the case for optimization tasks that are so-called mixed-integer programs where, for instance, $\{0,1\} \times [0,1]$ might be the search space, i. e., one of the *decision variables* is discrete (here 0-1-valued) and another one is continuous (here non-negative yet at most 1). In this dissertation, the optimization in "high-dimensional" Euclidean space is considered, i. e., the search space is $\mathbb{R}^n$. What "high-dimensional" means is usually anything but well-defined. A particular 10-dimensional problem in practice may already be considered "high-dimensional" by the ones who try to solve it. In this dissertation, the crucial aspect of the optimization is how the optimization time scales with the dimensionality of the search space $\mathbb{R}^n$, i. e., we consider the optimization time as a function of n. In other words, here we are interested in what happens when the dimensionality of the search space gets higher and higher. This viewpoint is typical for analyses in computer science. Unfortunately, it seems that the optimization in continuous search spaces is not one of the core topics in computer science. Rather it lies in the domain of operations research and mathematical programming. There, however, focusing on how the optimization time scales with the search space's dimension seems rather uncommon. Rather, the performance of an optimization method is described by means of convergence theory. As an example, let us take a closer look at "linear convergence." Let $\boldsymbol{x}^*$ denote the optimum search point of a unimodal function and $\boldsymbol{x}^{[k]}$ the approximate solution after k optimization steps. Then we have

$$\frac{\operatorname{dist}(\boldsymbol{x}^*, \boldsymbol{x}^{[k+1]})}{\operatorname{dist}(\boldsymbol{x}^*, \boldsymbol{x}^{[k]})} \to c \in \mathbb{R}_{<1} \quad \text{as} \quad k \to \infty$$

where $\operatorname{dist}(\cdot,\cdot)$ denotes some distance measure, most commonly the Euclidean distance between two points (when considering convergence towards $\boldsymbol{x}^*$ in the search space $\mathbb{R}^n$), or the absolute difference in function value (when considering convergence towards the optimum function value in the objective space). From a computer scientist's point of view, the first issue with such a result is that we do not know when k is large enough to actually ensure $\operatorname{dist}(\boldsymbol{x}^*, \boldsymbol{x}^{[k+1]}) \leq c' \cdot \operatorname{dist}(\boldsymbol{x}^*, \boldsymbol{x}^{[k]})$ for some constant $c' < 1$, i. e., to ensure progress of the optimization. The second issue is that there seems to be no connection to n, the dimension of the search space. Only if c is an absolute constant, there is actual independence of n; yet in general, the *convergence rate* c depends on n. When we are interested in, say, the number of steps necessary to halve the approximation error (given by the distance from $\boldsymbol{x}^*$), the order of this number with respect to n precisely depends on how c depends on n. For instance, if $c = 1 - 0.5/n$, we need $\Theta(n)$ steps; if $c = 1 - 0.5/n^2$, however, we need $\Theta(n^2)$ steps—when k is large enough, of course. Thus, the order of convergence ("linear" in the example above) tells us something about the "final speed" of the optimization, but in general nothing about the n-dependence of the number of steps necessary to ensure a certain

approximation error (unless c is an absolute constant; then it takes a constant number of steps to halve the distance from $\boldsymbol{x}^*$ independently of n).

Regarding the approximation error, for unconstrained optimization in $\mathbb{R}^n$ it is generally not clear how the optimization time can be measured solely with respect to the absolute error of the approximation. In contrast to discrete and finite problems (like CLIQUE), the initial error is generally not bounded (for CLIQUE the trivial solution consisting of a single vertex is an approximation with bounded error). Hence, the question how many steps it takes to get into the ε-ball around $\boldsymbol{x}^*$ does not make sense without specifying the starting conditions. Rather we must consider the optimization time with respect the relative improvement of the approximation.

The simple optimization problems that we will consider result in a somehow homogeneous optimization process which enables us to measure the performance of the algorithm by the number of steps which are necessary to halve the approximation error, i. e. the distance from $\boldsymbol{x}^*$. Starting at distance $2^b \cdot \varepsilon$ for some $b \in \mathbb{N}$, i. e., $\operatorname{dist}(\boldsymbol{x}^*, \boldsymbol{x}^{[0]}) = 2^b \cdot \varepsilon$, then gives an additional factor of b for the number of steps k which are necessary to obtain an ε-approximation, i. e., $\operatorname{dist}(\boldsymbol{x}^*, \boldsymbol{x}^{[k]}) \leq \varepsilon$.

Methods for solving optimization problems in continuous domains, essentially $S = \mathbb{R}^n$, are usually classified into first-order, second-order, and zeroth-order methods, depending on whether they utilize the gradient (the first derivative) of the objective function, the gradient and the Hessian (the second derivative), or neither of both. A zeroth-order method is also called *derivative-free* or *direct search method.* Newton's method is a classical second-order method; first-order methods can be (sub)classified into Quasi-Newton, steepest descent, and conjugate gradient methods. Classical zeroth-order methods try to approximate the gradient and to then plug this estimate into a first-order method. Finally, amongst the modern zeroth-order methods, evolutionary algorithms (EAs) come into play, which are (often general-purpose) search heuristics that mimic natural evolution—sometimes in a very broad sense. EAs for continuous optimization, however, are commonly subsumed under the term *evolution(ary) strategies (ESs).*[1]

When information about the gradient is not available, for instance if f relates to a property of some workpiece and is given by computer 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 expensive, in particular 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 are becoming more and more popular. However, the enthusiasm in practical EAs has led to an unclear variety of very sophisticated and problem-specific EAs. Unfortunately, from a theoretical point of view, the development of such EAs is solely driven by practical success, whereas the aspect of a theoretical analysis is left aside. Particularly "[i]n the early phase of ES[s], these EA[s] were mainly developed and analyzed by engineers. A more or less system-theoretic approach aiming at the prediction of the EA[s'] behavior as a dynamical system served as the central paradigm. That is, the usual way of thinking about a theory of EA[s] is considering the EA and the objective function $f\colon \mathbb{R}^n \to \mathbb{R}$ [...] in terms of a dynamical (or evolutionary) system" as noted by Beyer, Schwefel, and Wegener in their article "How to analyze evolutionary

[1] Beyer, Schwefel, and Wegener (2002, p. 107) point out: "It is common belief that evolutionary optimization of real-valued objective functions in $\mathbb{R}^n$ search spaces is a specialty of evolution strategies (ES). While there are indeed state-of-the-art ES versions specially tailored for $\mathbb{R}^n$ supporting this belief, it is historically not correct (for the history see Beyer and Schwefel (2002))."

algorithms" in *Theoretical Computer Science* (2002, p. 107). On page 108 the authors further note that even when the stochastic process which is induced by an ES is a Markov process, so that the Markov kernel "describes the dynamics of the EA system completely, its usefulness is rather limited: the analytical determination of the dynamics is almost always excluded. Even in the simplest cases the analytical determination of the Markov kernel is excluded. [...] When thinking of EA practice, the user often monitors the dynamics of the fitness values, e. g., expected average population fitness and expected best-so-far fitness come into mind. From a theoretical viewpoint also the expected distance to the optimum state (if there is a single one) is of interest. It should be the aim of theory to predict these mean-value dynamics for a given EA system analytically. However, up until now, even this task can only be accomplished for the simplest EA systems using asymptotic ($n \to \infty$) considerations or by relying on approximations."

To summarize, concerning EAs, theory has not kept up with practice, and thus, we should not try to analyze the most sophisticated EA en vogue, but concentrate on very basic, or call them "simple", EAs to build a sound and solid basis for EA-theory within the field of theoretical computer science.

For discrete search spaces, essentially $\{0,1\}^n$, such a theory has been started successfully in the 1990s, for instance Mühlenbein (1992), Rudolph (1997), Droste, Jansen, and Wegener (1998), and Garnier, Kallel, and Schoenauer (1999); cf. Wegener (2001) and Droste, Jansen, and Wegener (2002b). Meanwhile first results for non-artificial, but well-known problems have been obtained, e. g., for sorting and the shortest path problem by Scharnow, Tinnefeld, and Wegener (2002), for the maximum matching problem by Giel and Wegener (2003), for the minimum spanning tree problem by Neumann and Wegener (2004), and for a simple scheduling problem by Witt (2005b). Such results deal with the efficiency of concrete EAs for a concrete class of problems. Also complexity theoretical aspects have already been investigated:

When f is given to the optimization algorithm as an oracle for f-evaluations (zeroth-order oracle) and the cost of the optimization (the runtime) is defined as the number of queries to this oracle, we are in the so-called *black-box optimization* scenario. Nemirovsky and Yudin (1983, p. 333) state (w. r. t. the optimization in continuous search spaces) in their book *Problem Complexity and Method Efficiency in Optimization:* "From a practical point of view this situation would seem to be more typical. At the same time it is objectively more complicated and it has been studied in a far less extend than the one [with first-order oracles/methods] considered earlier." After more than two decades there still seems to be some truth in their statement—yet to a smaller extent. For discrete black-box optimization, a complexity theory has been successfully started by Droste, Jansen, Tinnefeld, and Wegener (2002a), cf. Wegener (2003) and Droste, Jansen, and Wegener (2006). Lower bounds on the number of f-evaluations (the *black-box complexity*) are proved with respect to classes of functions when an arbitrary(!) optimization heuristic (just for instance an EA) knows about the class $\mathcal{F}$ of functions to which f belongs, but nothing about f itself. The benefits of such results are obvious: They can prove that an allegedly poor performance of an apparently simple black-box algorithm on f is not due to the algorithm's simpleness, but due to the inherent black-box complexity of $\mathcal{F}$.

As mentioned above, the situation for evolutionary optimization in continuous search spaces is different. Besides the dynamical-system approach discussed above—Rechenberg (1973, 1994), Schwefel (1981, 1995), and in particular Beyer (2001)—the vast majority of the results are based on empiricism, i. e., experiments are performed and their outcomes are interpreted. However,

convergence properties of EAs have been studied to a certain extent, in particular by Rudolph (1997), by Bienvenue and Francois (2003), and already in 1989 by Rappl. Unfortunately, those results are "based on the assumption that the EA 'is able' to control the mutation strength (i. e. the expected step size) such that the conditions for the proofs are fulfilled. The mutation control part of the EA is usually not analyzed. The inclusion of the mutation control part in the analysis appears in all cases investigated until now as a difficult task" as noted by Beyer et al. (2002, p. 110). Just recently, Auger (2005) succeeded in proving the convergence of a basic evolution strategy (namely of the $(1, \lambda)$ ES using Schwefel's self-adaptation). As the minimization of the 1-dimensional function $f : \mathbb{R} \to \mathbb{R}$ with $f(x) = x^2$ is considered, also this very sophisticated proof does not reveal how the number of steps scales with dimension of the search space.

> The starting point of this dissertation was the aim to adopt and to enhance tools, methods, and techniques, which are known mainly from analyses of randomized approximation algorithms for discrete problems, in order to enable a probabilistic analysis of evolutionary algorithms for the continuous search space $\mathbb{R}^n$, so that theorems can be obtained—in particular on how the number of steps which an EA needs to realize a given approximation quality depends on n.

(Textbooks on randomized algorithms and their probabilistic analysis have been published by Hofri (1987), Motwani and Raghavan (1995), and Mitzenmacher and Upfal (2005), for instance.) In particular, the initial challenge was that the proofs should cover the adaptation mechanism that the ES uses. As it might have become clear from the discussion above, it would have been overconfident to start with a sophisticated adaptation mechanism which works particular well in today's practice. Rather the simplest one should be chosen as a starting point. In particular, it should be a deterministic adaptation mechanism to keep the "degree of randomness"—which usually makes an analysis hard—as small a possible. Thus, Rechenberg's 1/5-success-rule (1965) almost suggested itself as a candidate: it is deterministic and it is simple (as it originated in a time when computational resources were very limited).

Somewhat surprisingly, it turned out that for proving that the 1/5-rule "works"—at least in a very simple scenario—a general lower bound on the number of steps which are necessary to obtain a certain reduction of the approximation error would be a great help. As lower bounds (and complexity considerations; cf. the discussion on black-box complexity above) are of independent interest anyway, such lower-bound results will be presented in Chapter 4 before the analyses of concrete scenarios in Chapter 5 in which several ESs with 1/5-rule are considered.

In contrast to the results on the black-box complexity of certain classes of pseudo-Boolean functions discussed above, however, here the general lower bounds will be obtained with respect to particular types of evolution strategies (which are described in Section 1.2 (p. 8)). The restrictions can be roughly summarized as follows:

- "Mutation" is the only search operator (in particular, no crossover), where mutation consists in adding a random vector (sometimes called perturbation) to a search point in $\mathbb{R}^n$ in order to obtain a new candidate solution (a mutant).
- The random mutation vector is isotropically distributed, i. e., its distribution over $\mathbb{R}^n$ is rotationally/spherically symmetric (more precisely: invariant w. r. t. orthonormal transformations).

The rigorous analysis of such an "isotropic mutation" is the heart of the lower bounds (and also very important to obtain upper bounds for concrete scenarios, of course). In particular, the spatial gain of a mutation towards a fixed search point—usually the (or, a fixed) optimum—will be of utmost interest. And precisely this measure is covered by the *progress-rate theory* in the dynamical-system approach. A large number of results on progress rates exist, and many of them can be found in *The Theory of Evolution Strategies* by Beyer (2001). Situations in which noise disturbs the evaluation of the function to be optimized have been considered by Beyer and Meyer-Nieberg (2005, for instance) and particularly by Arnold (2002).

Unfortunately, those results cannot be (re)used to obtain the results we are aiming at here. The reason for this is the following: These progress rates have been obtained using the asymptotic simplification $n \to \infty$ (cf. the discussion above). Although the results that will be obtained in this dissertation are also asymptotic ones, here a different type of asymptotic will be used. To make the difference clear, we quote from *Asymptotic Methods in Analysis* by de Bruijn (1970, pp. 1–3):

> "A typical asymptotic result, and one of the oldest, is Stirling's formula [...]:
>
> $$\lim_{n\to\infty} n!\Big/\left(\mathrm{e}^{-n}n^n\sqrt{2\pi n}\right) = 1. \tag{1.1}$$
>
> For each n, the number $n!$ can be evaluated without any theoretical difficulty, and the larger n is, the larger the number of necessary operations becomes. But Stirling's formula gives a decent approximation $\mathrm{e}^{-n}n^n\sqrt{2\pi n}$, and the larger n is, the smaller its relative error becomes.
>
> [...]
>
> For no single special value of n can we draw any conclusion from (1.1) about $n!$. It is a statement about infinitely many values of n, which, remarkably enough, does not state anything about any special value of n.
>
> For the purpose of closer investigation of this feature, we abbreviate (1.1) to
>
> $$\lim_{n\to\infty} f(n) = 1, \quad \text{or} \quad f(n) \to 1 \quad (n \to \infty). \tag{1.2}$$
>
> This formula expresses the mere existence of a function $N(\varepsilon)$ with the property that:
>
> $$\text{for each } \varepsilon > 0\colon\ n > N(\varepsilon) \text{ implies } |f(n)-1| < \varepsilon. \tag{1.3}$$
>
> When proving $f(n) \to 1$, one usually produces, hidden or not, information of the form (1.3) with explicit construction of a suitable function $N(\varepsilon)$. It is clear that the knowledge of $N(\varepsilon)$ actually means numerical information about f. However, when using the notation $f(n) \to 1$, this information is suppressed. So if we write (1.2), the knowledge of a function $N(\varepsilon)$ with the property (1.3) is replaced by the knowledge of the existence of such a function.
>
> [...]
>
> A weaker form of suppression of information is given by the Bachmann-Landau O-notation [2]. It does not suppress a function, but only a number. That is to say, it replaces the knowledge of a number with certain properties by the knowledge that such a number exists. The O-notation suppresses much less information than with the limit notation, and yet it is easy enough to handle."

[2] See E. Landau, Vorlesung über die Zahlentheorie, Leipzig 1927, vol. 2, p. 3–5.

Obtaining asymptotic results with the help of the O-notation is common practice in computer science. Let f denote a function in $\mathbb{R}$ and g a function in $\mathbb{R}_{>0}$. Then we say "$f(x) = O(g(x))$ as x grows" if (and only if) there exists a constant c such that $|f(x)| \leq c \cdot g(x)$ for all $x \geq x' \in \mathbb{R}_{>0}$, so that the constant c is suppressed.[3]

The crucial difference that these two notions of "asymptotic" makes for the analysis of ESs (and in particular for the analysis of a mutation's spatial gain) is the following: If the variance of a random variable (which is normalized w. r. t. n) tends to zero as n grows, in the "$n \to \infty$" approach one may replace this random variable by its expectation, which can simplify the calculations significantly. When one aims at a probabilistic analysis and asymptotic results in the sense of "O", however, such a simplification is precluded. (This will be further discussed in Section 3.4 (p. 28).)

1.1 Overview

For the reason that has been discussed above, we have to (re)consider the random variable which corresponds to the spatial gain of an isotropic mutation in the search space $\mathbb{R}^n$. Before we come to this integral part of this dissertation, however, the framework of the evolution strategies considered in this work will be presented in the following Section 1.2 (p. 8). At the end of this introductory chapter, the publications that build the basis of this dissertation will be listed in Section 1.3 (p. 11).

Some preliminaries which may help to understand the following chapters are presented in Chapter 2. A few basic notions from probability theory are recapitulated, some notations are given, and well-known bounds on tail probabilities of random variables are quoted, namely the bounds/inequalities by Markov, Chebyshev, and Hoeffding.

Chapter 3 on "Isotropic Mutations" starts in Section 3.1 (p. 17) with a formal look at isotropic probability distributions. A very important type of isotropic mutations, namely so-called Gaussian mutations, are covered by Section 3.2 (p. 19). Subsequently, we start the analysis of the spatial gain of an isotropic mutation in Section 3.3 (p. 20). The chapter ends with some additional notes on isotropic mutations in Section 3.4 (p. 28).

The lower-bound results are presented in Chapter 4. Therefore, we proceed with the analysis of the spatial gain of an isotropic mutation in Section 4.1 (p. 31). Then the lower bounds are derived:

- In Section 4.2 (p. 35) we prove a lower bound of $\Omega(n)$ for the expected number of steps which a (1+1) ES needs to halve the approximation error in the search space (the Euclidean distance from a fixed search point in $\mathbb{R}^n$). This bound holds for any adaptation mechanism as long as isotropic mutations are used and for any function scenario.

- In Section 4.3 (p. 39) it is proved that (1+λ) ESs and $(1,\lambda)$ ESs that use a "global mutation strength" as well as $(1,\lambda)$ ES that use self-adaptive mutation strengths need with an overwhelming probability (of $1 - \mathrm{e}^{-\Omega(n)}$) $\Omega(n/\ln(1+\lambda))$ steps to halve the approximation error in the search space $\mathbb{R}^n$ (independently of the adaptation of isotropic mutations and for any function scenario).

[3] The O-notation is not limited to the case "as x grows", cf. de Bruijn (1970, Section 1.2: The O-symbol).

- In Section 4.4 (p. 43) (μ+1) ESs are considered and we prove that they need $\Omega(n \cdot \mu)$ steps/isotropic mutations with overwhelming probability to halve the approximation error in the search space (independently of the mutation adaptation and the function to be optimized).

In Section 4.5 (p. 47) we reconsider (1+λ) ESs and address the question how long it takes such elitist ESs to overcome "gaps" or "cliffs" in the fitness landscape. Lower bounds w. r. t. the size of a so-called "spherically separated gap" and of a so-called "linearly separated gap" are proved. The chapter on the lower bounds ends with additional comments and remarks in Section 4.6 (p. 54).

Chapter 5 deals with concrete optimization scenarios. In all scenarios Gaussian mutations adapted by a 1/5-rule will be considered, which are introduced in Section 5.1 (p. 57). Subsequently in Section 5.2 (p. 61) the class of SPHERE-like functions is defined and upper bounds on the runtimes of various ES are obtained for this scenario (given proper initialization):

- The (1+1) ES performs with overwhelming probability $O(n)$ steps to halve the approximation error in the search space.

- The (1+λ) ES as well as the (1, λ) ES get along with $O(n/\sqrt{\ln(1+\lambda)})$ steps with overwhelming probability—when the 1/5-rule bases on the number of successful mutations.

- The (1+λ) ES using a modified 1/5-rule, which bases on the number of successful *steps*, is proved to be indeed capable of getting along with $O(n/\ln(1+\lambda))$ steps with overwhelming probability, which is asymptotically optimal.

- The (μ+1) ES using Gaussian mutations adapted by the 1/5-rule performs $O(\mu \cdot n)$ steps with overwhelming probability to halve the approximation error in the search space, which is also asymptotically optimal.

In Section 5.3 (p. 84) a different function scenario, which can be considered a generalization of SPHERE-like functions, is investigated: positive definite quadratic forms (PDQFs). We restrict ourselves to the analysis of the (1+1) ES (using Gaussian mutations adapted by the 1/5-rule) for this scenario. It turns out that for PDQFs with a bounded condition number the upper bound of $O(n)$ obtained for SPHERE-like functions carries over. For PDQFs with a condition number that is not bounded but grows in n, a linear number of steps do not necessarily suffice to halve the approximation error. To show this, for the class of PDQFs $f_n^{\xi} : \mathbb{R}^n \to \mathbb{R}$ with

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

where $n \in 2\mathbb{N}$ and $\xi : \mathbb{N} \to \mathbb{R}_{>1}$ such that $\xi = \text{poly}(n)$ as well as $1/\xi_{(n)} \to 0$ as n grows, it is proved that the optimization process stabilizes such that $\Theta(\xi \cdot n)$ steps are necessary with overwhelming probability to halve the approximation error.

Finally, conclusions are drawn and an outlook is given in Chapter 6.

1.2 The Evolution Strategies under Consideration

1.2.1 (1+λ) Evolution Strategy

Let $\lambda\colon \mathbb{N} \to \mathbb{N}$ such that $\lambda = \text{poly}(n)$. "λ" may also abbreviate "$\lambda(n)$" in the following. The (1+λ) ES for minimization of $f\colon \mathbb{R}^n \to \mathbb{R}$ that we consider works as follows. A single/global *mutation strength* σ which takes values in $\mathbb{R}_{>0}$ is used for mutation adaptation—for the adaptation of *isotropic* mutations.

For a given initialization of the evolving search point $\boldsymbol{c} \in \mathbb{R}^n$ and the mutation strength $\sigma \in \mathbb{R}_{>0}$, the following *evolution loop* is performed:

1. FOR $i := 1$ TO λ DO
   Create a new search point $\boldsymbol{y}_{[i]} := \boldsymbol{c} + \boldsymbol{m} \in \mathbb{R}^n$, where the mutation vector $\boldsymbol{m}$ is drawn according to an isotropic mutation that depends only on σ.
2. IF $\min_{i\in\{1,\ldots,\lambda\}}\{f(\boldsymbol{y}_{[i]})\} \leq f(\boldsymbol{c})$ THEN $\boldsymbol{c} := \operatorname{argmin}_{i\in\{1,\ldots,\lambda\}}\{f(\boldsymbol{y}_{[i]})\}$ (when there are more than one mutant with minimum fitness, one of them is chosen uniformly at random).
3. Decide whether to increase, or to decrease, or to keep the mutation strength σ unchanged; adapt σ accordingly. (Details follow below.)
4. GOTO 1.

In practice, obviously, the GOTO is conditioned on a stopping criterion.[4] Fortunately, for the results we are aiming at, we need not define a reasonable stopping criterion. Rather we will consider a run of a (1+λ) ES as an infinite stochastic process. We are interested in how fast $\boldsymbol{c}$ evolves. Therefore, we let "$\boldsymbol{c}^{[i]}$" denote the current search point *after* the ith iteration of the evolution loop (so that "$\boldsymbol{c}^{[0]}$" denotes the initial search point). "$\sigma^{[i]}$" denotes the mutation strength that is used *in* the ith iteration.

Note that the (1+λ) ES is a so-called "hill climber" since mutants with a worse f-value are always discarded so that the sequence of f-values corresponding to the evolving search point is monotonic, i. e. non-increasing for minimization.

Concerning the generation of mutants in Instruction 1, we formally need a mapping from $\mathbb{R}_{>0}$ into the set of isotropic distributions which tells us (given a specific mutation strength σ) which isotropic distribution is to be used for the mutation vector. This mapping is fixed.

Concerning the adaptation of the mutation strength σ in Instruction 3, the decision (whether to increase, or to decrease, or to keep σ unchanged) may depend on the complete history of the optimization process, namely, in the kth step on the sequence $(\boldsymbol{c}^{[0]}, f(\boldsymbol{c}^{[0]})), \ldots, (\boldsymbol{c}^{[k-1]}, f(\boldsymbol{c}^{[k-1]}))$ given by the evolving search point $\boldsymbol{c}$ and also on the discarded mutants (including their f-values). The decision, however, must result in one of the three outcomes: "increase", "decrease", or "keep." Depending solely on this outcome, the mutation strength σ is updated—possibly in a randomized manner. For instance, the adaptation may be such that, when "increase" is the outcome, σ is multiplied by a factor that is uniformly chosen over the interval $[1, 2]$.

[4] In fact, since the evolution loop is repeated over and over again (no termination), this outline of a (1+λ) ES is formally not an algorithm. (Moreover, the concrete initialization is left open.) It seems that in such cases (when a framework for a class of algorithms is described) often the notion "method" is used (cf., for instance, "Newton's method").

1.2.2 (1,λ) Evolution Strategy

We obtain the "$(1,\lambda)$ ES with a global mutation strength" by dropping the IF-condition in Instruction 2 in the $(1+\lambda)$ ES above, implying that $\boldsymbol{c}$ is always replaced by (one of) the best of the λ mutants. Unlike the elitist $(1+\lambda)$ ES, the $(1,\lambda)$ ES may accept mutations that result in a search point with a worse f-value. Obviously, a (1,1) ES does not make much sense since selection becomes meaningless (in fact, no selection can take place). The search of a (1,1) ES is not completely random, i. e. independent of the function which is to be optimized, though. The function which is to be optimized does influence the search since it does affect the adaptation of the mutation strength.

In particular for the $(1,\lambda)$ ES, the concept of "self adaptation" ("SA") has been widely studied. The underlying idea is to evolve the mutation strength (or other parameters) along with the evolving search point (leading to the notion of "σSA" for self-adaptive mutation-strength control). Thus, an individual $\mathcal{C} = (\boldsymbol{c},\sigma) \in \mathbb{R}^n \times \mathbb{R}_{>0}$ consisting of a search point and an associated mutation strength is evolved. Self adaptation is sometimes also referred to as *mutative strategy-parameter control*.

For a given initialization of the evolving individual $\mathcal{C} = (\boldsymbol{c},\sigma) \in \mathbb{R}^n \times \mathbb{R}_{>0}$, the $(1,\lambda)\sigma$SA-ES (cf. Beyer (2001, p. 261)) performs the following evolution loop, where $f(\mathcal{C}) := f(\boldsymbol{c})$:

1. FOR $i := 1$ TO λ DO
   Create a new individual $\mathcal{Y}_{[i]} = (\boldsymbol{y}_{[i]},\sigma_{[i]})$, where
   $\sigma_{[i]} \in \mathbb{R}_{>0}$ depends only on σ (possibly, and usually, in a randomized manner), and where
   $\boldsymbol{y}_{[i]} := \boldsymbol{c} + \boldsymbol{m} \in \mathbb{R}^n$ with a mutation vector $\boldsymbol{m}$ drawn according to an isotropic mutation that depends only on the previously generated $\sigma_{[i]}$.
2. $(\boldsymbol{c},\sigma) := \operatorname{argmin}_{i\in\{1,\ldots,\lambda\}}\{f(\mathcal{Y}_{[i]})\}$ (when there are more than one mutant with minimum fitness, one of them is chosen uniformly at random).
3. GOTO 1.

For various operators to mutate the mutation strength σ, see Beyer (2001, Section 7.1.4). Presumably, the one that is most often used is scaling σ by multiplying it with a log-normally distributed random variable, which is due to Schwefel (1995, p. 143, for instance). For a general lower bound on the number of iterations which a $(1,\lambda)\sigma$SA-ES performs, however, the concrete σSA is not of interest. Thus, we will not go into further details of self adaptation here.

1.2.3 (μ+1) Evolution Strategy

$(\mu\overset{+}{,}\lambda)$ Evolution Strategies use a population consisting of μ individuals. As in our $(1,\lambda)\sigma$SA-ES, an individual $\mathcal{X} = (\boldsymbol{x},\sigma) \in \mathbb{R}^n \times \mathbb{R}_{>0}$ consists of a search point and an associated mutation strength.

Let $\mu : \mathbb{N} \to \mathbb{N}$ such that $\mu = \text{poly}(n)$. The (μ+1) ES for minimization of $f : \mathbb{R}^n \to \mathbb{R}$ works as follows: For a given initialization of the population of μ individuals, the following evolution loop is performed:

1. Choose one of the μ individuals in the (current) population uniformly at random. Let this be $\mathcal{C} = (\boldsymbol{c}, \sigma_{\boldsymbol{c}})$.
2. Create a new search point $\boldsymbol{y} := \boldsymbol{c} + \boldsymbol{m} \in \mathbb{R}^n$, where the distribution of the isotropic mutation (vector) $\boldsymbol{m}$ depends only on $\sigma_{\boldsymbol{c}}$.
3. Evaluate $f(\boldsymbol{y})$ and decide whether $\sigma_{\boldsymbol{c}}$ is to be increased, or decreased, or kept unchanged; adapt $\sigma_{\boldsymbol{c}}$ accordingly.
4. Create the mutant $\mathcal{Y} := (\boldsymbol{y}, \sigma_{\boldsymbol{c}})$
   (i. e., $\mathcal{Y}$ inherits the possibly updated/adapted mutation strength $\sigma_{\boldsymbol{c}}$ from its parent $\mathcal{C}$).
5. Discard one of the $\mu + 1$ individuals by uniformly choosing one of the worst individuals (maximal f-value when minimizing).
6. GOTO 1.

Again, in practice the GOTO would be conditioned on some termination criterion. Furthermore, for the generation of the mutant and the adaptation of the mutation strength σ the same properties as stated for the (1+λ) ES must be met.

We are interested in how fast the population, namely the best individual in the population, evolves. Which one of the individuals in the population is the best one changes (usually) over time, of course.

The (μ+1)-selection method is sometimes also referred to as *steady-state selection.*

1.2.4 Additional Notes, Notions, and Notations

We can obtain two (1+1) ES: the (1+λ) ES with $\lambda := 1$ and the (μ+1) ES with $\mu := 1$. These two (1+1) ES differ in one aspect: In the (1+λ) ES with $\lambda := 1$, whenever the mutant of the current search point is at least as good as its parent, the mutant replaces its parent and becomes the new/next current search point. In the (μ+1) ES with $\mu := 1$, however, if the mutant and its parent have equal f-values, both have a 50-50 chance to survive and to become the new current search point (in fact, the new single-individual population).

If the function to be optimized is such that the probability of hitting the level set of a search point (the set containing all search points with the same function value) with an isotropic mutation is zero anyway, this difference is meaningless, though. Namely, for such functions and a mutation adaptation that precludes mutation vectors with zero length, the mutant and its parent have different f-values (with probability one), so that the difference in the selection mechanism could not be observed anyway.

However, in this work, "(1+1) ES" means "(1+λ) ES with $\lambda := 1$." Moreover, "$(1, \lambda)$ ES" means "$(1, \lambda)$ ES with a global mutation strength." Whenever a self-adaptive variant is considered, we will explicitly use the term "$(1, \lambda)\sigma$SA-ES." "$(1\overset{+}{,}\lambda)$ ES" stands for "(1+λ) ES and/or $(1, \lambda)$ ES with global mutation strength."

Finally, note that the stochastic process induced by an $(1, \lambda)\sigma$SA-ES is necessarily Markovian, whereas the stochastic process induced by a $(1\overset{+}{,}\lambda)$ ES (with a global mutation strength) is not necessarily Markovian (and in most cases it is actually not).

1.3 Underlying Publications

This dissertation bases on the following publications:

1. J. J. (2003): Analysis of a Simple Evolutionary Algorithm for Minimization in Euclidean Spaces. In *Proceedings of the 30th International Colloquium on Automata, Languages, and Programming (ICALP 2003)*, Springer LNCS 2719, pp. 1068–1079.

2. J. J. (2005): Rigorous Runtime Analysis of the (1+1) ES: 1/5-Rule and Ellipsoidal Fitness Landscapes. In *Foundations of Genetic Algorithms: 8th International Workshop, FOGA 2005, Revised Selected Papers*, Springer LNCS 3469, pp. 260–281.

   This work has been expanded and extended:

3. J. J. (2006): How the (1+1) ES Using Isotropic Mutations Minimizes Positive Definite Quadratic Forms. *Theoretical Computer Science*, 361(1):38–56.

4. C. Witt and J. J. (2005): Rigorous Runtime Analysis of a (μ+1) ES for the Sphere Function. In *Proceedings of the 2005 Conference on Genetic and Evolutionary Computation (GECCO 2005)*, ACM Press, pp. 849–856.

5. J. J. (2005): On the Complexity of Overcoming Gaps with Isotropic Mutations and Elitist Selection. In *Proceedings of the 2005 IEEE Congress on Evolutionary Computation (CEC 2005)*, IEEE Press, pp. 206–213.

6. J. J. (2006): Probabilistic Runtime Analysis of ($1^+,\lambda$) Evolution Strategies Using Isotropic Mutations. In *Proceedings of the 2006 Conference on Genetic and Evolutionary Computation (GECCO 2006)*, ACM Press, pp. 461–468.

The article that has emerged from joint work with Carsten Witt is due to both authors to almost the same extent in ideas, proofs, and writing.

2 Preliminaries

We recall some notions concerning probability measures/distributions from Feller (1971).

Definition 2.1. Let F denote a probability distribution over $\mathbb{R}^n$. A point in $\mathbb{R}^n$ is called an *atom* (of F) if it carries positive mass (w. r. t. F). We call the distribution F
concentrated on a set $S \subseteq \mathbb{R}^n$ if $\mathbb{R}^n \setminus S$ has zero probability (i. e. zero mass w. r. t. F);
singular if it is concentrated on a set with Lebesgue measure zero;
atomic if it is concentrated on the set of its atoms;
absolutely continuous (w. r. t. Lebesgue's measure) if there exists a function $\mathsf{D}_F : \mathbb{R}^n \to \mathbb{R}$ such that for any Borel set $S \subseteq \mathbb{R}^n$ the probability of S (i. e. the mass of S w. r. t. F) is given by the Lebesgue integral $\int_{\boldsymbol{x} \in S} \mathsf{D}_F(\boldsymbol{x}) \mathrm{d}\boldsymbol{x}$. In such a case D_F is called the probability density (function) corresponding to the probability distribution F.

Note that any probability distribution can be decomposed into a linear combination of three distributions, one of which is absolutely continuous, one of which is singular without atoms, and one of which is atomic (Lebesgue/Jordan decomposition; cf. Feller (1971, pp. 138, 142)). (For distributions over $\mathbb{R}$, "atomic" and "singular" means the same; for distributions over $\mathbb{R}^n$ with $n \geq 2$, however, this is not the case.)

Definition 2.2. Let X and Y denote random variables.

- *X stochastically dominates Y*, in short "$X \succ Y$," if (and only if) $\mathsf{P}\{X \leq a\} \leq \mathsf{P}\{Y \leq a\}$ for all $a \in \mathbb{R}$.
- If $X \succ Y$ as well as $Y \succ X$, i. e., $\forall a \in \mathbb{R}\colon \mathsf{P}\{X \leq a\} = \mathsf{P}\{Y \leq a\}$, then we write "$X \sim Y$."
- We call a random variable X *symmetric* if (and only if) $-X \sim X$.

It is readily seen that, if $X \succ Y$ and $\mathsf{E}[X]$ exists, then $\mathsf{E}[Y] \leq \mathsf{E}[X]$. Obviously, stochastic dominance is a transitive relation.

Now we come to a very useful tool for probabilistic analyses: Hoeffding's bound; see also Hofri (1987, Section 2.6.2).

Theorem 2.3. Hoeffding (1963, Theorem 2): Let $X_1, \ldots, X_k$ denote independent random variables, each with bounded range. For $i \in \{1, \ldots, k\}$ let $[a_i, b_i] \subset \mathbb{R}$ be the range of X_i, where $a_i < b_i$. Let $S := X_1 + \cdots + X_k$. Then for any $x > 0$

$$\mathsf{P}\{S \geq \mathsf{E}[S] + x\} \quad \leq \quad \exp\left(\frac{-2x^2}{\sum_{i=1}^{k}(b_i - a_i)^2}\right).$$

Note that $\sum_{i=1}^{k}(b_i - a_i)^2 \leq k \cdot (b-a)^2$ with $a := \min\{a_i\}$ and $b := \max\{b_i\}$, i. e., the values of all X_i lie in $[a,b] \subset \mathbb{R}$ (with probability one). Furthermore, we directly obtain

$$\mathsf{P}\{S \leq \mathsf{E}[S] - x\} \quad \leq \quad \exp\left(\frac{-2x^2}{k \cdot (b-a)^2}\right)$$

and, consequently,

$$\mathsf{P}\{|S - \mathsf{E}[S]| \geq x\} \quad \leq \quad 2 \cdot \exp\left(\frac{-2x^2}{k \cdot (b-a)^2}\right).$$

In particular, if the range of $X_1, \ldots, X_k$ is $[0,1]$, for instance when considering the number of successful Bernoulli or Poisson trials, then $\mathsf{P}\{S \geq \mathsf{E}[S] + x\} \leq \mathrm{e}^{-2x^2/k}$. As an example, the probability of observing at least $n/2 + \sqrt{n}$ heads in n independent (and fair) coin flips is at most $\mathrm{e}^{-2} < 0.14$. As another example, the probability that at least $0.6n$ of the n flips show head is bounded from above by $\mathrm{e}^{-n/50}$. This might look like a weak bound. As n grows, however, the probability drops rapidly (as it is exponentially small).

In some cases, for discrete 0-1-variables the Chernoff bounds yield better estimates for the tail probability, cf. Motwani and Raghavan (1995, Section 4.1: The Chernoff Bound). However, when we apply Hoeffding's bound to discrete random variables, we will use the term "Chernoff's bound."

Another inequality which helps with the estimation of tail probabilities is due to Markov, cf. Motwani and Raghavan (1995, Theorem 3.2):

Theorem 2.4. (Markov's Inequality) Let X denote a non-negative random variable. Then for all $t \in \mathbb{R}_{>0}$: $\mathsf{P}\{X \geq t\} \leq \mathsf{E}[X]/t$.

If one knows about the variance of a random variable, then a result by Chebyshev can be useful, cf. Motwani and Raghavan (1995, Theorem 3.3):

Theorem 2.5. (Chebyshev's Inequality) Let X denote a random variable. If $\mathsf{E}[X]$ exists and $\mathsf{Var}[X] < \infty$, then for any $t \in \mathbb{R}_{>0}$: $\mathsf{P}\left\{|X - \mathsf{E}[X]| \geq t \cdot \sqrt{\mathsf{Var}[X]}\right\} \leq 1/t^2$.

Note that $\sqrt{\mathsf{Var}[X]}$ is the *standard deviation* of the random variable X.

An *indicator variable* $\mathbb{1}_S$ associated with a set $S \subseteq M$ is a mapping from M into $\{0,1\}$ (i. e. a 0-1-variable) such that $\forall x \in M\colon \mathbb{1}_S(x) = 1 \iff x \in S$. For instance, M may denote $\mathbb{R}$; then $\mathbb{1}_{\mathbb{R}_{\geq 0}}(x) = 1$ if $x \geq 0$ and $\mathbb{1}_{\mathbb{R}_{\geq 0}}(x) = 0$ if $x < 0$. Thus, in such cases (when M is clear from the context), we may write "$\mathbb{1}_{\{x \geq 0\}}$" instead of "$\mathbb{1}_{\mathbb{R}_{\geq 0}}(x)$" for instance. In particular, we may apply an indicator (variable) to a random variable X, and we let $X^+ := X \cdot \mathbb{1}_{\{X \geq 0\}}$ as well as $X^- := X \cdot \mathbb{1}_{\{X < 0\}}$, so that X^+ is a non-negative random variable and X^- is a non-positive random variable. Note that, as a consequence, $\mathsf{E}[\mathbb{1}_{\{X \leq a\}}] = \mathsf{P}\{X \leq a\}$ for all $a \in \mathbb{R}$. If $\mathsf{E}[X^+]$ exists, then $\mathsf{E}[X^+] \geq \mathsf{E}[X \cdot \mathbb{1}_{\{X \geq a\}}]$ for all $a \in \mathbb{R}$, and in particular, $\mathsf{E}[X^+] \geq \mathsf{E}[X]$.

For a symmetric random variable X, we have $-X^- \sim X^+$ (and in particular $\mathsf{E}[X] = 0$), so that applying Markov's inequality to X^+ (and $-X^-$) yields $\mathsf{P}\{|X| \geq t\} \leq \mathsf{E}[X^+]/t$ for all $t > 0$.

Definition 2.6. A probability $p(n)$ is *exponentially small* in n if there is a constant $\varepsilon > 0$ such that $p(n) = \exp(-\Omega(n^{\varepsilon}))$. An event $\mathcal{E}(n)$ happens *with overwhelming probability (w. o. p.)* with respect to n if $1 - \mathsf{P}\{\mathcal{E}(n)\}$ is exponentially small in n.

We say that a statement $Z(x)$, where $x \in \mathbb{R}$, holds *for x large enough* if $(\exists\, x' \in \mathbb{R})(\forall\, x \geq x')\, Z(x)$.

Let f and g denote functions in $\mathbb{R}$. Recall the following asymptotic notations (as $x \to \infty$) when $g(x), h(x) > 0$ for x large enough, cf. Motwani and Raghavan (1995, Definition B.1):

- $g(x) = O(h(x))$ if there exists a constant $\kappa > 0$ such that $g(x) \leq \kappa \cdot h(x)$ for x large enough,
- $g(x) = \Omega(h(x))$ if $h(x) = O(g(x))$,
- $g(x) = \Theta(h(x))$ if $g(x)$ is $O(h(x))$ as well as $\Omega(h(x))$,
- $g(x) = o(h(x))$ if $g(x)/h(x) \to 0$ as $x \to \infty$,
- $g(x) = \omega(h(x))$ if $h(x) = o(g(x))$,
- $g(x) \asymp h(x)$ if $g(x)/h(x) \to 1$ as $x \to \infty$,
- $g(x) = \mathrm{poly}(x)$ if there exists a constant c such that $g(x) = O(x^c)$.

Note that $g \asymp h$ implies $g = \Theta(h)$ (as well as $h = \Theta(g)$, of course), yet that $g = \Theta(h)$ does not even imply the existence of $\lim_{x\to\infty} g(x)/h(x)$ as shown by the example $g(x) := x \cdot (2 + \sin x)$ and $h(x) := x$.

3 Isotropic Mutations

3.1 Isotropic Probability Distributions

Definition 3.1. Let a vector $\boldsymbol{x}$ be distributed according to some distribution F over $\mathbb{R}^n$. Then F is *spherically symmetric* (or *isotropic*) if it is invariant w. r. t. orthonormal transformations, i. e., for any orthogonal matrix $\boldsymbol{M}$ (i. e. $\boldsymbol{M}^\top \boldsymbol{M} = \boldsymbol{I}$) the distribution of $\boldsymbol{M}\boldsymbol{x}$ equals the one of $\boldsymbol{x}$, namely F. Then $\boldsymbol{x}$ is called *isotropically distributed* over $\mathbb{R}^n$.

The nice property of isotropically distributed vectors is that their (possibly) random length is independent of their random direction and that the direction is "uniformly random." Formally, this can be stated as follows:

Proposition 3.2. Let $\boldsymbol{u} \in \mathbb{R}^n$ be uniformly distributed over the unit hyper-sphere[a]. A vector $\boldsymbol{x}$ is isotropically distributed if and only if there exists a non-negative random variable ℓ (independent of $\boldsymbol{u}$) such that the distribution of $\boldsymbol{x}$ equals the one of $\ell \cdot \boldsymbol{u}$.

[a] By "hyper-sphere" we mean the *geometrical* n-dimensional sphere (n-sphere) in Euclidean n-space. From a topologist's point of view, however, our geometric n-sphere is an instance of a topological $(n-1)$-sphere (since our geometric n-sphere is an $(n-1)$-dimensional sub-manifold of an n-space, namely of $\mathbb{R}^n$).

A proof can be found in Fang, Kotz, and Ng (1990, Sec. 2.1). That the direction is "uniformly random" is intuitive. The main idea why the length of an isotropically distributed vector $\boldsymbol{x}$ is independent of its direction reads in short: We pick a direction by picking a half-line L starting at the origin. Then we obtain a conditional distribution by assuming that $\boldsymbol{x} \in L$. Since the mapping $\boldsymbol{x} \mapsto \boldsymbol{M}\boldsymbol{x}$ defined by the multiplication with an orthogonal matrix $\boldsymbol{M}$ (an orthonormal transformation) is a bijection in $\mathbb{R}^n$ which preserves the inner product (implying $|\boldsymbol{x}| = |\boldsymbol{M}\boldsymbol{x}|$), this conditional distribution is invariant w. r. t. the choice of L. Namely, we obtain the same conditional distribution *independent* of the choice of "the direction" L. Hence, we have just found the distribution of ℓ.

Definition 3.3. We call a vector $\boldsymbol{u}$ which is uniformly distributed upon the unit hyper-sphere $\{\boldsymbol{x} \in \mathbb{R}^n \mid |\boldsymbol{x}| = 1\}$ a *unit isotropic mutation (vector).*

If the distribution of an isotropically distributed vector is singular (like the one of a unit isotropic mutation), then ℓ's distribution is atomic (for instance, for a unit isotropic mutation, ℓ is concentrated on the singleton $\{1\}$). If the distribution is absolutely continuous, then also the distribution of the corresponding random variable ℓ is absolutely continuous. There are more direct consequences of the definition of isotropy:

Proposition 3.4. An atomic distribution is isotropic if and only if it is concentrated on the origin.

An absolutely continuous probability distribution F over $\mathbb{R}^n$ is isotropic if (and only if) for all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$: $|\boldsymbol{x}| = |\boldsymbol{y}| \implies \mathsf{D}_F(\boldsymbol{x}) = \mathsf{D}_F(\boldsymbol{y})$.

Let the random vector $\boldsymbol{m} \in \mathbb{R}^n$ be distributed according to a distribution $F_{\boldsymbol{m}}$ which is singular over $\mathbb{R}^n$ and has no atoms. Then $\boldsymbol{m}$ is isotropically distributed if and only if there exists a countable set $L \subset \mathbb{R}_{>0}$ such that $F_{\boldsymbol{m}}$ is concentrated on $\{\boldsymbol{x} \in \mathbb{R}^n \mid |\boldsymbol{x}| \in L\}$ such that, under the condition $|\boldsymbol{m}| = \ell \in L$, the vector $\boldsymbol{m}$ is uniformly distributed upon the hyper-sphere $\{\boldsymbol{x} \in \mathbb{R}^n \mid |\boldsymbol{x}| = \ell\}$.

Lemma 3.5. Let the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ be independently (not necessarily identically) isotropically distributed over $\mathbb{R}^n$. Then $\boldsymbol{z} := \boldsymbol{x} + \boldsymbol{y}$ is also isotropically distributed over $\mathbb{R}^n$.

Proof. Since $\boldsymbol{x}$ and $\boldsymbol{y}$ are isotropically distributed, respectively, for any choice of an orthogonal matrix $\boldsymbol{M}$, the distribution of $\boldsymbol{x}$ equals the one $\boldsymbol{M}\boldsymbol{x}$ and the one of $\boldsymbol{y}$ equals the one of $\boldsymbol{M}\boldsymbol{y}$. Because of the independence, the distribution of $\boldsymbol{x} + \boldsymbol{y}$ equals the one of $\boldsymbol{M}\boldsymbol{x} + \boldsymbol{M}\boldsymbol{y}$, and since $\boldsymbol{M}\boldsymbol{x} + \boldsymbol{M}\boldsymbol{y} = \boldsymbol{M}(\boldsymbol{x} + \boldsymbol{y})$, the distribution of $\boldsymbol{x} + \boldsymbol{y}$ in fact equals the one of $\boldsymbol{M}(\boldsymbol{x} + \boldsymbol{y})$—for any choice of an orthogonal matrix $\boldsymbol{M}$, precisely matching the definition of isotropy. □

By induction, we directly obtain

Corollary 3.6. Let the vectors $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ be independently (not necessarily identically) isotropically distributed over $\mathbb{R}^n$. Then the distribution of the vector $\boldsymbol{y} := \boldsymbol{x}_1 + \cdots + \boldsymbol{x}_k$ is also isotropic.

So, we know that adding two independent isotropically distributed vectors results in a vector that is also isotropically distributed. Hence, we know that all directions are "equiprobable" (actually "equidense"). However, the result tells us nothing about the distribution of the length. And in fact, isotropy is preserved already when the directions of the isotropically distributed vectors that we add are independent, i. e., the length distributions need not necessarily be independent. Therefore, let $\boldsymbol{x}$ be isotropically distributed over $\mathbb{R}^n$, and let $\boldsymbol{y}$ be distributed according to an isotropic mutation that may depend on $|\boldsymbol{x}|$ but that is independent of $\boldsymbol{x}$'s direction, i. e., $\boldsymbol{y}$'s distribution is parameterized and we use the notation "$\boldsymbol{y}_{\boldsymbol{x}}$" to indicate this. Then, given an orthogonal matrix $\boldsymbol{M}$, we have $\boldsymbol{M}\boldsymbol{x} \sim \boldsymbol{x}$ and, in particular, $|\boldsymbol{M}\boldsymbol{x}| \sim |\boldsymbol{x}|$. Consequently, due to our assumptions on how $\boldsymbol{y}$'s distribution may depend on $\boldsymbol{x}$, we have $\boldsymbol{y}_{\boldsymbol{x}} \sim \boldsymbol{y}_{\boldsymbol{M}\boldsymbol{x}}$. Thus, $\boldsymbol{x} + \boldsymbol{y}_{\boldsymbol{x}} \sim \boldsymbol{M}\boldsymbol{x} + \boldsymbol{y}_{\boldsymbol{M}\boldsymbol{x}} \sim \boldsymbol{M}\boldsymbol{x} + \boldsymbol{y}_{\boldsymbol{x}}$. Since moreover $\boldsymbol{y}_{\boldsymbol{x}} \sim \boldsymbol{M}\boldsymbol{y}_{\boldsymbol{x}}$ whatever the value of $\boldsymbol{x}$, we have $\boldsymbol{x} + \boldsymbol{y}_{\boldsymbol{x}} \sim \boldsymbol{M}\boldsymbol{x} + \boldsymbol{M}\boldsymbol{y}_{\boldsymbol{x}}$, i. e. $\boldsymbol{x} + \boldsymbol{y}_{\boldsymbol{x}} \sim \boldsymbol{M}(\boldsymbol{x} + \boldsymbol{y}_{\boldsymbol{x}})$. Since this holds for any choice of the orthogonal matrix $\boldsymbol{M}$, we have just shown that $\boldsymbol{x} + \boldsymbol{y}_{\boldsymbol{x}}$ is isotropically distributed. By induction, we obtain

Lemma 3.7. Consider a sequence $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_k$ of isotropically distributed vectors, where the distribution of $\boldsymbol{x}_i$ may depend on $|\boldsymbol{x}_{i-1}|$ (but not on the direction of $\boldsymbol{x}_{i-1}$) for $i \in \{2, \ldots, k\}$. Then the vector obtained by subsequently adding these vectors is isotropically distributed.

This property will be very useful in the reasoning for the lower bounds on the number of isotropic mutations which are necessary to obtain a reduction of the approximation error.

3.2 Gaussian Mutations

Gaussian mutations date back to the very first application of evolutionary strategies. Namely, they were used in the original (1+1) ES by Rechenberg and Schwefel.

Definition 3.8. Let each of the n components of the random vector $\widetilde{\boldsymbol{m}}$ over $\mathbb{R}^n$ be independently standard-normally distributed.

We call the random vector $\widetilde{\boldsymbol{m}}$ a *Gaussian mutation (vector).* For a given $\sigma \in \mathbb{R}_{>0}$, the random vector $\sigma \cdot \widetilde{\boldsymbol{m}}$ is called a *scaled Gaussian mutation (vector).*

As one may have already guessed, Gaussian mutations bear the following property:

Proposition 3.9. A (scaled) Gaussian mutation vector is isotropically distributed.

Proof. As the components of $\widetilde{\boldsymbol{m}}$ are independently standard-normally distributed, the density at $\boldsymbol{x} \in \mathbb{R}^n$ equals

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

Hence, vectors of equal length have the same density; obviously, the scaling does not affect this property. □

The distribution of $|\widetilde{\boldsymbol{m}}|$, of the random length of a Gaussian mutation vector, is well known. It is a χ-distribution with n degrees of freedom, cf. Arfken (1990). Its density at $x \in \mathbb{R}_{\geq 0}$ equals $x^{n-1} \cdot \mathrm{e}^{-x^2/2} \cdot 2^{1-n/2} / \Gamma(n/2)$ (where "Γ" denotes the well-know Gamma-function), forming a unimodal density having its mode at $\sqrt{n-1}$ and two inflection points at $\sqrt{n-1/2 \pm \sqrt{2n-7/4}}$ for $n \geq 3$. As a consequence, for $x \geq \sqrt{2n}$ the density drops exponentially so that large deviations are not probable. More precisely:

Lemma 3.10. For a scaled Gaussian mutation $\boldsymbol{m} = \sigma \cdot \widetilde{\boldsymbol{m}}$ over $\mathbb{R}^n$ with $\sigma \in \mathbb{R}_{>0}$

$$\mathsf{E}[|\boldsymbol{m}|] \begin{cases} \asymp & \sigma \cdot \sqrt{n} \\ \leq & \sigma \cdot \sqrt{n} \\ \geq & \sigma \cdot \sqrt{n-1/2}. \end{cases}$$

Let $\bar{\ell}$ abbreviate $\mathsf{E}[|\boldsymbol{m}|]$. For $\delta > 0$

$$\mathsf{P}\left\{ \left| |\boldsymbol{m}| - \bar{\ell} \right| \geq \delta \cdot \bar{\ell} \right\} \;\leq\; \frac{1}{\delta^2 \cdot (2n-1)}.$$

Let $\boldsymbol{m}_1, \ldots, \boldsymbol{m}_k$ denote k independent instances of $\boldsymbol{m}$. For any constant $\varepsilon > 0$ there exist two constants $a_\varepsilon, b_\varepsilon > 0$ such that, for the index set $I := \{i \in \{1, \ldots, k\} \mid a_\varepsilon \cdot \bar{\ell} \leq |\boldsymbol{m}_i| \leq b_\varepsilon \cdot \bar{\ell}\}$, we have $\mathsf{P}\{\#I < k \cdot (1-\varepsilon)\} = \mathrm{e}^{-\Omega(k)}$.

Proof. The random variable $|\widetilde{\boldsymbol{m}}|$ is χ-distributed (with n degrees of freedom), and hence,

$$\mathsf{E}[|\widetilde{\boldsymbol{m}}|] \;=\; \sqrt{2} \cdot \frac{\Gamma(n/2+1/2)}{\Gamma(n/2)} \;\in\; \left[\sqrt{n-1/2}, \sqrt{n}\right]$$

(cf. Haagerup (1982) for the bracketing of the fraction involving the Gamma function).

Furthermore, since the random variable $|\widetilde{\boldsymbol{m}}|^2$ is χ^2-distributed, we have $\mathsf{E}\left[|\widetilde{\boldsymbol{m}}|^2\right] = n$, and hence, we can bound the variance of the length of a Gaussian mutation:

$$\mathsf{Var}[|\widetilde{\boldsymbol{m}}|] \;=\; \mathsf{E}\left[|\widetilde{\boldsymbol{m}}|^2\right] - \mathsf{E}[|\widetilde{\boldsymbol{m}}|]^2 \;\leq\; n - \left(\sqrt{n-1/2}\right)^2 \;=\; 1/2$$

(in fact, it has been shown that $\mathsf{Var}[|\widetilde{\boldsymbol{m}}|] \nearrow 1/2$ as $n \to \infty$).

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

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

Since $\mathsf{E}[|\boldsymbol{m}|] = \sigma \cdot \mathsf{E}[|\widetilde{\boldsymbol{m}}|]$ and $\mathsf{Var}[|\boldsymbol{m}|] = \sigma^2 \cdot \mathsf{Var}[|\widetilde{\boldsymbol{m}}|]$, applying this bound to $|\boldsymbol{m}|$, the random length of a scaled Gaussian mutation, yields

$$\mathsf{P}\left\{\left||\boldsymbol{m}| - \bar{\ell}\right| \geq \delta \cdot \bar{\ell}\right\} \;\leq\; \frac{\sigma^2 \cdot 1/2}{(\delta \cdot \sigma \cdot \mathsf{E}[|\widetilde{\boldsymbol{m}}|])^2} \;\leq\; \frac{1/2}{\delta^2 \cdot (n-1/2)}.$$

Finally, we consider k i. i. d. scaled Gaussian mutations. Since $|\boldsymbol{m}| = \Theta(\mathsf{E}[|\boldsymbol{m}|])$ with probability $1 - O(1/n)$ as we have just seen, $\mathsf{E}[\#I] = k - O(k/n)$. Applying Chernoff's bound yields that $\#I$ deviates by a positive constant fraction below its expectation only with probability $\mathrm{e}^{-\Omega(\mathsf{E}[\#I])}$, which is $\mathrm{e}^{-\Omega(k)}$ as n grows. □

3.3 Spatial Gain of an Isotropic Mutation

Since any isotropic mutation can be decomposed into a random direction, on the one hand, and an independent distribution for its length on the other hand, we focus on unit isotropic mutations first.

3.3.1 Spatial Gain of a Unit Isotropic Mutation

Consider an arbitrary but fixed search point $\boldsymbol{c} \in \mathbb{R}^n$ and a unit isotropic mutation $\boldsymbol{u}$ over $\mathbb{R}^n$, and let $\boldsymbol{c}' := \boldsymbol{c} + \boldsymbol{u}$ denote the random mutant. Then this mutant $\boldsymbol{c}'$ is isotropically distributed upon the hyper-sphere $S_{\boldsymbol{c}} := \{\boldsymbol{x} \in \mathbb{R}^n \mid \mathrm{dist}(\boldsymbol{x}, \boldsymbol{c}) = 1\}$, the so-called *mutation sphere*. Furthermore, consider the linear function $\textsc{Sum}_n \colon \mathbb{R}^n \to \mathbb{R}$ defined by

$$\textsc{Sum}_n(\boldsymbol{x}) \;:=\; \sum_{i=1}^{n} x_i \tag{3.1}$$

which is also called OneMax when $\boldsymbol{x} \in \{0,1\}^n$. For a given $a \in \mathbb{R}$ let "$H_{\textsc{Sum}=a}$" denote the hyper-plane $\{\boldsymbol{x} \in \mathbb{R}^n \mid \textsc{Sum}(\boldsymbol{x}) = a\}$, and let "$H_{\boldsymbol{c}}$" abbreviate $H_{\textsc{Sum}=\textsc{Sum}(\boldsymbol{c})}$. Furthermore, for $\rhd \in \{<, >, \leq, \geq\}$ we let "$H_{\textsc{Sum}\rhd a}$" denote the open/closed half-space $\{\boldsymbol{x} \in \mathbb{R}^n \mid \textsc{Sum}(\boldsymbol{x}) \rhd a\}$, and let "$H_{\rhd \boldsymbol{c}}$" abbreviate $H_{\textsc{Sum}\rhd\textsc{Sum}(\boldsymbol{c})}$.

When talking about "the gain" of a mutation or a step, in this section we mean the *spatial gain* of a mutation. The change in the Sum-value is merely used as an indicator whether the mutant of $\boldsymbol{c}$ lies in the one half-space w. r. t. the hyper-plane $H_{\boldsymbol{c}}$ or in the other. In particular, instead of Sum we could have chosen any other linear function that essentially depends on all n

components. In fact, we may chose an arbitrary but fixed hyper-plane containing $\boldsymbol{c}$ since we may rotate $H_{\boldsymbol{c}}$ around $\boldsymbol{c}$. Because of the isotropy of the mutation vector's distribution, nothing would change.

As we focus on isotropically distributed mutation vectors, the larger the length of the mutation vector, the larger the expected distance between the mutant $\boldsymbol{c}'$ and $H_{\boldsymbol{c}}$. Recall that, to focus on the core of the reasoning, we decided to consider unit isotropic mutations for the present. (Later we show how to extend the calculations to (scaled) Gaussian mutations, the length of which follows a (scaled) χ-distribution.) So, the random variable G defined[1] by

$$G := \begin{cases} \operatorname{dist}(\boldsymbol{c}', H_{\boldsymbol{c}}) & \text{if } \boldsymbol{c}' \text{ lies in the (closed) half-space } H_{\leq \boldsymbol{c}} \\ -\operatorname{dist}(\boldsymbol{c}', H_{\boldsymbol{c}}) & \text{if } \boldsymbol{c}' \text{ lies in the (open) half-space } H_{> \boldsymbol{c}} \end{cases} \tag{3.2}$$

corresponds to the *signed distance* of the mutant (generated by a unit isotropic mutation) from the hyper-plane $H_{\boldsymbol{c}}$ (or from any other predefined hyper-plane containing its parent $\boldsymbol{c}$, as we have seen). The nice property of this random variable G is that it maps an "n-dimensional randomness" to a single dimension—leaving just enough information to obtain interesting results as we shall see. As we consider unit isotropic mutations for now, G is concentrated on the interval $[-1,1]$, and naturally, we would like to know G's distribution. In particular, we are interested in how this distribution changes with n, the dimensionality of the search space.

Recall the mutation sphere $S_{\boldsymbol{c}} := \{\boldsymbol{x} \in \mathbb{R}^n \mid \operatorname{dist}(\boldsymbol{x}, \boldsymbol{c}) = 1\}$ in which the mutant $\boldsymbol{c}'$ must lie. Then we have $G \geq g$ for some fixed $g \geq 0$ if $\boldsymbol{c}'$ lies in the hyper-hemisphere $S_{\boldsymbol{c}} \cap H_{\leq \boldsymbol{c}}$ such that $\operatorname{dist}(\boldsymbol{c}', H_{\boldsymbol{c}}) \geq g$. Since all points in the hyper-hemisphere $S_{\boldsymbol{c}} \cap H_{\leq \boldsymbol{c}}$ that have distance g from the hyper-plane $H_{\boldsymbol{c}}$ form an $(n-1)$-sphere lying in some hyper-plane J which is parallel to $H_{\boldsymbol{c}}$ with distance g, the set consisting of all potential mutants that result in $G \geq g$ in fact forms a hyper-spherical cap with height $h := 1 - g$ (cf. the figure on page 22); let $C_{\boldsymbol{c},g}$ denote this cap. For $g = 1$, the cap $C_{\boldsymbol{c},g}$ degenerates to a singleton, and for $g > 1$, obviously $C_{\boldsymbol{c},g}$ is no longer a cap but the empty set. Thus, we concentrate on $g \in [1,0]$ in the following, and since $\boldsymbol{c}'$ is uniformly distributed upon $S_{\boldsymbol{c}}$, we have

$$\mathsf{P}\{G \geq g\} = \frac{(n-1)\text{-volume of } C_{\boldsymbol{c},g}}{(n-1)\text{-volume of } S_{\boldsymbol{c}}}. \tag{3.3}$$

Since G is symmetric, i. e. G and $-G$ follow the same distribution, we have $\mathsf{P}\{G \leq -g\} = \mathsf{P}\{G \geq g\}$ for any $g \in \mathbb{R}$. In particular, $\mathsf{P}\{G \geq g\} = \mathsf{P}\{G > g\}$ because the hyper-plane J containing the boundary of the cap $C_{\boldsymbol{c},g}$ is hit with zero probability (just like any other predefined hyper-plane).

In the following we concentrate on the ratio of the hyper-surface area of a hyper-spherical cap to the one of the hyper-sphere of which this cap is cut off by the intersection with some hyper-plane, namely J. In particular, we are interested in how this ratio depends on the height of the cap and on n, the dimension of the search space.

Therefore, we assume that $\boldsymbol{c}$ coincides with the origin $\boldsymbol{o}$ and use polar/spherical coordinates: Let r denote the distance from the origin, α the azimuthal angle with range $[0, 2\pi)$, and $\beta_3, \ldots, \beta_n$ the remaining angles with range $[0, \pi]$. Here, for a given $\boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{o}\}$, β_i is the angle between

[1] The probability space that underlies G consists of $\mathbb{R}^n$, the corresponding Borel σ-algebra, and the probability measure induced by the distribution of the random mutation vector over $\mathbb{R}^n$.

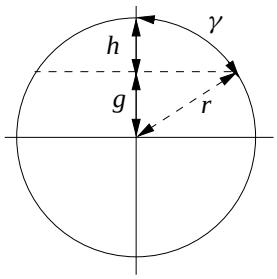
(the positive half of) the ith axis (in the Cartesian coordinate system) and the half-line starting at $\boldsymbol{o}$ and passing through $\boldsymbol{x}$. Let ρ denote an arbitrary permutation on $\{3,\ldots,n\}$. Fixing r in n-space, but none of the angles, defines an n-sphere $S_r^{[n]}$ with radius r; additionally fixing $\beta_{\rho(n)}$ results in an $(n-1)$-sphere $S_r^{[n-1]} \subseteq S_r^{[n]}$ having radius $r \cdot \sin\beta_{\rho(n)}$; fixing $\beta_{\rho(n-1)}$ in addition to r and $\beta_{\rho(n)}$ results in an $(n-2)$-sphere $S_r^{[n-2]} \subseteq S_r^{[n-1]} \subseteq S_r^{[n]}$ with radius $r \cdot \sin\beta_{\rho(n)} \cdot \sin\beta_{\rho(n-1)}$, and so on (cf. Kendall (1961)). Thus, the hyper-surface area of an n-sphere with radius r is given by

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

Re-grouping the factors and solving the α-integral, namely $\int_0^{2\pi}\mathrm{d}\alpha = 2\pi$, yields

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

for the area of an n-sphere with radius r. Naturally, we could have looked up the formula for the hyper-surface area of an n-sphere in a formulary, but we also need a formula for the cap. The formula for the cap can easily be derived from the one above—yet only if one knows about the derivation of the latter.



The area of an n-dimensional spherical cap is calculated by adjusting the upper limit on the angle β_n appropriately. In the figure on the right, the interdependence between the upper limit (γ) on the angle β_n and the height (h) of a spherical cap is shown (where the sheet this figure is drawn on corresponds to the plane spanned by the first and the nth axis when $\alpha = 0$). Consequently, the area of a hyper-spherical cap with radius r and height $h = r\cdot(1-\cos\gamma) \in [0,2r]$, i. e. $\gamma \in [0,\pi]$, is in fact given by

$$r^{n-1}\cdot 2\pi\cdot\left(\int_0^{\gamma}(\sin\beta)^{n-2}\,\mathrm{d}\beta\right)\cdot\left(\prod_{i=1}^{n-3}\int_0^{\pi}(\sin\beta)^i\,\mathrm{d}\beta\right).$$

All in all, in n-space, $n \geq 3$, the ratio of the hyper-surface area of a spherical cap with height $h \in [0,2r]$, on the one hand, to the hyper-surface area of the hyper-sphere with radius r the cap is cut off, on the other hand, reduces to

$$\frac{\int_0^{\gamma}(\sin\beta)^{n-2}\,\mathrm{d}\beta}{\int_0^{\pi}(\sin\beta)^{n-2}\,\mathrm{d}\beta} \quad \text{with} \quad \gamma = \arccos(1-h/r).$$

Since the mutation sphere $S_{\boldsymbol{c}}$ in which the mutant lies has unit radius (i. e. $r = 1$), we have $1-h/r = 1-(1-g)/1 = g$. Thus, for $n \geq 3$, Equation (3.3) on page 21 reads in fact

$$\mathsf{P}\{G \geq g\} \;=\; \frac{\int_0^{\arccos g}(\sin\beta)^{n-2}\,\mathrm{d}\beta}{\int_0^{\pi}(\sin\beta)^{n-2}\,\mathrm{d}\beta} \qquad (\text{for } n \geq 3).$$

Unfortunately, the anti-derivative of $\sin^k$ (i. e. the indefinite integral $\int (\sin x)^k \,\mathrm{d}x$) does not have an "algebraically closed" form. Nevertheless, this identity tells us something: Since arccos and sin are differentiable, we may try to find the density D_G of G, namely $\frac{\mathrm{d}}{\mathrm{d}g}\mathsf{P}\{G \le g\}$.

Therefore, we will now transform the formula that we have just derived into one which makes such an estimation simple and which will turn out useful also in the analysis of the expected spatial gain. Namely, we will concentrate on the probability density of hitting the boundary of the cap $C_{\boldsymbol{c},g} \subset S_{\boldsymbol{c}}$. With the help of this density, we will obtain an alternative formula for the probability of hitting a cap.

Let $\Psi_n(x) := \int_0^x (\sin\beta)^{n-2}\,\mathrm{d}\beta$ and let "Ψ" abbreviate $\Psi_n(\pi)$. Then for the probability distribution of G we obtain

$$\mathsf{P}\{G \le g\} \;=\; 1-\mathsf{P}\{G > g\} \;=\; 1-\frac{\Psi_n(\arccos g)}{\Psi_n(\pi)},$$

and hence, for $g \in (-1,1)$,

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}g}\mathsf{P}\{G \le g\} &= \frac{-1}{\Psi_n(\pi)} \cdot \frac{\mathrm{d}\Psi_n(\arccos g)}{\mathrm{d}g} \\ &= \frac{-1}{\Psi} \cdot \frac{\mathrm{d}}{\mathrm{d}g}\int_0^{\arccos g} (\sin\beta)^{n-2}\,\mathrm{d}\beta.\end{aligned}$$

Let Sin_k denote the anti-derivative of $\sin^k$, i. e. the indefinite integral $\int (\sin x)^k \,\mathrm{d}x$, such that $\mathrm{Sin}_k(0) = 0$. Then

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}g}\int_0^{\arccos g} (\sin\beta)^k\,\mathrm{d}\beta &= \frac{\mathrm{d}\mathrm{Sin}_k(\arccos g)}{\mathrm{d}g} \\ &= \mathrm{Sin}_k'(\arccos g)\cdot \arccos' g \\ &= (\sin(\arccos g))^k \cdot \arccos' g\,,\end{aligned}$$

and since $\sin(\arccos g) = \sqrt{1-g^2}$ and $\arccos' g = -1/\sqrt{1-g^2}$, we obtain for $k \ge 2$

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

All in all, we finally obtain for $n \ge 4$ the probability density of G at $g \in (-1,1)$ in n-space

$$\mathsf{D}_G(g) \;=\; \tfrac{\mathrm{d}}{\mathrm{d}g}\mathsf{P}\{G \le g\} \;=\; \frac{1}{\Psi}\cdot\left(1-g^2\right)^{(n-3)/2} \qquad (\text{for } n \ge 4). \tag{3.4}$$

This density function can now be used to derive an alternative formula for the probability that G is at least g, namely

$$\mathsf{P}\{G \ge g\} \;=\; \frac{1}{\Psi}\cdot\int_g^1 \left(1-x^2\right)^{(n-3)/2}\,\mathrm{d}x \qquad \text{for } g \in [-1,1] \text{ and } n \ge 4. \tag{3.5}$$

Moreover, as a by-product, we obtain $\Psi = \Psi_n(\pi) = \int_{-1}^1 (1-x^2)^{(n-3)/2}\,\mathrm{d}x$. The value of this definite integral equals $\sqrt{\pi}\cdot\Gamma(n/2-1/2)/\,\Gamma(n/2)$, cf. Gradshteyn and Ryzhik (1994) for instance.

Using the bracketing of the ratio of the two Gamma-function values already applied in the proof of Lemma 3.10 (p. 19), we obtain the following bracketing for the normalization factor $1/\Psi$

$$\sqrt{\frac{n-2}{2\pi}} \;\leq\; \frac{1}{\Psi} \;\leq\; \sqrt{\frac{n-1}{2\pi}} \qquad \text{(for } n \geq 4\text{)}, \tag{3.6}$$

which implies $1/\Psi \asymp \sqrt{n}/\sqrt{2\pi} \approx 0.4\sqrt{n}$.

Unfortunately, as one may expect, also $(1-x^2)^{(n-3)/2}$—like the function Ψ_n—does not have an "algebraically closed" anti-derivative. (We see clearly now that the probability $\mathsf{P}\{G \geq g\}$ drops exponentially as $g \to 1$, though.) However, the function $x \cdot (1-x^2)^{(n-3)/2}$ has an anti-derivative, namely $(1-x^2)^{(n-1)/2}/(1-n)$ for $n \geq 4$. Thus, for instance, we can compute the expected distance of the mutant $\boldsymbol{c}'$ from the hyper-plane $H_{\boldsymbol{c}}$, which equals $\mathsf{E}[G^+] - \mathsf{E}[G^-] = 2 \cdot \mathsf{E}[G^+]$, where we use the symmetry of the random variable G (recall that G^+ and G^- abbreviate $G \cdot \mathbb{1}_{\{G \geq 0\}}$ resp. $G \cdot \mathbb{1}_{\{G \leq 0\}}$). More generally, we obtain the following result.

Lemma 3.11. Let G denote the random variable as defined in Equation (3.2) on page 21. Then for $g \in [0,1]$ and $n \geq 4$

$$\mathsf{E}[G \cdot \mathbb{1}_{\{G \geq g\}}] \;=\; \frac{(1-g^2)^{(n-1)/2}}{(n-1)\cdot\Psi} \begin{cases} \asymp & (1-g^2)^{(n-1)/2}/\sqrt{2\pi n} \\ < & (1-g^2)^{(n-1)/2} \cdot 0.4/\sqrt{n-1} \\ > & (1-g^2)^{(n-1)/2} \cdot 0.3989/\sqrt{n+1}. \end{cases}$$

Proof. As we have already noted above, $(1-x^2)^{(n-1)/2}/(1-n)$ is an anti-derivative of the function $x \cdot (1-x^2)^{(n-3)/2}$. Hence,

$$\begin{aligned} \mathsf{E}[G \cdot \mathbb{1}_{\{G \geq g\}}] \;&=\; \frac{1}{\Psi} \cdot \int_g^1 x \cdot (1-x^2)^{(n-1)/2}\,\mathrm{d}x \\ &=\; \frac{1}{\Psi} \cdot \left[\frac{-1}{n-1} \cdot (1-x^2)^{(n-1)/2}\right]_g^1 \\ &=\; \frac{1}{\Psi \cdot (n-1)} \cdot (1-g^2)^{(n-1)/2}. \end{aligned}$$

Using the bracketing of $1/\Psi$ (Inequality (3.6) on page 24), we obtain

$$\frac{1}{\Psi\cdot(n-1)} \begin{cases} \leq & \sqrt{(n-1)/2\pi}\,\big/(n-1) \;=\; 1/\sqrt{2\pi(n-1)} \;<\; 0.4/\sqrt{n-1} \\ \geq & \sqrt{(n-2)/2\pi}\,\big/(n-1) \;\geq\; 1/\sqrt{2\pi(n+1)} \;>\; 0.3989/\sqrt{n+1} \end{cases}$$

(using $\sqrt{n-2}\big/(n-1) \geq 1/\sqrt{n+1}$ for $n \geq 3$). □

This lemma tells us that for the expected distance of the mutant from $H_{\boldsymbol{c}}$ (or from any other predefined hyper-plane containing its parent) $\mathsf{E}[|G|] \asymp 2/\sqrt{2\pi n} \approx 0.8/\sqrt{n}$. This might appear bewildering (at first) since this implies that, as the search space's dimensionality increases, the expected distance from $H_{\boldsymbol{c}}$ tends to zero—although the distance of $\boldsymbol{c}'$ form $\boldsymbol{c}$ is fixed to one and $H_{\boldsymbol{c}}$ is hit with zero probability. However, noting that $H_{\boldsymbol{c}}$ is an affine subspace with dimension $n-1$ (i. e. codimension 1), it may become more plausible that getting far away from $H_{\boldsymbol{c}}$ becomes

less and less probable as n increases. It might also help to recall that an n-hypercube with unit diameter (longest diagonal) has edges of length $1/\sqrt{n}$.

Let us come back to the probability $\mathsf{P}\{G \geq g\}$ as given in Equation (3.5) on page 23. Although we may not be able to compute the integral (algebraically in a closed form), we may approximate the integral's value. Namely, upper and lower bounds on the value of the integral $\int_g^1 (1-x^2)^{(n-3)/2}\,\mathrm{d}x$ must be derived—in dependence on g and n.

Lemma 3.12. Let G denote the random variable as defined in Equation (3.2) on page 21.

1. For $n \geq 9$ and $g\colon \mathbb{N} \to \mathbb{R}$ such that $g(n) \in [\varepsilon/\sqrt{n}, 1/3]$ for some constant $\varepsilon > 0$,

$$\mathsf{P}\{G \geq g\} \begin{cases} > & \frac{g}{\Psi} \cdot \exp(-g^2 \cdot 4n) \\ < & \frac{g}{\Psi} \cdot \frac{\exp(-g^2 \cdot n/3)}{1-\exp(-g^2 \cdot n)} \end{cases} = \frac{g}{\Psi} \cdot \exp(-g^2 \cdot n/3) \cdot \Theta(1)$$

   so that $\mathsf{P}\{G \geq g\} = \sqrt{n} \cdot g \cdot \mathrm{e}^{-\Theta(g^2 \cdot n)}$. Furthermore,

2. $0 \geq g = o(1/\sqrt{n}) \implies \mathsf{P}\{G \geq g\} \to 1/2$ as $n \to \infty$,

3. $g \geq 1/3 \implies \mathsf{P}\{G \geq g\} = \mathrm{e}^{-\Omega(n)}$,

4. $\mathsf{P}\{G \geq g\} = \Omega(1) \iff g = O(1/\sqrt{n})$,

5. $1/2 - \mathsf{P}\{G \geq g\} = \Omega(1) \iff g = \Omega(1/\sqrt{n})$.

Proof. Let $\beta/\sqrt{n}$ substitute g. Then for $\beta \in [\varepsilon, \sqrt{n}/3]$ and $n \geq 9$, on the one hand,

$$\begin{aligned}
\Psi \cdot \mathsf{P}\{G \geq \beta/\sqrt{n}\} &= \int_{\beta/\sqrt{n}}^{1} (1-x^2)^{(n-3)/2}\,\mathrm{d}x \\
&\geq \int_{\beta/\sqrt{n}}^{2\beta/\sqrt{n}} (1-x^2)^{(n-3)/2}\,\mathrm{d}x \\
&> \frac{\beta}{\sqrt{n}} \cdot (1-(2\beta)^2/n)^{(n-3)/2} \\
&> \frac{\beta}{\sqrt{n}} \cdot \exp\left(-\frac{(n-3)/2}{n/(2\beta)^2 - 1}\right) \qquad \text{(because } (1-1/m)^{m-1} > 1/\mathrm{e}\text{)} \\
&= \frac{\beta}{\sqrt{n}} \cdot \exp\left(-2\beta^2 \frac{n-3}{n-4\beta^2}\right) \\
&\geq \frac{\beta}{\sqrt{n}} \cdot \exp\left(-4\beta^2\right) \qquad \text{(because } \beta \leq \sqrt{n}/3 \text{ and } n \geq 9\text{)},
\end{aligned}$$

and on the other hand,

$$
\begin{aligned}
\Psi \cdot \mathsf{P}\big\{G \geq \beta/\sqrt{n}\big\} &= \int_{\beta/\sqrt{n}}^{1} (1-x^2)^{(n-3)/2}\,\mathrm{d}x \\
&\leq \sum_{i=1}^{\lfloor\sqrt{n}/\beta\rfloor} \frac{\beta}{\sqrt{n}} \cdot \big(1-(i\beta/\sqrt{n})^2\big)^{(n-3)/2} \qquad \text{(upper sum; width } \beta/\sqrt{n}\text{)} \\
&= \frac{\beta}{\sqrt{n}} \cdot \sum_{i=1}^{\lfloor\sqrt{n}/\beta\rfloor} \big(1-(i\beta)^2/n\big)^{(n-3)/2} \\
&< \frac{\beta}{\sqrt{n}} \cdot \sum_{i=1}^{\infty} \exp\left(-\frac{(n-3)/2}{n/(i\beta)^2}\right) \qquad \text{(because } (1-1/m)^m < 1/\mathrm{e}\text{)} \\
&\leq \frac{\beta}{\sqrt{n}} \cdot \sum_{i=1}^{\infty} \exp(-(i\beta)^2/3) \qquad \text{(because } \frac{n-3}{2n} \geq \frac{1}{3} \text{ for } n \geq 9\text{)} \\
&< \frac{\beta}{\sqrt{n}} \cdot \exp(-\beta^2/3) \cdot \underbrace{\frac{1}{1-\exp(-\beta^2)}}_{=\Theta(1) \text{ since } \beta \geq \varepsilon \in \mathbb{R}_{>0}}
\end{aligned}
$$

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

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

Thus, for $\beta \in [\varepsilon, \sqrt{n}/3]$ (i. e., for $g \in [\varepsilon/\sqrt{n}, 1/3]$ since $\beta = \sqrt{n}\cdot g$) we obtain

$$
\mathsf{P}\{G \geq g\} = \frac{1}{\Psi}\int_{\beta/\sqrt{n}}^{1} (1-x^2)^{(n-3)/2}\,\mathrm{d}x = \frac{1}{\Psi}\cdot\frac{\beta}{\sqrt{n}}\cdot \mathrm{e}^{-\Theta(\beta^2)} = \sqrt{n}\cdot g\cdot \mathrm{e}^{-\Theta(g^2\cdot n)}
$$

since $1/\Psi = \Theta(\sqrt{n})$ (cf. Inequality (3.6) on page 24).

Concerning the second claim, note that $g = o(1/\sqrt{n})$ implies $(1-g^2)^{(n-3)/2} \to 1$ as n grows, and concerning the third claim, we have $(2/3)^{(n-3)/2}/\Psi = \mathrm{e}^{-\Omega(n)}\cdot O(\sqrt{n})$, which is bounded above by $\mathrm{e}^{-\Omega(n)}$.

Finally, for the proof of the fourth and the fifth claim, note that $\sqrt{n}\cdot g\cdot\exp(-\Theta(g^2\cdot n)) = \Theta(1)$ if and only if $g = \Theta(1/\sqrt{n})$. □

3.3.2 Spatial Gain of a Gaussian Mutation

As we have seen, a Gaussian mutation is in fact a unit isotropic mutation which is scaled by multiplying it with a χ-distributed random variable ℓ_χ (with n degrees of freedom and which is independent of the direction given by the unit isotropic mutation). Analogously to the definition of the random variable G (Equation (3.2) on page 21), let $\widetilde{G}$ denote the "signed distance" of $\boldsymbol{c}+\widetilde{\boldsymbol{m}}$ from the hyper-plane $H_{\boldsymbol{c}}$, where $\widetilde{\boldsymbol{m}}$ is a Gaussian mutation vector. Then $\widetilde{G}$'s distribution indeed equals the one of the random variable $\ell_\chi \cdot G$. In particular, we have (for $n \geq 4$ since we apply Lemma 3.11 (p. 24) for the value of $\mathsf{E}[G^+]$)

$$\begin{aligned}
\mathsf{E}\big[\widetilde{G}^+\big] &= \mathsf{E}\big[\ell_\chi\big]\cdot\mathsf{E}\big[G^+\big] \\
&= \sqrt{2}\cdot\frac{\Gamma(n/2+1/2)}{\Gamma(n/2)}\cdot\frac{1}{\Psi\cdot(n-1)} \\
&= \sqrt{2}\cdot\frac{\Gamma(n/2+1/2)}{\Gamma(n/2)}\cdot\frac{\Gamma(n/2)}{\Gamma(n/2-1/2)\cdot\sqrt{\pi}\cdot(n-1)} \\
&= \sqrt{2}\cdot\frac{n/2-1/2}{\sqrt{\pi}\cdot(n-1)} \;=\; \frac{1}{\sqrt{2\pi}} \;=\; 0.3989\ldots
\end{aligned}$$

where we use $\Gamma(n/2+1/2) = (n/2-1/2)\cdot\Gamma(n/2-1/2)$ and the result on $\mathsf{E}[\ell_\chi]$ from the proof of Lemma 3.10 (p. 19). In this case, multiplying the expectations is indeed allowed since we investigate $\mathsf{E}\big[\widetilde{G}\cdot\mathbb{1}_{\{\widetilde{G}>0\}}\big]$. Namely, whether the indicator variable is one or zero is independent of the random variable ℓ_χ since $\mathsf{P}\{\ell_\chi > 0\} = 1$. Or in other words, the indicator variable merely checks whether the random direction points into the half-space $H_{<\boldsymbol{c}}$ or into $H_{>\boldsymbol{c}}$, which is—per definition—independent of the (distribution of the) length. When we are interested in $\mathsf{E}\big[\widetilde{G}\cdot\mathbb{1}_{\{\widetilde{G}\geq g\}}\big]$ for some $g \neq 0$, things become more complicated, of course. Clearly, for $g > 0$, the larger the length of the isotropically distributed vector, the larger the probability that $\widetilde{G}$ exceeds g. Formally, we have the convolution involving the density of the χ-distributed length. Namely, for $g > 0$,

$$\begin{aligned}
\mathsf{P}\big\{\widetilde{G}\geq g\big\} &= \int_g^\infty \mathsf{D}_\chi(x)\cdot\mathsf{P}\{G\geq g/x\}\,\mathrm{d}x \\
&= \frac{2^{1-n/2}}{\Gamma(n/2)}\cdot\int_g^\infty \frac{x^{n-1}}{\mathrm{e}^{x^2/2}}\cdot\mathsf{P}\{G\geq g/x\}\,\mathrm{d}x,
\end{aligned}$$

where the integration starts at g (rather than 0) because $g/x > 1$ for $x < g$ and $\mathsf{P}\{G \geq 1\} = 0$ anyway (in less formal words, if the mutation's length is smaller than g then the mutant's distance from $H_{\boldsymbol{c}}$ must also be smaller than g).

Since $\mathsf{P}\big\{\ell_\chi \in [\sqrt{n}/2, 2\sqrt{n}\,]\big\}$ equals $\int_{\sqrt{n}/2}^{2\sqrt{n}} \mathsf{D}_\chi(x)\,\mathrm{d}x = 1 - O(1/n)$ as implied by Lemma 3.10 (p. 19), by substituting "1" for "$\mathsf{P}\{G \geq g/x\}$" when $x \notin [\sqrt{n}/2, 2\sqrt{n}\,]$ we obtain the upper bound

$$\mathsf{P}\big\{\widetilde{G}\geq g\big\} \;=\; \frac{2^{1-n/2}}{\Gamma(n/2)}\cdot\int_{\sqrt{n}/2}^{2\sqrt{n}} \frac{x^{n-1}}{\mathrm{e}^{x^2/2}}\cdot\mathsf{P}\{G\geq g/x\}\,\mathrm{d}x \quad+\quad O(1/n).$$

By substituting "0" for "$\mathsf{P}\{G \geq g/x\}$" when $x \notin [\sqrt{n}/2, 2\sqrt{n}\,]$ we trivially obtain the lower bound

$$\mathsf{P}\big\{\widetilde{G}\geq g\big\} \;\geq\; \frac{2^{1-n/2}}{\Gamma(n/2)}\cdot\int_{\sqrt{n}/2}^{2\sqrt{n}} \frac{x^{n-1}}{\mathrm{e}^{x^2/2}}\cdot\mathsf{P}\{G\geq g/x\}\,\mathrm{d}x.$$

Thus, in the remaining part of the convolution of $\mathsf{P}\{G \geq g/x)\}$ with the distribution of the random length we have $x \in [\sqrt{n}/2, 2\sqrt{n}\,]$, i. e., $g/x = \Theta(g/\sqrt{n})$. Since $\mathsf{P}\{G \geq g/x)\}$ is bounded from below by $\Omega(1)$ and from above by $1/2 - \Omega(1)$ if and only if $g/x = \Theta(1/\sqrt{n})$ as shown in Lemma 3.12 (p. 25) (items 4 and 5), we directly obtain

Corollary 3.13. Let $\widetilde{G}$ denote (analogously to G given in Equation (3.2) on page 21) the random variable corresponding to the "signed distance" of $\boldsymbol{c}+\widetilde{\boldsymbol{m}}$ from $H_{\boldsymbol{c}}$, where $\widetilde{\boldsymbol{m}}$ is a Gaussian mutation. Then

- $\mathsf{P}\big\{\widetilde{G}\geq g\big\} = \Omega(1) \iff g = O(1)$,
- $1/2 - \mathsf{P}\big\{\widetilde{G}\geq g\big\} = \Omega(1) \iff g = \Omega(1)$.

We note again that this result is no surprise as the χ-distribution shows only very small deviations from its expectation. Therefore recall that its variance is upper bounded by $1/2$, whereas the expectation is $\Theta(\sqrt{n})$. This may become even more clear when we consider a scaled Gaussian mutation $\sigma \cdot \widetilde{\boldsymbol{m}}$ which is scaled such that we expect unit length, which implies $\sigma = \Theta(1/\sqrt{n})$. Then the variance is $O(\sigma^2)$, i. e. $O(1/n)$—which obviously tends to zero as n grows.

3.4 Additional Notes

The random variables G and $\widetilde{G}$ are not tailored to the analysis of a specific function—although we use the linear function SUM in its definition. As has been already noted several times, $H_{\boldsymbol{c}}$ could denote any predefined hyper-plane containing $\boldsymbol{c}$, rather than $\{\boldsymbol{x} \in \mathbb{R}^n \mid \text{SUM}(\boldsymbol{x}) = \text{SUM}(\boldsymbol{c})\}$. Due to the isotropy of a unit isotropic mutation, we would actually end up with the same random variable—or, more precisely, with a random variable having the same distribution as G.

Furthermore, we would like to stress that the random variable G differs from the random variable $\Delta_{\boldsymbol{x}^*}$ corresponding to a unit mutation's spatial gain towards a fixed point $\boldsymbol{x}^* \in \mathbb{R}^n$ (usually the/an optimum). However, as $\text{dist}(\boldsymbol{c}, \boldsymbol{x}^*) \to \infty$, the (sequence of) random variable(s) $\Delta_{\boldsymbol{x}^*}$ converges in distribution to the random variable G. In fact, G stochastically dominates $\Delta_{\boldsymbol{x}^*}$ as we shall see.

Finally, the approach of using $\widetilde{G}$ when Gaussian mutations are considered differs from the commonly followed progress-rate approach at least in one crucial aspect: The reasoning in most progress-rate results is the following: Assume for a moment that $\boldsymbol{c}$ coincides with the origin and that the optimum $\boldsymbol{x}^*$ lies on the positive halve of the first axis. Then the mutation vector can be decomposed into a component pointing towards $\boldsymbol{x}^*$ along the first axis, called *central component* (or *radial component)*, and into a so-called *lateral component* (or *traversal component*) given by the mutant's distance from the first axis. Then the central component of the gain towards $\boldsymbol{x}^*$ is indeed normally distributed—because it is just the first component of the Gaussian mutation vector. The lateral component, however, lies in the hyper-plane spanned by the remaining $n-1$ axes (in fact an $(n-1)$-subspace since $\boldsymbol{c}$ coincides with the origin by assumption). The length of the mutation vector's lateral component, i. e. the mutant's distance from the first axis, is again χ-distributed—with $n-1$ degrees of freedom rather than n, though. As we have seen, the variance of the lateral component's length is by an $O(1/n)$-factor smaller than its expectation. In the very most progress-rate results, this fact is taken as a reason to substitute the expectation of the lateral component's length for the random variable in the calculations. This does significantly ease the calculations since the central component follows an ordinary normal distribution—presumably, one of the best known and best investigated distributions. This simplification, namely the assumption that the lateral component's length were not random, however, rules out the possibility of obtaining theorems on the algorithm's behavior. Rather the results are actually obtained for/in a simplifying model of the stochastic process that is induced by the algorithm under consideration, and simulations become necessary to justify this simplification.

When we consider the random variable $\widetilde{G}$, then the randomness "in all n dimensions" is regarded—rather than only the central component—, and the way to theorems on the algorithm's "true" behavior is still open.

A Note on Isotropic Mutations for Bit-Strings

When EAs for the search space $\{0,1\}^n$ are investigated, the commonly used mutation operator flips each of the n bits independently with some fixed probability p_{mut}, usually $p_{\text{mut}} := 1/n$. However, just like (scaled) Gaussian mutations for $\mathbb{R}^n$, this can be considered a particular type of isotropic binary mutation: The number of bits that flip follows a binomial distribution. However, one may say that the mutation remains isotropic when we choose an arbitrary distribution over $\{0,1,\ldots,n\}$ for the number of bits to be flipped. Let k be distributed according to this distribution (which might depend on the course of the optimization), then a subset of k of the n bits is *uniformly* chosen, and those k bits are flipped. The reason why we may call this an isotropic binary mutation is the following: If we pick a particular bit (and disregard the other $n-1$ bits), then the probability that this bit is actually flipped is independent of our choice. Formally, the mutation of an n-bit-string is associated with a distribution over the power-set of $\{1,\ldots,n\}$. Then we call a mutation isotropic if (and only if) any two subsets of equal cardinality are equiprobable. This implies that the distribution is invariant w. r. t. permutations of the bits' positions in the string (cf. the invariance w. r. t. rotations of the search space in $\mathbb{R}^n$).

Considering adaptation of the mutation operator is rather uncommon when $\{0,1\}^n$ is the search space. In some cases, one wants to consider the best case w. r. t. the mutation operator, and then considering this general notion of isotropic mutations might by useful. In general, in the best case there is a particular number k of bits such that flipping k (uniformly chosen) bits results in maximum success probability for an isotropic mutation, and hence, the best-case assumption would just be to assume that with probability 1 we flip k uniformly distributed bits.

For constant k, choosing $p_{\text{mut}} := k/n$ results in a k-bit-mutation to occur with probability $\Omega(1)$, so that for most (of the interesting?) asymptotic analyses, there might be only small differences between an "optimally adapted isotropic binary mutation" and an optimally chosen p_{mut} for independent bit-flips. However, for large k there might be a substantial difference.

We will come back to this in Section 4.5 (p. 47).

4 General Lower Bounds

In this chapter we will derive lower bounds on the number of isotropic mutations which are necessary to reduce the approximation error in the search space. Namely, in the following, the approximation error (in the search space) is given by $d := \mathrm{dist}(\boldsymbol{c}, \boldsymbol{x}^*)$, the Euclidean distance of the evolving search point $\boldsymbol{c}$ from a fixed search point $\boldsymbol{x}^* \in \mathbb{R}^n$ —for instance the (or a fixed) optimum of a function to be optimized. In particular, we consider the number of mutations to halve this approximation error. The lower bounds that we will obtain hold independently of the function to be optimized, i. e., they are valid for any function scenario. To follow the reasoning, however, one may keep in mind the minimization of $\textsc{Sphere}\colon \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ defined as the squared distance from the origin. This function scenario is in some sense a best case since a mutation results in a gain w. r. t. the SPHERE-value if and only if it results in a spatial gain towards the minimum in the search space. Halving the distance from the optimum corresponds to a SPHERE-gain of 75%, i. e., the function-value is quartered. Besides, this example shows that a lower bound on the runtime w. r. t. the reduction of the approximation error in the search space usually implies a lower-bound result on the reduction of the approximation error w. r. t. the function value (this implied bound may be weak, though; we shall see an example for this effect later, namely when we investigate positive definite quadratic forms).

Moreover, the lower bounds on the runtime we are going to show will be valid independently of the adaptation of the mutation strength. In fact, they will be independent of the length-distribution of the isotropic distribution that is used to generate mutants. For instance, the length could be distributed according to a (scaled) Cauchy distribution, rather than according to a (scaled) χ-distribution (with n degrees of freedom) when Gaussian mutations are used.

One may ask whether lower bounds that hold in such a general sense may be too general, i. e. too weak, so that common concrete mutation mechanisms just cannot achieve a runtime which is upper bounded by the same order, i. e., which is at most by a constant factor larger than the lower bound. This is not necessarily the case as we will see in the chapter where concrete scenarios are investigated.

We will start off with a closer look at the spatial gain which a single isotropic mutation may yield, since a general upper bound on the expectation of this gain will enable us to obtain various lower-bound results.

4.1 Spatial Gain Towards a Fixed Search Point

When we want to prove a lower bound on the number of mutations which are necessary to realize a certain reduction of the approximation error, an upper bound on the expected spatial gain towards $\boldsymbol{x}^*$ in a single step is needed. So far we have considered the signed distance of the mutant from a fixed hyper-plane which contains its parent. In the following reasoning, let $H_{\boldsymbol{c}}$ denote the hyper-

plane that contains $\boldsymbol{c}$ and lies perpendicular to the line passing through $\boldsymbol{c}$ and $\boldsymbol{x}^*$. Essentially, we have considered the random variable G (defined in Equation (3.2) on page 21) which bases on a unit isotropic mutation. Let G_ℓ denote the random variable defined just like G except for the length of the mutation vector $\boldsymbol{m}$ being fixed to $\ell > 0$ rather than to 1, i. e., $\boldsymbol{m}$ is isotropically distributed such that $\mathsf{P}\{|\boldsymbol{m}| = \ell\} = 1$. Then $G_\ell \sim \ell \cdot G$ since this is just a rescaling of the situation. Then the random variable

$$\Delta_{\boldsymbol{x}^*,\ell} \quad := \quad \operatorname{dist}(\boldsymbol{c}, \boldsymbol{x}^*) - \operatorname{dist}(\boldsymbol{c} + \boldsymbol{m}, \boldsymbol{x}^*) \tag{4.1}$$

corresponds to the spatial gain towards $\boldsymbol{x}^*$. (Note that this is not to be mixed up with the so-called "central component" of a mutation as discussed in Section 3.4 (p. 28).)

The interdependence between the signed distance (g) from $H_{\boldsymbol{c}}$ and the gain (δ) towards $\boldsymbol{x}^*$ is depicted in the following figure.

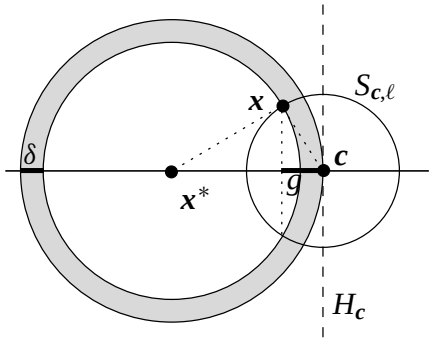


Figure 4.1: Interrelation between δ (gain towards $\boldsymbol{x}^*$) and g (signed distance from $H_{\boldsymbol{c}}$)

Obviously (and as we have seen), the larger the length of an isotropic mutation, the larger the expected distance from the hyper-plane $H_{\boldsymbol{c}}$. Recall that d is defined as $\operatorname{dist}(\boldsymbol{c}, \boldsymbol{x}^*)$. The best possible gain towards $\boldsymbol{x}^*$ is ℓ—if $\ell \leq d$. If $\ell > d$, however, the best possible gain towards $\boldsymbol{x}^*$ is $2d - \ell$ since all mutants have distance at least $\ell - d$ from $\boldsymbol{x}^*$. The least possible gain towards $\boldsymbol{x}^*$ is $-\ell$, independently of how ℓ relates to d. All in all, the range of $\Delta_{\boldsymbol{x}^*,\ell}$ is $[-\ell, \min\{\ell, 2d - \ell\}]$. (Hence, in particular, the gain towards $\boldsymbol{x}^*$ is always negative if $\ell > 2d$.)

Now, note this trivial but essential geometric fact:

Fact 4.1. The spatial gain δ towards $\boldsymbol{x}^*$ corresponding to the signed distance g (from the hyper-plane that contains $\boldsymbol{c}$ and lies perpendicular to the line passing through $\boldsymbol{c}$ and $\boldsymbol{x}^*$) cannot be larger than g.

Since $\Delta_{\boldsymbol{x}^*,\ell} \geq \delta$ implies $G_\ell \geq g(\delta)$, where $g(\delta)$ denotes the g that corresponds to the specified $\delta \in [\ell, \min\{\ell, 2d - \ell\}]$, and since $g(\delta) \geq \delta$ as just noticed, this trivial observation directly implies that $\mathsf{P}\left\{\Delta_{\boldsymbol{x}^*,\ell} \geq \delta\right\} \leq \mathsf{P}\{G_\ell \geq g(\delta)\} \leq \mathsf{P}\{G_\ell \geq \delta\}$. In other words, $\Delta_{\boldsymbol{x}^*,\ell} \prec G_\ell$. Note that this stochastic dominance holds for any fixed length ℓ. As a consequence, the dominance indeed holds for any distribution of $|\boldsymbol{m}|$, i. e., for arbitrary isotropic mutations. We have just obtained

Proposition 4.2. Consider an arbitrary but fixed search point $\boldsymbol{x}^* \in \mathbb{R}^n$. Let the mutation vector be distributed according to an arbitrary isotropic distribution F.

Let the random variable $\Delta_{\boldsymbol{x}^*,F}$ be defined (analogously to Equation (4.1) on page 32) as the mutation's spatial gain towards $\boldsymbol{x}^*$, and let the random variable G_F be defined (analogously to Equation (3.2) on page 21) as the mutant's signed distance from the hyper-plane $H_{\boldsymbol{c}}$.

Then $\Delta_{\boldsymbol{x}^*,F} \prec G_F$ (i. e., G_F stochastically dominates $\Delta_{\boldsymbol{x}^*,F}$).

Due to the isotropy of the mutation vector $\boldsymbol{m}$, for any point $\boldsymbol{x}^{**} \in \mathbb{R}^n$ that has the same distance (namely d) from $\boldsymbol{c}$ as $\boldsymbol{x}^*$, we have $\Delta_{\boldsymbol{x}^*,F} \sim \Delta_{\boldsymbol{x}^{**},F}$. Because of this invariance, it makes sense to use the subscript "d" rather than "$\boldsymbol{x}^*$." Furthermore, we may drop the subscript F since the dominance holds for any F (as long as F is isotropic, of course).

Naturally, one may ask how the random variables Δ_{d_1} and Δ_{d_2} relate when, say, $d_1 < d_2$. One may already guess that $\Delta_{d_1} \prec \Delta_{d_2}$. As this might not be that obvious, the concrete correspondence between δ and g will be derived in the following. Therefore, reconsider Figure 4.1 (p. 32) and assume that the length of the isotropic mutation happens to be ℓ. Furthermore, we define $M_\delta := \{\boldsymbol{x} \mid \operatorname{dist}(\boldsymbol{x},\boldsymbol{c}) = \ell \wedge \operatorname{dist}(\boldsymbol{x},\boldsymbol{x}^*) = d - \delta\}$ as the set which consists of all potential mutants that are exactly δ closer to $\boldsymbol{x}^*$ than $\boldsymbol{c}$. For $\delta < -\ell$ and/or $\delta > \min\{\ell, 2d-\ell\}\}$, M_δ is empty since such gains are impossible, and for $\delta = -\ell$ and/or $\delta = \min\{\ell, 2d-\ell\}\}$, M_δ is a singleton. Finally, for $-\ell < \delta < \min\{\ell, 2d-\ell\}$, M_δ forms an $(n-1)$-sphere; namely, M_δ is the intersection of the two hyper-spheres $S_{\boldsymbol{c},\ell}$ (the mutation sphere) and $S_{\boldsymbol{x}^*,d-\delta}$ (consisting of all points having distance $d-\delta$ from $\boldsymbol{x}^*$).

Now, using Pythagoras, we obtain that $\ell^2 - g^2$ as well as $(d-\delta)^2 - (g-d)^2$ equal the squared radius of M_δ. Solving the equation $\ell^2 - g^2 = (d-\delta)^2 - (g-d)^2$ for g yields the correspondence

$$g \quad = \quad \delta + \frac{\ell^2 - \delta^2}{2d} \qquad \text{for } \delta \in [-\ell, \min\{\ell, 2d-\ell\}]. \tag{4.2}$$

As we can see now, the additive term by which g (the gain away from $H_{\boldsymbol{c}}$) must be larger than the corresponding δ (the spatial gain towards $\boldsymbol{x}^*$), namely $\ell^2 - \delta^2/(2d)$, is indeed anti-proportional to d, the distance from $\boldsymbol{x}^*$. Since, on the one hand, $\mathsf{P}\{\Delta_{d,\ell} \geq \delta\} = 1$ for any $\delta \leq -\ell$ and, on the other hand, $\mathsf{P}\{\Delta_{d,\ell} \geq \delta\} = 0$ for any $\delta \geq \min\{\ell, 2d-\ell\}$ anyway, we have indeed

$$\mathsf{P}\{\Delta_{d_1} \geq \delta\} \quad \leq \quad \mathsf{P}\{\Delta_{d_2} \geq \delta\} \quad \text{when} \quad d_1 \leq d_2$$

for any/arbitrary $\delta \in \mathbb{R}$. As our choice of ℓ in the above reasoning was again arbitrary, the inequality that we derived above does not only hold for any isotropic mutation of an arbitrarily fixed length but for arbitrary isotropic mutations. We obtain the following result (which is not at all a surprise, yet it will be of great help):

Proposition 4.3. Consider two arbitrary but fixed search points $\boldsymbol{x}^*, \boldsymbol{x}^{**} \in \mathbb{R}^n$. The search point $\boldsymbol{c}$ is mutated by adding a vector which is distributed according to an arbitrary isotropic distribution F. Then $\operatorname{dist}(\boldsymbol{x}^*,\boldsymbol{c}) \leq \operatorname{dist}(\boldsymbol{x}^{**},\boldsymbol{c})$ implies $\Delta_{\boldsymbol{x}^*,F} \prec \Delta_{\boldsymbol{x}^{**},F}$.

The stochastic-dominance relations that we have derived for the various random variables induced by an isotropic mutation will be frequently used in numerous reasonings and calculations.

As another consequence of the interrelation between the signed distance from $H_{\boldsymbol{c}}$ and the spatial gain towards a fixed search point (Equation (4.2) on page 33), we see that $\Delta_{d,\ell} \geq 0$ implies

$G_\ell \geq \ell^2/(2d)$, and hence, $\mathsf{P}\big\{\Delta_{d,\ell} \geq 0\big\}$, the probability of the mutant being at least as close to $\boldsymbol{x}^*$ as $\boldsymbol{c}$, is upper bounded by $\mathsf{P}\big\{G_\ell \geq \ell^2/2d\big\}$. Furthermore, utilizing the stochastic-dominance relation, we directly obtain that

$$\mathsf{E}\big[\Delta^+_{d,\ell}\big] \quad \leq \quad \mathsf{E}[G_\ell \cdot \mathbb{1}_{\{G_\ell \geq \ell^2/(2d)\}}].$$

Using Lemma 3.11 (p. 24) (and the fact that $G_\ell \sim \ell \cdot G$), we obtain for $n \geq 4$

$$\mathsf{E}[G_\ell \cdot \mathbb{1}_{\{G_\ell \geq \ell^2/(2d)\}}] \quad \leq \quad \ell \cdot 0.4 \cdot \big(1-(\ell/(2d))^2\big)^{(n-1)/2} \big/ \sqrt{n-1},$$

and thus, by substituting x for $\ell/(2d)$, we have (for $n \geq 4$)

$$\mathsf{E}\big[\Delta^+_{d,\ell}\big] \quad \leq \quad \frac{0.8\,d}{\sqrt{n-1}} \cdot x \cdot (1-x^2)^{(n-1)/2} \tag{4.3}$$

where $x \in (0,1)$. Consider n to be fixed for a moment. It is readily seen that $x \cdot (1-x^2)^{(n-1)/2}$ has a unique maximum, and since

$$\begin{aligned}
\frac{\mathrm{d}}{\mathrm{d}x}\, x \cdot (1-x^2)^{(n-1)/2} &= (1-x^2)^{(n-1)/2} - x^2 \cdot (n-1) \cdot (1-x^2)^{(n-1)/2-1} \\
&= (1-x^2)^{(n-1)/2-1} \cdot \big((1-x^2) - x^2 \cdot (n-1)\big) \\
&= (1-x^2)^{(n-1)/2-1} \cdot (1-x^2 \cdot n),
\end{aligned}$$

solving $1-x^2 \cdot n = 0$ for x yields that $x \cdot (1-x^2)^{(n-1)/2}$ takes its maximum at $1/\sqrt{n}$. Substituting "$1/\sqrt{n}$" for "$x \cdot (1-x^2)^{(n-1)/2}$" in the RHS of Inequality (4.3) on page 34, we obtain for $n \geq 4$

$$\mathsf{E}\big[\Delta^+_{d,\ell}\big] \quad \leq \quad \frac{0.8\,d}{\sqrt{n-1}} \cdot \frac{1}{\sqrt{n}} \cdot (1-1/n)^{(n-1)/2} \quad \leq \quad \frac{0.8\,d}{n-1} \cdot (3/4)^{3/2} \quad < \quad \frac{0.52\,d}{n-1}.$$

Note that also this bound holds independently of the length ℓ of the isotropic mutation, i. e., it holds for any isotropic mutation $\boldsymbol{m}$ with $\mathsf{P}\{|\boldsymbol{m}| = \ell\} = 1$. Thus, the bound indeed holds for arbitrary distributions of $|\boldsymbol{m}|$, i. e., for any isotropic mutation. Finally, note that—for any random variable X—we have $\mathsf{E}[X^+] \geq \mathsf{E}[X \cdot \mathbb{1}_{\{X \geq a\}}]$ for any $a \in \mathbb{R}$. Thus, we have shown the following result:

Lemma 4.4. Consider the optimization of an arbitrary function $f : \mathbb{R}^n \to \mathbb{R}$. Let $\boldsymbol{x}^* \in \mathbb{R}^n$ denote an arbitrary but fixed point (for instance an optimum of f, if one exists). Let $\boldsymbol{c}$ denote the current search point to which an isotropic mutation $\boldsymbol{m}$ is added, resulting in the mutant $\boldsymbol{c}' = \boldsymbol{c} + \boldsymbol{m}$. Then—independently of the distribution of $|\boldsymbol{m}|$ and independently of the selection rule, which decides whether $\boldsymbol{c}'$ replaces $\boldsymbol{c}$ or not—the expected spatial gain of this step (mutation followed by selection) towards $\boldsymbol{x}^*$ is smaller than $0.52 \cdot \mathrm{dist}(\boldsymbol{c}, \boldsymbol{x}^*)/(n-1)$ for $n \geq 4$.

This lemma tells us that, even when the length of an isotropic mutation and the selection rule are chosen optimally (i. e. such that the expected gain of the mutation followed by selection is maximum), the approximation error (in the search space, w. r. t. $\boldsymbol{x}^*$) is reduced at most by a $\frac{0.52}{n-1}$-fraction. This general *upper* bound on the expected best-case one-step gain of a mutation can now be turned into a general *lower* bound on the expected number of steps which are necessary to realize a certain reduction of the approximation error in the search space (defined as the Euclidean distance from a certain search point).

Before we will do this, however, we will prove that a gain which is "considerably" larger than the best-case expected gain is very unlikely:

Lemma 4.5. Let $\boldsymbol{x}^* \in \mathbb{R}^n$ denote an arbitrary but fixed point and $\boldsymbol{c} \neq \boldsymbol{x}^*$ the current search point to which an isotropic mutation $\boldsymbol{m}$ is added, i. e., $d := \operatorname{dist}(\boldsymbol{c}, \boldsymbol{x}^*) > 0$. Then, for any constant $\varepsilon \in (0,1]$, independently of the distribution of $|\boldsymbol{m}|$, the probability that the mutant is such that $d - \operatorname{dist}(\boldsymbol{c}+\boldsymbol{m}, \boldsymbol{x}^*) = \Omega(d \cdot n^\varepsilon/n)$ is bounded from above by $\mathrm{e}^{-\Omega(n^\varepsilon)}$, i. e., the mutant's distance from $\boldsymbol{x}^*$ is by an $\Omega(n^\varepsilon/n)$-fraction smaller than the one of its parent only with an exponentially small probability.

Proof. Assume that the length of the isotropic mutation $\boldsymbol{m}$ is $\ell > 0$. Then with $\delta := d \cdot n^\varepsilon/n$ Equation (4.2) on page 33 tells us that

$$g_\delta \;=\; \frac{\ell^2}{2d} + \delta - \frac{\delta^2}{2d} \;=\; \frac{\ell^2}{2d} + \frac{dn^\varepsilon}{n}\left(1 - \frac{n^\varepsilon}{2n}\right) \;\geq\; \frac{\ell^2}{2d} + \frac{dn^\varepsilon}{2n}.$$

Since the two summands in our lower bound on g_δ are equal for $\ell = d \cdot n^{(\varepsilon-1)/2}$, we obtain that when $\ell > d \cdot n^{(\varepsilon-1)/2}$, then "$\frac{\ell^2}{2d}$" is the larger summand, and when $\ell < d \cdot n^{(\varepsilon-1)/2}$, then "$\frac{dn^\varepsilon}{2n}$" is the larger summand.

For $\ell \geq d \cdot n^{(\varepsilon-1)/2}$, i. e. $d \leq \ell \cdot n^{(1-\varepsilon)/2}$, we have $g_\delta \geq \ell^2/(2d) \geq (\ell/2) \cdot n^{(\varepsilon-1)/2}$, whereas for $\ell \leq d \cdot n^{(\varepsilon-1)/2}$, i. e. $d \geq \ell \cdot n^{(1-\varepsilon)/2}$, we have $g_\delta \geq dn^\varepsilon/(2n) \geq (\ell/2) \cdot n^{(\varepsilon-1)/2}$. In other words, $g_\delta \geq (\ell/2) \cdot n^{(\varepsilon-1)/2}$ for any length $\ell > 0$ of the mutation vector $\boldsymbol{m}$, and thus,

$$\mathsf{P}\big\{\operatorname{dist}(\boldsymbol{c}+\boldsymbol{m}, \boldsymbol{x}^*) \leq d - d \cdot n^\varepsilon/n\big\} \;\leq\; \mathsf{P}\big\{G_\ell \geq (\ell/2) \cdot n^{(\varepsilon-1)/2}\big\}.$$

This probability is bounded from above by $\mathrm{e}^{-\Omega(n^\varepsilon)}$ according to Lemma 3.12 (p. 25).

Finally, it is readily checked that this asymptotic upper bound on the probability does not only hold for a δ of exactly $d \cdot n^\varepsilon/n$, but for any δ that is $\Omega(d \cdot n^\varepsilon/n)$. □

Like the upper bound on the expected gain of a mutation, also this bound on the gain of a mutation can be turned into a lower bound on the number of mutations which are necessary to reach a certain reduction of the approximation error. Before we do so, however, we focus on the expected gain and on the expected number of steps again.

4.2 Lower Bound on the Expected Number of Steps of (1+1) ESs

Recall our framework for (1+1) ESs from Section 1.2 (p. 8). In the following, $\boldsymbol{c}^{[i]}$ denotes the evolving individual after i steps and we let $d^{[i]}$ denote the approximation error in the search space given by $\operatorname{dist}(\boldsymbol{c}^{[i]}, \boldsymbol{x}^*)$ after i steps. Then $d^{[0]}$ is the initial approximation error. Moreover, in this section let α be such that $\alpha \cdot d$ is the best-case expected one-step gain (i. e., $\alpha = \max_{\ell>0} \mathsf{E}[\Delta^+_{d=1,\ell}]$) for which we have just proved that $\alpha < 0.52/(n-1) = O(1/n)$. Note that, because of the scaling invariance of the situation, $\alpha = \max_{d>0} \mathsf{E}[\Delta^+_d]/d$ with the length of the underlying mutation being fixed to an arbitrary positive length, i. e., α is well defined.

Our best-case assumptions on the step length and the selection rule obviously result in the largest possible expected one-step gain—yet one may ask whether the "greedy" assumption of

assuming the best case for each of a number of steps does indeed result in a best-case multi-step assumption. Therefore, consider two successive steps and assume that in the first step a (possibly negative) spatial gain of $\delta^{[1]}$ is realized. Then, assuming the best-case for the second/final step, we obtain for the minimum (best possible) expected approximation error after the two steps (under the condition of the gain in the first step being $\delta^{[1]}$)

$$\mathsf{E}\left[d^{[2]} \mid \delta^{[1]}\right] \;=\; (1-\alpha)\cdot d^{[1]} \;=\; (1-\alpha)\cdot(d^{[0]}-\delta^{[1]}) \;=\; (1-\alpha)\cdot d^{[0]} - (1-\alpha)\cdot\delta^{[1]}.$$

Obviously, the larger $\delta^{[1]}$, the smaller the expected final approximation error. Let $\Delta^{[1]}$ denote the random variable corresponding to the spatial gain of the first step (mutation *and* selection). Using the linearity of expectation, we obtain $\mathsf{E}[d^{[2]}] = (1-\alpha)\cdot d^{[0]} - (1-\alpha)\cdot\mathsf{E}[\Delta^{[1]}]$, and hence, applying the one-step best-case assumption also to the first step indeed results in the expected final approximation error to be minimum. Namely, after two steps we have in the best case (w. r. t. the expected approximation error)

$$\mathsf{E}\left[d^{[2]}\right] = (1-\alpha)\cdot d^{[0]} - (1-\alpha)\cdot(\alpha\cdot d^{[0]}) \;=\; (1-\alpha)^2\cdot d^{[0]}.$$

By induction we obtain that in the best case—namely when in each step the length of the mutation was such that $\mathsf{E}[\Delta^+]$ is maximum and the selection was such that a mutation is accepted if and only if the approximation error is decreased—after k steps the expected approximation error is $(1-\alpha)^k\cdot d^{[0]}$. Since $(1-\alpha)^k \geq 1-\alpha\cdot k$, the smallest number of steps k such that $\mathsf{E}[d^{[k]}] \leq d^{[0]}/2$ is at least $\frac{1/2}{\alpha} > \frac{1/2}{0.52/(n-1)} > 0.96\,(n-1)$.

So, now we know a lower bound on the number of steps which are necessary until we expect the approximation error to be halved. However, in general, maximizing the *expected total gain* need not necessarily result in minimizing the *expected number of steps* to realize a specified gain (for instance, to halve the approximation error). Nevertheless, $0.5/\alpha$ (which is larger than $0.96\,(n-1)$ as we have already seen) will turn out to be a lower bound on the expected number of steps which are necessary to halve the approximation error. The proof will be easy once we know about the following lemma, which is a modification of Wald's equation (see Feller (1971, Formula (2.8) in Chapter 12), for instance).

Lemma 4.6. Let $X_1, X_2, \ldots$ denote random variables with bounded range and S the random variable defined by $S = \min\{t \mid X_1 + \cdots + X_t \geq g\}$ for a given $g > 0$. Given that S is a stopping time (i. e., the event $\{S = t\}$ depends only on $X_1, \ldots, X_t$), if $\mathsf{E}[S] < \infty$ and $\mathsf{E}[X_i \mid S \geq i] \leq u \neq 0$ for $i \in \mathbb{N}$, then $\mathsf{E}[S] \geq g/u$.

Proof. First of all note that (unlike in Wald's equation) the X_i need not be independent—making the assumption necessary that S is a stopping time, though.

Obviously $S \geq 1$, and for $i \geq 2$, the condition "$S \geq i$" is equivalent to "$X_1 + \cdots + X_k < g$ for $k \in \{1, \ldots, i-1\}$." Since the X_i are bounded, $\mathsf{E}[X_1 + \cdots + X_S] < \infty$ if $\mathsf{E}[S] < \infty$. The proof follows the one of Wald's equation (up to the point where the upper bound on $\mathsf{E}[X_i \mid S \geq i]$ is utilized rather than the original assumption that the X_i are i. i. d.).

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

□

Before we apply this lemma to prove the lower bound on the expected number of steps which are necessary to halve the approximation error, however, we will show that also when assuming the best case w. r. t. the expected number of steps, we can assume that mutations which result in a larger approximation error are always discarded. Therefore, let $\boldsymbol{x}^* \in \mathbb{R}^n$ be an arbitrary (but fixed) point and assume that a (1+1) ES minimizes the function $f : \mathbb{R}^n \to \mathbb{R}$ with $\boldsymbol{c} \mapsto \operatorname{dist}(\boldsymbol{c}, \boldsymbol{x}^*)$ using isotropic mutations.

Assume that the spatial gain towards $\boldsymbol{x}^*$ in the first step (mutation *and* selection) is $\delta^{[1]} < 0$ so that $d^{[1]} = d^{[0]} - \delta^{[1]} > d^{[0]}$. Let L be the distribution of the mutation's length which is used by

the original (1+1) ES in the second step to mutate the search point (which was generated in the first step) at distance $d^{[1]} = d^{[0]} - \delta^{[1]} (> d^{[0]})$ from $\boldsymbol{x}^*$. Then we claim that discarding the first mutation and instead using the scaled length distribution $L' := L \cdot (d^{[0]}/d^{[1]})$ in the second step to mutate the initial individual anew results in a "better" distribution of the mutant that is generated in the second step (*before* selection). Formally, we consider the random variable $\delta^{[1]} + \Delta_{d^{[1]},L}$ for the original process (where $\delta^{[1]} < 0$ is fixed) and the random variable $\Delta_{d^{[0]},L'}$ for the alternative process. We will show that the latter one is "better" in the sense that it stochastically dominates the random variable $\delta^{[1]} + \Delta_{d^{[1]},L}$ (which describes the original process). Therefore, note that, because of the scaling invariance of the situation, we have

$$\Delta_{d^{[1]},L} \quad \sim \quad \tfrac{d^{[1]}}{d^{[0]}} \cdot \Delta_{d^{[0]},L'}.$$

Thus, in the alternative process the total spatial gain after the second mutation (*before* selection) is at least δ' with exactly the same probability with which in the original process a total spatial gain of at least $\delta^{[1]} + \delta' \cdot d^{[1]}/d^{[0]}$ occurs—for any $\delta' \in \mathbb{R}$. Since $\delta^{[1]} < 0$ (by assumption) and $d^{[0]} > 0$, the following inequalities are equivalent:

$$\begin{aligned}
\delta^{[1]} + \delta' \cdot d^{[1]}/d^{[0]} &< \delta' \\
\delta' \cdot (d^{[0]} - \delta^{[1]})/d^{[0]} &< \delta' - \delta^{[1]} \\
\delta' \cdot (1 - \delta^{[1]}/d^{[0]}) &< \delta' - \delta^{[1]} \\
\delta' \cdot (-\delta^{[1]}/d^{[0]}) &< -\delta^{[1]} \\
\delta' &< d^{[0]}.
\end{aligned}$$

Obviously, reducing the approximation error by more than the distance from $\boldsymbol{x}^*$ is impossible, and $\boldsymbol{x}^*$ is hit with zero probability anyway. Thus, indeed $\delta' < d^{[0]}$ with probability one. Consequently, a gain of at least δ' is realized in the alternative process *with exactly the same probability* with which in the original process the *smaller* gain of at least $\delta^{[1]} + \delta' \cdot d^{[1]}/d^{[0]}$ is realized. This directly implies the claimed stochastic dominance relation:

Proposition 4.7. Let $d^{[0]} > 0$ as well as $\delta^{[1]} < 0$ be fixed, and let $d^{[1]} := d^{[0]} - \delta^{[1]}$ $(> d^{[0]})$. For any length distribution (non-negative random variable) L, we have $\Delta_{d^{[0]},L \cdot d^{[0]}/d^{[1]}} \succ \delta^{[1]} + \Delta_{d^{[1]},L}$.

So, up to now we considered the first step (consisting of a mutation followed by selection) and the second mutation (without selection). For the selection in the second step, we obtain by the same reasoning that it is again "best" to discard the mutation in this second step if it results in a negative gain, and so on. By induction, we obtain that after *any* number of steps, the *total* gain of the alternative (imaginary) process (in which mutations resulting in a negative gain are always discarded) stochastically dominates the total gain of the original process. In other words, for *any* number of steps k, the probability of realizing a predefined reduction of the approximation error within the first k steps is at least as large for the alternative process as for the original process. This directly implies that the random number of steps which are necessary to realize this reduction for the original process stochastically dominates the respective random variable for the alternative process. As a simple consequence, we obtain that we expect the original process to perform at least as many steps (to realize the predefined reduction of the approximation error) as the alternative process needs in expectation.

Now we can easily prove the lower bound on the expected number of steps:

Theorem 4.8. Let $\boldsymbol{x}^* \in \mathbb{R}^n$ be an arbitrary (but fixed) point. Let a (1+1) ES minimize the function $f\colon \mathbb{R}^n \to \mathbb{R}$, $n \geq 4$, with $\boldsymbol{c} \mapsto \operatorname{dist}(\boldsymbol{c}, \boldsymbol{x}^*)$ (or any other function) using isotropic mutations and assume that in each step the distribution of the mutation's length as well as the selection rule are such that the expected number of steps until $\operatorname{dist}(\boldsymbol{c}^{[t]}, \boldsymbol{x}^*) \leq \operatorname{dist}(\boldsymbol{c}^{[0]}, \boldsymbol{x}^*)/2$ for the first time is minimum. Then this expected number of steps is larger than $0.96 \cdot (n-1)$.

Correspondingly, the expected number of steps until $\operatorname{dist}(\boldsymbol{c}^{[t]}, \boldsymbol{x}^*) \leq \operatorname{dist}(\boldsymbol{c}^{[0]}, \boldsymbol{x}^*)/2^{b(n)}$ for the first time, where $b\colon \mathbb{N} \to \mathbb{N}$, is larger than $b(n) \cdot (0.96n - 2) + 1$.

Proof. For the application of Lemma 4.6 (p. 36) we let X_i denote the random variable which corresponds to the spatial gain in ith step (mutation and selection). As we have just seen, we can assume that mutations which result in a negative gain are always discarded. Consequently, the distance from $\boldsymbol{x}^*$, i. e. the approximation error, will never exceed $d^{[0]}$ (the initial approximation error). As a further consequence, the X_i are bounded, namely $0 \leq X_i \leq d^{[0]}$.

We choose $g := d^{[0]}/2$ and note that S is a stopping time in our case. Lemma 4.4 (p. 34) gives the upper bound $\mathsf{E}[X_i] \leq d^{[0]} \cdot \alpha < d^{[0]} \cdot \frac{0.52}{n-1}$, and hence, we choose $u := d^{[0]} \cdot \frac{0.52}{n-1}$. Then the lower bound g/u on the expected number of steps necessary to halve the approximation error (from Lemma 4.6 (p. 36)) finally solves to $(d^{[0]}/2)\big/(d^{[0]} \cdot \frac{0.52}{n-1}) > 0.96 \cdot (n-1) > 0.96n - 1$.

Due to the linearity of expectation, the expected number of steps to halve the approximation error b times is lower bounded by $(0.96n-1) + (b-1) \cdot (0.96n - 1 - 1)$, where the rightmost "$-1$" emerges because the last step within a halving-phase is also (and must be counted as) the first step of the following halving-phase. □

Now that we know that $\Omega(n)$ steps are necessary *in expectation* to halve the approximation error in the search space, we would like to know whether there is a good chance of getting by with considerably fewer steps, i. e., we want a bound on the probability that a certain number of steps does—or, does not—suffice to halve the approximation error.

4.3 Lower Bound for (1+,λ) ESs which Holds with Overwhelming Probability

As in the previous section, we concentrate on the number of steps to halve the approximation error in the search space, i. e. the distance from a predefined search point $\boldsymbol{x}^* \in \mathbb{R}^n$. However, now we want to obtain a lower bound on the number of steps which holds with a certain probability, namely with overwhelming probability, i. e., the probability that fewer steps suffice is exponentially small.

Therefore recall Lemma 4.5 (p. 35). This lemma indeed almost directly implies the following lower-bound result:

Theorem 4.9. Let a (1+,λ) ES using isotropic mutations and an arbitrary mutation adaptation optimize an arbitrary function. Let $\boldsymbol{x}^*$ denote some fixed point (for instance the/a fixed optimum). Given that $d := \operatorname{dist}(\boldsymbol{c}^{[0]}, \boldsymbol{x}^*) > 0$, for $b : \mathbb{N} \to \mathbb{N}$ such that $b = \operatorname{poly}(n)$ and any two constants $\kappa, \varepsilon > 0$, the probability that within $\kappa \cdot b(n) \cdot n^{1-\varepsilon}$ steps (i. e. $\lambda \cdot \kappa \cdot b(n) \cdot n^{1-\varepsilon}$ mutations) a search point $\boldsymbol{c}^{[i]}$ with $\operatorname{dist}(\boldsymbol{c}^{[i]}, \boldsymbol{x}^*) \leq d/2^{b(n)}$ is generated is upper bounded by $\mathrm{e}^{-\Omega(n^{\varepsilon})}$.

Proof. We can focus on the number of iterations to halve the approximation error since the total error probability is bounded from above by $b \cdot \mathrm{e}^{-\Omega(n^\varepsilon)}$, which is $\mathrm{e}^{-\Omega(n^\varepsilon)}$.

Assume that $\kappa\, n^{1-\varepsilon}$ steps suffice to halve the approximation error. Then at least one step must yield a gain of at least $(d/2)/(\kappa\, n^{1-\varepsilon}) = \Omega(d \cdot n^\varepsilon / n)$. Using Lemma 4.5 (p. 35), the probability that at least one of the $\lambda = \mathrm{poly}(n)$ mutants in a step yields such a gain is upper bounded by $\lambda \cdot \mathrm{e}^{-\Omega(n^\varepsilon)}$, which is bounded by $\mathrm{e}^{-\Omega(n^\varepsilon)}$. □

The proof is appealingly simple. One feels, however, that the "true" lower bound on the number of steps should become smaller when λ is increased. Beyer (2001, p. 77) states that "an increase in the number of offspring of the $(1,\lambda)$ ES yields a logarithmic increase of the progress rate."[1] So the proof of our lower-bound result may be so simple because the bound is weak. To obtain a better lower bound, however, a more sophisticated reasoning than a simple application of the pigeonhole principle seems necessary.

As a starting point, one may ask with what probability a (1+1) ES might halve the approximation error in a single step. In other words, we aim at an *upper* bound on the success probability of an isotropic mutation to result in a spatial gain of at least $d/2$, where d denotes the distance from $\boldsymbol{x}^*$. More precisely, we want to bound $\mathsf{P}\{\Delta_{d,L} \geq d/2\}$ from above, where the length distribution L is arbitrary, i. e., we must again assume that the best length distribution was chosen. Clearly, there is one particular length ℓ^* of an isotropic mutation that results in the best chance of halving the approximation error. Therefore, recall Equation (4.2) on page 33 which tells us the correspondence between the distance from the hyperplane containing the parent (and lying perpendicular to the line passing through $\boldsymbol{c}$ and $\boldsymbol{x}^*$) and the spatial gain towards $\boldsymbol{x}^*$, where ℓ denotes the length of the isotropic mutation. For $g > 0$, the larger ℓ compared to g, the larger $\mathsf{P}\{G_\ell \geq g\}$, and thus, we need to minimize

$$\frac{g}{\ell} \;=\; \frac{\delta}{\ell} + \frac{\ell^2 - \delta^2}{\ell \cdot 2d} \;=\; \frac{\delta}{\ell} + \frac{\ell}{2d} - \frac{\delta^2}{\ell \cdot 2d}$$

(where we assume $\ell > 0$). As

$$\frac{\mathrm{d}}{\mathrm{d}\ell}\frac{g}{\ell} \;=\; \frac{\mathrm{d}}{\mathrm{d}\ell}\left(\frac{\delta}{\ell} + \frac{\ell}{2d} - \frac{\delta^2}{\ell\, 2d}\right) \;=\; \frac{-\delta}{\ell^2} + \frac{1}{2d} + \frac{\delta^2}{\ell^2 2d} \;=\; \frac{1}{2d} - \frac{\delta(2d-\delta)}{\ell^2 2d},$$

solving the equation $\frac{\mathrm{d}}{\mathrm{d}\ell}\, g/\ell = 0$ for ℓ yields that, for $0 < \delta < d$, the length

$$\ell^* \;:=\; \sqrt{\delta \cdot (2d - \delta)} \tag{4.4}$$

results in maximum success probability. Since $\delta < 2d - \delta$, we have $\ell^* \leq 2d - \delta$, and consequently,

$$\max_{\ell > 0} \mathsf{P}\{\Delta_{\boldsymbol{x}^*,\ell} \geq \delta\} \;=\; \max_{\ell > 0} \mathsf{P}\{G_\ell \geq g_{(d,\delta,\ell)}\} \;=\; \mathsf{P}\{G_{\ell^*} \geq g_{(d,\delta,\ell^*)}\}, \tag{4.5}$$

where $g_{(d,\delta,\ell)} = \delta + (\ell^2 - \delta^2)/(2d)$ and d is the distance from $\boldsymbol{x}^*$.

So, as we want to know the probability of halving d, we substitute $d/2$ for δ and obtain that in this case $\ell^* = \sqrt{\delta \cdot (2d - \delta)} = \sqrt{(d/2)(2d - d/2)} = d \cdot \sqrt{3/4}$ and $g_{(d,\delta = d/2,\ell^*)} = d \cdot 3/4$. Since

[1] where the progress rate "measures the expected change of the population with respect to a reference point in the parameter space from generation g to generation $g+1$", describing a "microscopic aspect of the local evolution" (Beyer, 2001, p. 17)

(for any fixed $\ell > 0$) $G_\ell \sim \ell \cdot G$, we have $\mathsf{P}\{G_{\ell^*} \geq g_{(d,\delta,\ell^*)}\} = \mathsf{P}\{G \geq g_{(d,\delta,\ell^*)}/\ell^*\}$, and hence, the probability that an isotropic mutation halves the approximation error equals $\mathsf{P}\left\{G \geq \sqrt{3/4}\right\}$ in the best case, i. e., when the mutation's length is chosen optimally.

Lemma 3.12 (p. 25) tells us that this probability is $\mathrm{e}^{-\Omega(n)}$. We obtain a more precise upper bound by recalling Equation (3.5) on page 23, which tells us that for $n \geq 4$

$$\mathsf{P}\left\{G \geq \sqrt{3/4}\right\} \;=\; \frac{1}{\Psi} \cdot \int_{\sqrt{3/4}}^{1} \left(1-x^2\right)^{(n-3)/2} \mathrm{d}x \;<\; (1-3/4)^{(n-3)/2}/\Psi \;=\; 2^{-n+3}/\Psi.$$

Thus (using the upper bound on $1/\Psi$ given in Inequality (3.6) on page 24) we have just proved

Lemma 4.10. Let $\boldsymbol{x}^* \in \mathbb{R}^n$ and $\boldsymbol{c} \in \mathbb{R}^n \setminus \{\boldsymbol{x}^*\}$ be fixed search points and let $\boldsymbol{m}$ be arbitrarily isotropically distributed over $\mathbb{R}^n$. Then, for $n \geq 4$, the probability $\mathsf{P}\{\mathrm{dist}(\boldsymbol{c}+\boldsymbol{m}, \boldsymbol{x}^*) \leq \mathrm{dist}(\boldsymbol{c}, \boldsymbol{x}^*)/2\}$ is bounded above by $2^{-n+3}/\Psi \;<\; 2^{-n+3} \cdot \sqrt{n-1}/\sqrt{2\pi} \;<\; 2^{-n} \cdot 3.2\sqrt{n}$.

So, what does this lemma tell us? Though it is no surprise that the chance of halving the approximation error with a single mutation drops when the dimensionality increases, we now know a concrete (exponentially small) upper bound on that probability. And indeed, this upper bound will enable us to also obtain an upper bound on the success probability within multiple steps of a (1+,λ) ES.

The idea behind this bound is the "curse of dimensionality" in $\mathbb{R}^n$. Therefore, firstly consider the search space $\{0,1\}^n$ and the standard mutation operator, which flips each of the n bits independently with probability $1/n$. When we repeatedly mutate a search point without doing selection, then each point in the search space is hit infinitely often as the number of mutations approaches infinity. In particular, the number of steps it takes this random search to visit a certain search point is finite. Now consider $\mathbb{R}^n$ for $n \geq 3$. Let us start with a fixed point and repeatedly add an isotropically distributed vector (with an arbitrary distribution of the length that is not concentrated on 0) to this point. Despite the fact that our starting point is never exactly hit again, even the probability of ever getting close again to our starting point tends to zero as the dimensionality increases, even if the number of mutations approaches infinity; cf. Grinstead and Snell (1997, Section 12.1).

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

In the following, we modify the (1+,λ) ES (with a global mutation strength as described in Section 1.2 (p. 8)) such that we end up with a search procedure that is independent of the function to be optimized and, thus, purely random: Consider the (1+,λ) ES after initialization, i. e., an initial starting point and an initial mutation strength are given. In the first step λ mutants are generated, each by adding an isotropic mutation (the distribution of which depends solely on the current σ) to the starting point. In contrast to the original (1+,λ) ES, we now do *not* select one of the λ(+1) individuals, yet keep all $1+\lambda$ search points as a population $P^{[1]}$. After the first step σ may be up- or down-scaled—depending on the individuals' function values. Thus, to also get rid of this function-dependency, each of the $1+\lambda$ points in $P^{[1]}$ is mutated 3 times: once without changing σ, once with an up-scaled σ, and once with a down-scaled mutation strength. Again we keep all $(1+\lambda) \cdot 3\lambda$ newly generated individuals (each of which consists of a search point and

the σ that was use to generate this search point). Consequently, we have $(1+\lambda)+(1+\lambda)\cdot 3\lambda = (1+\lambda)(1+3\lambda)$ individuals after the second step in the population $P^{[2]}$. Repeating this procedure, after i iterations a population $P^{[i]}$ is generated which contains

$$(1+\lambda)(1+3\lambda)^{i-1} \quad \leq \quad (1+3\lambda)^i \quad = \quad \mathrm{e}^{\ln(1+3\lambda)\cdot i}$$

individuals. The crucial point is that $P^{[i]}$ is built without any dependency on the function to be optimized, and that all search paths of the original $(1\overset{+}{,}\lambda)$ ES emerge in this modified search procedure with the same probability density. Let $S \subset \mathbb{R}^n$ denote an arbitrary Borel set. Then the probability that $P^{[i]}$ hits S, namely $\mathsf{P}\{S \cap P^{[i]} \neq \emptyset\}$, is an upper bound on the probability that the search point evolved within i iterations by the original $(1\overset{+}{,}\lambda)$ ES is in S. This is readily proved by induction on the number of steps; it is crucial that the initialization is done in the same way for both search procedures, of course.

Since each search point $\boldsymbol{x} \in P^{[i]}$ is generated by successively adding i isotropically distributed vectors to the initial search point, Lemma 3.7 (p. 18) tells us that $\boldsymbol{x}$ is indeed isotropically distributed w. r. t. the initial search point. We do not know the (distribution of the) distance between $\boldsymbol{x}$ and the initial search point, yet this does not matter—namely, we may assume the best case.

Now, if we choose the "target set" S as the hyper-ball containing all search points that have a distance of at most half the initial distance from $\boldsymbol{x}^*$, and if we know that the probability that an individual in $P^{[i]}$ hits S is very small, say, upper bounded by $2^{-n+3}/\Psi = \mathrm{e}^{-(\ln 2)(n-3)}/\Psi$ (which is at most $\mathrm{e}^{-0.692n}$ for n large enough since $\ln 2 > 0.693$), then the probability that $P^{[i]}$ contains at least one point from S is bounded above by

$$\#P^{[i]} \cdot \mathrm{e}^{-0.692n} \quad \leq \quad \mathrm{e}^{\ln(1+3\lambda)\cdot i} \cdot \mathrm{e}^{-0.692n} \quad = \quad \mathrm{e}^{\ln(1+3\lambda)\cdot i - 0.692n}$$

for n large enough (using the union bound). Then choosing $i := 0.69n/\ln(1+3\lambda)$ finally yields an upper bound of $\mathrm{e}^{-0.002n} = \mathrm{e}^{-\Omega(n)}$ on the probability that after $0.69n/\ln(1+3\lambda)$ steps the population contains an individual that lies in S. In other words, more than $0.69n/\ln(1+3\lambda)$ steps are necessary with probability $1-\mathrm{e}^{-\Omega(n)}$ to halve the approximation error. Since adding up a polynomial number of "error probabilities" each of which is $\mathrm{e}^{-\Omega(n)}$ results in a total error probability that is still $\mathrm{e}^{-\Omega(n)}$, we obtain the following lower-bound result:

Theorem 4.11. Let a $(1\overset{+}{,}\lambda)$ ES optimize an arbitrary function $f\colon \mathbb{R}^n \to \mathbb{R}$, and let $\boldsymbol{x}^* \in \mathbb{R}^n$ be some fixed point (for instance an optimum). Let $b : \mathbb{N} \to \mathbb{N}$ such that $b = \mathrm{poly}(n)$. Given that the initial search point has distance $d > 0$ from $\boldsymbol{x}^*$, with probability $1-\mathrm{e}^{-\Omega(n)}$ more than $b(n)\cdot 0.69n/\ln(1+3\lambda)$ steps (i. e. $\lambda \cdot b(n)\cdot 0.69n/\ln(1+3\lambda)$ f-evaluations) are necessary until (for the first time) the current search point has a distance of at most $d/2^{b(n)}$ from $\boldsymbol{x}^*$.

In particular, for the (1+1) ES we obtain that at least $0.69n/\ln 4 > 0.497n$ steps/f-evaluations are necessary with probability $1-\mathrm{e}^{-\Omega(n)}$ to halve the approximation error. Recall that we obtained $0.96n-1$ as a lower bound on the *expected* number of steps to halve the approximation error in Theorem 4.8 (p. 39).

So, what about the $(1,\lambda)\sigma$SA-ES, i. e. $(1,\lambda)$ ESs that use σ-self-adaptation instead of a global mutation strength, one might ask. In fact, the same reasoning applies: We drop selection and end up with a purely random search (since the way how σ is updated/mutated is independent of the function to be optimized). The population generated by this search procedure contains

$(1+\lambda)^i = e^{\ln(1+\lambda)\cdot i}$ individuals after i steps (rather than $(1+\lambda)(1+3\lambda)^{i-1} \leq e^{\ln(1+3\lambda)\cdot i}$), so that we obtain a slightly better lower bound:

Theorem 4.12. Let a $(1,\lambda)\sigma$SA-ES optimize an arbitrary function $f : \mathbb{R}^n \to \mathbb{R}$, and let $\boldsymbol{x}^* \in \mathbb{R}^n$ be some fixed point (for instance an optimum). Let $b : \mathbb{N} \to \mathbb{N}$ such that $b = \text{poly}(n)$. Given that the initial search point has distance $d > 0$ from $\boldsymbol{x}^*$, with probability $1 - e^{-\Omega(n)}$ more than $b(n) \cdot 0.69n/\ln(1+\lambda)$ steps (i. e. $\lambda \cdot b(n) \cdot 0.69n/\ln(1+\lambda)$ f-evaluations) are necessary until (for the first time) the current search point has a distance of at most $d/2^{b(n)}$ from $\boldsymbol{x}^*$.

4.4 Lower Bound for $(\mu+1)$ ESs which Holds with Overwhelming Probability

Recall the selection mechanism for reproduction in the $(\mu+1)$ ES: In each iteration of the evolution loop one of the μ individuals in the population is selected uniformly at random. Thus, if we pick one individual in advance (and disregard the other $\mu-1$ individuals), this one is actually selected with probability $1/\mu$. We assign to each individual, which is generated in a run of the $(\mu+1)$ ES, a unique number. Therefore, let the individuals in the initial population be numbered $-(\mu-1),\ldots,0$. The mutant that is generated in the first iteration of the evolution loop is numbered with "1" and so on.

Then each potential lineage of an individual of depth ℓ corresponds to a sequence $(i_0,\ldots,i_\ell) \in \mathbb{Z}^{\ell+1}$ such that $i_\ell > \cdots > i_0 \in \{-\mu+1,\ldots,0\}$. We will address the question with what probability a fixed such sequence emerges within the first k iterations of the evolution loop in a run of the $(\mu+1)$ ES. For $j \in \{1,\ldots,\ell\}$, the probability that the individual i_{j-1} is selected (for reproduction) in the i_jth step is either 0 or $1/\mu$, depending on whether this individual has already been removed from the population or not. Thus, $\mu^{-\ell}$ is an upper bound on the unconditional probability that the lineage corresponding to our fixed sequence emerges (we disregard that an individual may already have been deleted).

> Obviously, two such events, e. g., that the lineages respectively corresponding to the sequences "$0,1$" and "$-1,1$" emerge, are not independent (since the label "1" is assigned only once; in other words, the mutant generated in the first step cannot be a mutant of both, of individual "0" and of individual "-1").

Besides the μ choices for $i_0 \in \{-\mu+1,\ldots,0\}$, there are $\binom{k}{\ell}$ choices for $i_1,\ldots,i_\ell \in \{1,\ldots,k\}$, and thus, the number of sequences which cover all potential lineages of depth ℓ equals $\mu \cdot \binom{k}{\ell}$. Since the probability of a union of events is upper bounded by the sum of the probabilities of the single events (union bound), the probability that a lineage of depth ℓ emerges within the first k steps is upper bounded by

$$\mu \cdot \binom{k}{\ell} \cdot \mu^{-\ell} \quad \leq \quad \left(\frac{e \cdot k}{\ell}\right)^\ell \cdot \mu^{-\ell+1}.$$

This way of bounding the probability that a specific lineage emerges has already been proposed by Witt (2005a, Lemma 2).

Obviously, if no lineage of depth ℓ exists (after k steps), then the depth of each of the μ family trees (each of which is rooted at one of the μ initial individuals) is smaller than ℓ.

Theorem 4.13. Let a $(\mu+1)$ ES, where $\mu = \text{poly}(n)$, optimize an arbitrary function in $\mathbb{R}^n$. Let $\alpha : \mathbb{N} \to \mathbb{R}_{>0}$ such that $\alpha(n) \geq 1/n$, and let "α" abbreviate "$\alpha(n)$." Then the probability that after $\alpha\mu n$ iterations/mutations there is an individual in the population which has at least $\alpha 3n$ ancestors is upper bounded by $\mu \cdot 0.744^{n\cdot\alpha}$.

If $\alpha = \Omega(n^\varepsilon/n)$ for a constant $\varepsilon > 0$, then, for n large enough, $\alpha\, 3n$ is an upper bound on the expected depth of the forest after $\alpha\mu n$ steps.

Proof. Choosing $\ell := 3k/\mu$, the upper bound on the probability which we derived above becomes $\mu \cdot (\mathrm{e}/3)^{3k/\mu}$. When we choose $k := \alpha \cdot n \cdot \mu$ (implying that $\ell = 3k/\mu = \alpha 3n$), this upper bound becomes $\mu \cdot (\mathrm{e}/3)^{\alpha 3n}$. Finally, $(\mathrm{e}/3)^3 < 0.744$.

Substituting "2.9" for "3" in the preceding arguments yields that after $\alpha n\mu$ steps with a probability of less than $\mu \cdot 0.83^{\alpha n}$ the depth of the forest is at most $\alpha\, 2.9n$. Hence, the expected depth of the forest is upper bounded by $\alpha\, 2.9n + \alpha n\mu \cdot (\mu \cdot 0.83^{\alpha n})$, which is smaller than $\alpha\, 3n$ for n large enough when $\alpha \cdot n = \Omega(n^\varepsilon)$ and $\mu = \text{poly}(n)$ (because then $\mu^2 \cdot 0.83^{\alpha n} = \mu^2 \cdot \mathrm{e}^{-\Omega(n^\varepsilon)} \leq 0.1$ for n large enough). □

This theorem tells us that, if we want a lineage to emerge the depth of which is linear in the dimensionality of the search space, then w. o. p. $\Omega(\mu\, n)$ steps are necessary. Consequently, if we knew that a lineage of linear depth is necessary w. o. p. for a certain progress of the optimization, then w. o. p. $\Omega(\mu\, n)$ steps would be necessary to obtain such a progress.

Reconsider the (1+1) ES for a moment. As we have shown, it needs (even in the best case) more than $0.96(n-1)$ steps until the expected gain towards $\boldsymbol{x}^*$ is at least halve the initial distance from $\boldsymbol{x}^*$. As we have just seen, for the $(\mu+1)$ ES the number of steps until we expect a linage of length at least $0.96(n-1)$ to emerge is by a factor of at least $\mu/3$ larger. Thus, if the best-case progress along a linage of the $(\mu+1)$ ES was somehow "bounded" by the best-case progress in the (1+1) ES, we would obtain for the $(\mu+1)$ ES a lower bound of $(\mu/3) \cdot 0.96(n-1) = \mu\, 0.32(n-1)$ on the expected number of steps necessary to halve the approximation error.

Unfortunately, this first rough idea of a reasoning about how to show a lower bound cannot be extended to a formal proof. The selection mechanism for replacement raises dependencies between the events which correspond to the emergence of certain lineages. Namely, on the one hand, if a mutant makes it into the population, then there must be at least one individual in the population which is not better than the mutant. If, on the other hand, an individual $\mathcal{X}$ is eliminated from the population, this event tells us that the respective progress along the lineages of *all* other $\mu - 1$ individuals has been at least as good as the progress along the lineage of $\mathcal{X}$. These dependencies among the individuals in the population (and among their lineages) make an analysis very hard, possibly impractical. (In particular, we cannot multiply the expected depth with the expected best-case one-step progress to obtain an upper bound on the expected total progress.) Nevertheless, in particular the bound on the depth (of the lineages to emerge within a certain number of steps) which holds w. o. p. will later be useful in the analysis of the $(\mu+1)$ ES in a concrete scenario.

To obtain a general lower bound for the $(\mu+1)$ ES, however, and to get around this kind of dependencies, we may imagine that elimination in the $(\mu+1)$ ES was omitted (just as we did in the derivation of the general lower bound for $(1\dot{+}\lambda)$ ESs). As a consequence, the population grows in each iteration. Let $\mu^{[i]}$ denote the population's size *after* the ith iteration, so that $\mu^{[0]} = \mu$. Instead of generating one mutant per iteration, we now choose a set of $\lceil \mu^{[i-1]}/\mu \rceil$ individuals uniformly

at random in the ith iteration each of which is mutated. This ensures that, when having a look at a fixed individual in the population after a fixed number of steps (and disregarding all the other individuals), then this individual is selected for reproduction with a probability of at least $1/\mu$.

Besides the selection for elimination, there is another instruction within the evolution loop of the (μ+1) ES that raises dependencies when observing the decisions which are made within this instruction: the mutation adaptation. Whether the mutation strength is increased or decreased tells us something about the course of the optimization process so far. To get around this kind of dependencies we replace the mutation adaptation by the following procedure: In the ith iteration $3 \cdot \lceil \mu^{[i-1]}/\mu \rceil$ new individuals are generated; namely, $\lceil \mu^{[i-1]}/\mu \rceil$ new search points are generated (by mutating each of the randomly selected individuals once), yet each new search point bears three new individuals: one with the scaling factor decreased, one with the scaling factor increased, and one adopts the unmodified scaling factor of its parent.

Since the population grows (in each step i) by a factor that is at least $1+3/\mu$ but smaller than $1+6/\mu$, the population's size after $i \geq 1$ steps is bracketed by

$$\mu \cdot (1+3/\mu)^i \;\leq\; \mu^{[i]} \;<\; \mu \cdot (1+6/\mu)^i \;\leq\; \mu \cdot \mathrm{e}^{6i/\mu}.$$

All in all, our modifications to the (μ+1) ES lead to the following search procedure which we may call "(μ+1) Random Search" ("(μ+1) RS"), where the g- and b-counters are useless and, hence, omitted: For a given initialization of the population of μ individuals, the (μ+1) RS performs the following loop:

1. Choose $k := \lceil \text{current population size}/\mu \rceil$ of the individuals in the current population uniformly at random (without replacement). Let those be $\mathcal{X}_1, \ldots, \mathcal{X}_k$.
2. For each $(\boldsymbol{x}, \sigma) \in \{\mathcal{X}_1, \ldots, \mathcal{X}_k\}$ do
   a) create a new search point $\boldsymbol{y} := \boldsymbol{x} + \boldsymbol{m} \in \mathbb{R}^n$ with an isotropic mutation vector $\boldsymbol{m}$ (the distribution of which depends solely on σ);
   b) add the individuals $(\boldsymbol{y}, \sigma), (\boldsymbol{y}, 2\sigma), (\boldsymbol{y}, \sigma/2)$ to the population.
3. GOTO 1.

Obviously, this algorithm does not take the function to be optimized into account, yet performs some kind of "non-guided" random search. Nevertheless, it will be useful in the analysis of the (μ+1) ES. Namely, for any Borel set $S \subset \mathbb{R}^n$, the probability that the (μ+1) ES hits S (i. e., at least one individual from the population lies in S) within i steps is upper bounded by the probability that after i iterations the population of the (μ+1) RS contains an individual in S. This is again readily proved by induction on the number of steps, and again it is crucial that for both search procedures the population is initialized in the same way.

Hence, if this "hitting-probability" of the (μ+1) RS is bad, namely exponentially small, after i iterations, then the (μ+1) ES needs at least i iterations w. o. p. The main advantage, however, is the following: Since the random search of the (μ+1) RS is unbiased, each lineage corresponds to an "independent-mutation sequence" (this notion was coined by Witt (2005a)), i. e., each member in the population has evolved from some individual $\mathcal{X} = (\boldsymbol{x}, \sigma)$ in the initial population by adding

independently isotropically distributed vectors to $\boldsymbol{x}$. Thus, each search point in the population is isotropically distributed around the initial individual from which it descends. This enables us to prove the following lower-bound result:

Theorem 4.14. Let a $(\mu{+}1)$ ES, $\mu = \text{poly}(n)$, optimize an arbitrary function $f\colon \mathbb{R}^n \to \mathbb{R}$ using isotropic mutations, and let $\boldsymbol{x}^* \in \mathbb{R}^n$ be a fixed point (for instance an optimum). Let $b\colon \mathbb{N} \to \mathbb{N}$ such that $b = \text{poly}(n)$. Given that each initial search point has distance $d > 0$ from $\boldsymbol{x}^*$, with probability $1 - \mathrm{e}^{-\Omega(n)}$ more than $b(n) \cdot \mu \cdot 0.115n$ steps/f-evaluations are necessary until (for the first time) there is a search point in the population that has a distance of at most $d/2^{b(n)}$ from $\boldsymbol{x}^*$.

Proof. We firstly concentrate on halving the approximation error. Therefore, recall Lemma 4.10 (p. 41) and let S again denote the hyper-ball containing all search points with a distance of at most $d/2$ from $\boldsymbol{x}^*$. Since after i steps there are less than $\mu \cdot \mathrm{e}^{6i/\mu}$ individuals in the population that is generated by the $(\mu+1)$ RS, and since each of the search points is isotropically distributed around one of the μ initial search points (each of which has a distance of at least d from $\boldsymbol{x}^*$), the probability that this population hits S is smaller than

$$\mu \cdot \mathrm{e}^{6i/\mu} \cdot 2^{-n+3}/\Psi \;=\; \mathrm{e}^{6i/\mu - n\cdot\ln 2} \cdot 8\mu/\Psi \;=\; \mathrm{e}^{6i/\mu - n\cdot\ln 2} \cdot O(\mu\sqrt{n}). \tag{4.6}$$

Since $\ln 2 > 0.693$, choosing $i := \mu \cdot 0.115n$ results in an upper bound of $\mathrm{e}^{-0.003n} \cdot O(\mu\sqrt{n}) = \mathrm{e}^{-\Omega(n)}$ on the probability that after i steps the population contains a search point that lies in S. Hence, with probability $1 - \mathrm{e}^{-\Omega(n)}$ more than $\mu \cdot 0.115n$ steps are necessary for the population to halve the approximation error.

Finally, concerning halving the approximation error b times, summing up $b = \text{poly}(n)$ error probabilities each of which is $\mathrm{e}^{-\Omega(n)}$ results in a probability of $\mathrm{e}^{-\Omega(n)}$ that at least one of b halvings is accomplished within at most $\mu \cdot 0.115n$ steps. □

In particular, for the "$(\mu{+}1)$ ES with $\mu := 1$" this bound becomes $0.115n$ for the number of steps that are necessary w. o. p. to halve the approximation error, and since we dropped selection, this bound also holds for the (1+1) ES, i. e. the "$(1{+}\lambda)$ ES with $\lambda := 1$." This lower bound is worse than the bound of $0.497n$ implied for the (1+1) ES by the lower-bound result for the $(1^+,\lambda)$ ES in the previous section (namely Theorem 4.11 (p. 42)), though.

However, it has not been our aim to obtain a good bound for $\mu = 1$. We are interested in how the lower bound scales with the population size μ, that is the point. And we see that in the best case w. r. t. the minimization of the approximation error in the search space, the number of steps does indeed grow linearly in the population size μ for the $(\mu{+}1)$ ES.

The lower bound tells us that w. o. p. at least $\mu \cdot 0.115n$ steps are necessary to halve the approximation error. Yet what about the number of steps that are necessary to reduce the approximation error by, say, 1%? Therefore, recall Equation (4.6) on page 46 in the proof of the lower bound, and in particular the term "$2^{-n+3}/\Psi$." This is an upper bound on the best-case probability to halve the approximation error with an isotropic mutation. Now, Lemma 4.5 (p. 35) tells us that the probability that an isotropic mutation (in particular, in the best case) reduces the approximation error by $0.01d$ is bounded above by $\mathrm{e}^{-\Omega(n)}$. Assume, this probability is at most $\mathrm{e}^{-\varepsilon n}$ for n large enough. Then we can modify Equation (4.6) on page 46 and obtain an upper bound of

$$\mu \cdot \mathrm{e}^{6i/\mu} \cdot \mathrm{e}^{-\varepsilon n} \;=\; \mu \cdot \mathrm{e}^{6i/\mu - \varepsilon n}$$

on the probability that there is at least one individual in the population $P^{[i]}$ whose distance from $\boldsymbol{x}^*$ is at most $0.99d$. Choosing $i := \mu \cdot n \cdot \varepsilon/7$, this upper bound becomes $\mu \cdot \mathrm{e}^{-n\cdot\varepsilon/7} = \mathrm{e}^{-\Omega(n)}$. As all arguments hold not only for the reduction of the approximation error by 1%, but for any positive constant fraction, we obtain the following:

Corollary 4.15. Let a $(\mu$+1) ES, $\mu = \mathrm{poly}(n)$, optimize an arbitrary function $f : \mathbb{R}^n \to \mathbb{R}$ using isotropic mutations, and let $\boldsymbol{x}^* \in \mathbb{R}^n$ be some fixed point (for instance an optimum). Assume that each initial search point has distance $d > 0$ from $\boldsymbol{x}^*$. Then, for any constant $\varepsilon > 0$, with probability $1 - \mathrm{e}^{-\Omega(n)}$ the $(\mu$+1) ES needs $\Omega(\mu n)$ steps/f-evaluations until (for the first time) there is a search point in the population that has a distance of at most $(1-\varepsilon)\cdot d$ from $\boldsymbol{x}^*$.

4.5 Overcoming Gaps with Elitist Selection

We (re)consider $(1+\lambda)$ ESs in this section, and the crucial aspect to keep in mind is the following: When elitist selection is used (as in the $(1+\lambda)$ ES), then a mutant must be at least as good as its parent (w. r. t. to the function value) to have a chance to become selected. In other words, mutants with a worse function value are always discarded.

To get an idea of the problem which we want to deal with, consider the finite search space $\{0,1\}^n$ for a moment. One of the first functions that have been considered in a theoretical runtime analysis is $\textsc{Jump}_m : \{0,1\}^n \to \mathbb{N}$ with $m : \mathbb{N} \to \mathbb{N}$ such that $2 \le m(n) \le n/3$, defined by

$$\textsc{Jump}_m(\boldsymbol{x}) := \begin{cases} 2n & \text{if } 1 \le \textsc{Sum}(\boldsymbol{x}) \le m-1, \\ \textsc{Sum}(\boldsymbol{x}) & \text{otherwise,} \end{cases}$$

which is to be minimized (note that $\textsc{Sum}(\boldsymbol{x})$ equals the number of 1-bits in $\boldsymbol{x}$). We call the plateau of worst Jump-value $2n$ "the gap" as it separates the global minimum, namely the origin (the all-zero string), from the L^1-norm based part of the fitness landscape; all bit-strings with exactly m ones are locally but not globally optimal. Since the (1+1) EA chooses the initial search point $\boldsymbol{x}$ uniformly at random, $\mathsf{E}[\textsc{Sum}(\boldsymbol{x})] = n/2$ and, by Chernoff's bound, $\mathsf{P}\{|\boldsymbol{x}| \le n/3\} = \mathrm{e}^{-\Omega(n)}$. Consequently, the initial search point is located in the gap only with an exponentially small probability; the probability that the initial search point is the optimum equals 2^{-n}.

Droste, Jansen, and Wegener (2002b) prove that the expected runtime of the (1+1) EA using the static mutation probability $p = 1/n$ on $\textsc{Jump}_m$ is $\Theta(n^m)$ (in fact, a slightly different function which is to be maximized is investigated, yet the proof carries over). Roughly speaking, the (1+1) EA minimizes $\textsc{Jump}_m$ as it minimizes the L^1-norm up to the point when a locally but not globally optimal point with Hamming distance m from the origin is created. Then a mutation must exactly flip the remaining m ones for the (1+1) EA to overcome the gap, i. e., to obtain a search point with smaller Jump-value (namely the global minimum). The probability of this event (called "success" in the following) equals $p^m(1-p)^{n-m}$, where p denotes the mutation probability (recall that a mutation consists in flipping each bit independently with probability p). Since $\frac{\mathrm{d}}{\mathrm{d}p}\, p^m(1-p)^{n-m} = 0$ for $p = m/n$, the success probability is maximum when using the mutation probability $p = m/n$, and hence, even if the (1+1) EA could adapt p optimally, the success probability is upper bounded by $(m/n)^m(1-m/n)^{n-m}$. Since the number of trials until a mutation actually creates a better point is geometrically distributed, the expected runtime of the (1+1) EA

on $\textsc{Jump}_m$ is lower bounded by the reciprocal of the success probability. Thus, we expect a super-polynomial number of steps if $(n/m)^m$ is super-polynomial or if $(1-m/n)^{n-m}$ is super-polynomially small. For $m \in [n^\varepsilon, n/3]$ with $\varepsilon \in (0,1)$, we have $(1-m/n)^{n-m} \leq (1-n^\varepsilon/n)^{n\cdot 2/3} < \mathrm{e}^{-n^\varepsilon 2/3}$, and hence, the success probability is exponentially small, so that the expected runtime is exponential. For $m \in [\log n, n^\varepsilon]$ with $\varepsilon \in (0,1)$, we have $(n/m)^m \geq (n^{1-\varepsilon})^{\log n} = n^{(1-\varepsilon)\cdot\log n}$, and thus, the expected runtime is super-polynomial. Finally, we consider the case $m \leq \log n$. Then $(n/m)^m = 2^{m(\log n - \log m)} \geq 2^{m(\log n - \log\log n)} = 2^{m\cdot\Omega(\log n)} = n^{m\cdot\Omega(1)}$, and hence, the expected runtime is super-polynomial unless $m = O(1)$.

All in all, the expected runtime of the (1+1) EA on $\textsc{Jump}_m$ is polynomial (in n) if $m = O(1)$ when using the standard mutation probability $1/n$, and—as we we have just shown—it is super-polynomial if m is not $O(1)$ *even when the mutation probability could be adapted optimally*, i. e., our lower bound applies also, for instance, to the dynamic (1+1) EA introduced by Droste, Jansen, and Wegener (2001), which varies the mutation probability according to a static periodic schedule. Moreover, this remains true when considering arbitrary isotropic binary mutations (cf. the discussion on page 29): In the best case, a uniformly chosen subset of m bits would be flipped, resulting in a success probability of $1/\binom{n}{m}$. And, since $m \leq n/3$, we have $\binom{n}{m} = \mathrm{poly}(n)$ only if $m = O(1)$. In other words, an efficient optimization, i. e. a polynomial (expected) runtime, is possible only for a gap corresponding to a constant number of specific bits which have to be flipped simultaneously by a single mutation.

The aim of this section is to prove a similar result for minimization in the search space $\mathbb{R}^n$ when using "isotropic-mutation hill-climbing", i. e., when applying (1+λ) ESs that use isotropic mutations.

4.5.1 Linearly Separated Gaps

Consider a search point $\boldsymbol{c} \in \mathbb{R}^n$ and its lower-level set $A_{<\boldsymbol{c}} := \{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) < f(\boldsymbol{c})\}$ for a given function f. Assume that the set $A_{<\boldsymbol{c}}$ is bounded (finite diameter) and that it has a positive Lebesgue measure (a positive n-volume). Then we say that $\boldsymbol{c}$ faces a *linearly separated gap* in the search space if there is a hyper-plane $H_{\boldsymbol{c}}$ containing $\boldsymbol{c}$ such that $A_{<\boldsymbol{c}}$ lies completely in one of the two half-spaces w. r. t. $H_{\boldsymbol{c}}$. Then $\mathrm{dist}(H_{\boldsymbol{c}}, A_{<\boldsymbol{c}}) = \inf\{\mathrm{dist}(\boldsymbol{x},\boldsymbol{y}) \mid \boldsymbol{x} \in H_{\boldsymbol{c}}, \boldsymbol{y} \in A_{<\boldsymbol{c}}\}$ is the (absolute) size of the gap and we assume that the hyper-plane $H_{\boldsymbol{c}}$ is oriented such that this gap is as large as possible. Let $r := \sup\{\mathrm{dist}(\boldsymbol{c},\boldsymbol{x}) \mid \boldsymbol{x} \in A_{<\boldsymbol{c}}\}$. We define the relative size of the linearly separated gap as $\mathrm{dist}(H_{\boldsymbol{c}}, A_{<\boldsymbol{c}})/r$ for $r > 0$, and otherwise, the gap's relative size is zero.

So, assume that a (1+λ) ES minimizes some function f in $\mathbb{R}^n$ and that the evolving search point $\boldsymbol{c}$ does face a linearly separated gap of relative size $s > 0$. If f is such that $\boldsymbol{c}$'s level-set $A_{=\boldsymbol{c}} := \{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) = f(\boldsymbol{c})\}$ has zero Lebesgue measure (or such that any point in $A_{=\boldsymbol{c}}$ faces a linearly separated gap of relative size at least s), the only chance to overcome the gap, i. e. to leave $\boldsymbol{c}$ (resp. $A_{=\boldsymbol{c}}$), is to generate a mutant in $A_{<\boldsymbol{c}}$. Depending on the gap's relative size, we can now ask for an upper bound on the success probability of an isotropic mutation, i. e. on the probability that the mutant $\boldsymbol{c}+\boldsymbol{m}$ lies in $A_{<\boldsymbol{c}}$ (which is the mass of $A_{<\boldsymbol{c}}$ w. r. t. to the measure induced by the distribution of the mutation vector $\boldsymbol{m}$). However, depending on the shape of $A_{<\boldsymbol{c}}$ and/or the distribution of $\boldsymbol{m}$ this might actually be intractable, and thus, we are going to make best-case assumptions:

1. Consider the hyper-ball B centered at $\boldsymbol{c}$ with radius r (cf. above) which is cut in half by the hyperplane $H_{\boldsymbol{c}}$. One of the two parts contains $A_{<\boldsymbol{c}}$ completely, let this part be denoted by M, i. e., $B \supset M \supseteq A_{<\boldsymbol{c}}$. Let $C := \{\boldsymbol{x} \in M \mid \operatorname{dist}(\boldsymbol{x}, H_{\boldsymbol{c}}) = \operatorname{dist}(H_{\boldsymbol{c}}, A_{<\boldsymbol{c}})\}$, implying that $A_{<\boldsymbol{c}} \subseteq C \subseteq M \subset B$. ($C \neq M$ iff the gap's absolute size is non-zero.) The set C is a solid cap of the ball B.
   Then we assume that hitting $C \supseteq A_{<\boldsymbol{c}}$ is a success, in other words, we assume the best case that the "success region" is "as large as possible" for the given relative gap size.

2. We assume that the distribution of the isotropic mutation is such that the probability of hitting $C \supseteq A_{<\boldsymbol{c}}$ is maximum.

Assume that this "hitting probability" is $p_{\text{best}} > 0$ (under these best-case assumptions). Then again assuming the best case that the (1+λ) ES repeats doing best-case mutations over and over again, the number of trials necessary to get away from $\boldsymbol{c}$ (namely to generate a mutant that lies in $C \supseteq A_{<\boldsymbol{c}}$) is geometrically distributed. Consequently, the expected number of trials to leave $\boldsymbol{c}$ equals $1/p_{\text{best}}$ in the very best case, so that the expected number of isotropic mutations performed by an (1+λ) ES is lower bounded by $\lambda \cdot \lceil (1/p_{\text{best}})/\lambda \rceil$, which is at least $1/p_{\text{best}}$ and considerably larger than $1/p_{\text{best}}$ only if λ is considerably larger than $1/p_{\text{best}}$. Thus, we could add another best-case assumption; namely, we may concentrate on (1+1) ESs.

Consider an isotropic mutation with a fixed length of $\ell \in (0, r]$, i. e., for the isotropic mutation $\boldsymbol{m}$ we have $\mathsf{P}\{|\boldsymbol{m}| = \ell\} = 1$. Then the probability of hitting C equals

$$\mathsf{P}\{\boldsymbol{c} + \boldsymbol{m} \in C \mid |\boldsymbol{m}| = \ell\} \quad = \quad \mathsf{P}\{G_\ell \geq \operatorname{dist}(H_{\boldsymbol{c}}, A_{<\boldsymbol{c}})\} \quad = \quad \mathsf{P}\{G \geq \operatorname{dist}(H_{\boldsymbol{c}}, A_{<\boldsymbol{c}})/\ell\}$$

(recall the definition of the random variable G in Equation (3.2) on page 21). Thus, the larger ℓ, the larger the hitting probability, and hence we assume that the length of the isotropic mutation is concentrated on r (the best case; cf. above). Recall that the relative gap size equals $s = \operatorname{dist}(H_{\boldsymbol{c}}, A_{<\boldsymbol{c}})/r$. Using Equation (3.5) on page 23, we obtain a best-case hitting-probability of

$$\mathsf{P}\{\boldsymbol{c} + \boldsymbol{m} \in C \mid |\boldsymbol{m}| = r\} \quad = \quad \mathsf{P}\{G \geq s\} \quad = \quad \frac{1}{\Psi} \int_s^1 \left(1 - x^2\right)^{(n-3)/2} \mathrm{d}x.$$

Since $(1-x^2)^{(n-3)/2}$ is decreasing (in x for $0 < x < 1$), the integral's value is in fact bounded from above by $(1-s^2)^{(n-3)/2}/\Psi$ and it is super-polynomially small if s^2 is not $O(\log n/n)$ because $(1-t/k)^k \leq \mathrm{e}^{-t}$ for $0 \leq t \leq k \geq 1$ (and $1/\Psi = \Theta(\sqrt{n})$; cf. Inequality (3.6) on page 24).

On the other hand, for any $a \in (0, 1/2)$,

$$\int_a^{2a} \left(1 - x^2\right)^{(n-3)/2} \mathrm{d}x \quad \geq \quad a \cdot \left(1 - (2a)^2\right)^{(n-3)/2},$$

and hence, $\int_s^1 (1-x^2)^{(n-3)/2} \mathrm{d}x$ is bounded also from below by a polynomial (of negative degree) for $s^2 = O(\log n/n)$. (Note that the (negative) degree of the polynomial depends on the disguised constant in the O-notation.) In shorter words, we have proved

$$1/\mathsf{P}\{G \geq s\} = \operatorname{poly}(n) \quad \Longleftrightarrow \quad s^2 = O(\log n/n).$$

All in all, we obtain

Theorem 4.16. Let a $(1+\lambda)$ ES, $\lambda = \text{poly}(n)$, optimize some function $f : \mathbb{R}^n \to \mathbb{R}$ using isotropic mutations. Assume that the current search point $\boldsymbol{c}$ faces a linearly separated gap of relative size s and that f is such that $\boldsymbol{c}$'s level set $\{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) = f(\boldsymbol{c})\}$ has zero Lebesgue measure or that any $\boldsymbol{x}$ in $\boldsymbol{c}$'s level set faces a linearly separated gap of relative size at least s. Then, independently of the mutation adaptation, the expected number of mutations until a better (w. r. t. the f-value) search point is generated cannot be polynomial in n unless $s = O(\sqrt{\log n/n})$.

If in this situation $s = \Omega(n^\varepsilon/\sqrt{n})$ for some positive constant ε, then, in expectation as well as with probability $1 - \exp(-\Omega(n^{2\varepsilon}))$, the number of mutations which are necessary to generate a better search point is $\exp(\Omega(n^{2\varepsilon}))$.

Proof. That the expected number of steps cannot be polynomial unless $s^2 = O(\log n/n)$ has just been shown in the reasoning preceding the theorem.

For the proof of the second claim, let $s = \Omega(n^\varepsilon/\sqrt{n})$, so that

$$(1-\Omega(n^{2\varepsilon})/n)^{(n-3)/2}/\Psi \quad \leq \quad \exp\left(-\frac{(n-3)\cdot\Omega(n^{2\varepsilon})}{2 \quad \cdot \quad n}\right)\cdot O(\sqrt{n}) \quad = \quad \exp(-\Omega(n^{2\varepsilon}))$$

is an upper bound on the best-case hitting-probability; assume that $\alpha\colon \mathbb{N} \to \mathbb{R}$ is such that $\exp(-\alpha(n)\cdot n^{2\varepsilon})$ is this upper bound, i. e., $\alpha = \Omega(1)$. Then the probability of having at least one hit in $\exp(\alpha(n)\cdot n^{2\varepsilon}/2) = \exp(\Omega(n^{2\varepsilon}))$ trials/mutations is upper bounded by $\exp(-\alpha(n)\cdot n^{2\varepsilon}/2) = \exp(-\Omega(n^{2\varepsilon}))$ (using the union bound). □

Note that, since $\lambda = \text{poly}(n)$, this theorem remains valid when substituting "number of steps" for "number of mutations," which makes sense when all λ mutations in a step can be performed in parallel.

4.5.2 Spherically Separated Gaps

Consider again a search point $\boldsymbol{c} \in \mathbb{R}^n$ and its lower-level set $A_{<\boldsymbol{c}} := \{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) < f(\boldsymbol{c})\}$, and assume again that the set $A_{<\boldsymbol{c}}$ is bounded (finite diameter) and that it has a positive Lebesgue measure. Then there is a hyper-ball $B_{<\boldsymbol{c}} \supseteq A_{<\boldsymbol{c}}$ of smallest size (i. e. with smallest radius), and we say that the search point $\boldsymbol{c} \in \mathbb{R}^n$ faces a *spherically separated gap* in the search space of absolute size $\text{dist}(\boldsymbol{c}, B_{<\boldsymbol{c}})$ and relative size $\text{dist}(\boldsymbol{c}, B_{<\boldsymbol{c}})/\text{dist}(\boldsymbol{c}, \text{center of } B_{<\boldsymbol{c}})$ (if defined).

So, we assume that a $(1+\lambda)$ ES minimizes some function f and that the evolving search point $\boldsymbol{c}$ faces a spherically separated gap of relative size $s > 0$. If f is such that the level-set $A_{=\boldsymbol{c}}$ has zero Lebesgue measure (or such that any point in $A_{=\boldsymbol{c}}$ faces a spherically separated gap of relative size at least s), the only chance to overcome the gap, i. e. to leave $\boldsymbol{c}$ (resp. $A_{=\boldsymbol{c}}$), is to generate a mutant in $A_{<\boldsymbol{c}}$. Again we make best-case assumptions:

1. We assume that hitting the hyper-ball $B_{<\boldsymbol{c}} \supseteq A_{<\boldsymbol{c}}$ is a success and that, in addition,
2. the distribution of the isotropic mutation is such that the probability of hitting $B_{\boldsymbol{c}}$ is maximum.

Assume that this hitting probability is $p_{\text{best}} > 0$ under the best-case assumptions. Then, again, the expected number of trials to leave $\boldsymbol{c}$ (resp. $A_{=\boldsymbol{c}}$) equals $1/p_{\text{best}}$ in the very best case.

Recall that we have already tackled the question of the best-case probability to overcome a spherically separated gap of relative size 0.5. Namely, Lemma 4.10 (p. 41) tells us (by letting $\boldsymbol{x}^*$ denote the center of the hyper-ball $B_{<\boldsymbol{c}}$) that for $s = 0.5$ the probability of hitting $B_{<\boldsymbol{c}}$—namely of halving the distance from the center of $B_{<\boldsymbol{c}}$—is bounded above by $2^{-n} \cdot 3.2\sqrt{n}$ for any isotropic mutation when $n \geq 4$. Thus, in our scenario the expected number of mutations to overcome the spherically separated gap of relative size 0.5 is bounded below by $2^{n-O(\log n)}$. The reasoning that has led to the previously mentioned lemma can also be used to upper bound the hitting probability for other gap sizes. Therefore, reconsider Figure 4.1 (p. 32): $\boldsymbol{x}^*$ can be considered the center of the ball $B_{<\boldsymbol{c}}$ and δ the absolute size of the spherically separated gap which $\boldsymbol{c}$ faces. Then Equation (4.4) on page 40 tells us the length ℓ^* which makes an isotropic mutation hit $B_{<\boldsymbol{c}}$ with the largest possible probability.

Namely, the optimal length of an isotropic mutation (under the best-case assumptions) equals $\sqrt{\delta \cdot (2d - \delta)}$, where here δ denotes the absolute size of the spherically separated gap and d the distance between $\boldsymbol{c}$ and $\boldsymbol{x}^*$ (here the center of $B_{<\boldsymbol{c}}$). Moreover, Equation (4.5) on page 40 tells us that the best-case hitting probability in this case equals $\mathsf{P}\{G_{\ell^*} \geq g_{(d,\delta,\ell^*)}\} = \mathsf{P}\{G \geq g_{(d,\delta,\ell^*)}/\ell^*\}$, where $g_{(d,\delta,\ell^*)} = \delta + ((\ell^*)^2 - \delta^2)/(2d)$. Since

$$\frac{g_{(d,\delta,\ell^*)}}{\ell^*} \;=\; \frac{\delta + \frac{(\ell^*)^2 - \delta^2}{2d}}{\ell^*} \;=\; \frac{\delta + \frac{\delta \cdot (2d-\delta) - \delta^2}{2d}}{\sqrt{\delta \cdot (2d-\delta)}} \;=\; \frac{\delta \cdot (2 - \delta/d)}{\sqrt{\delta \cdot (2d-\delta)}} \;=\; \frac{\sqrt{\delta \cdot (2d-\delta)}}{d} \;=\; \frac{\ell^*}{d},$$

the best-case probability of hitting $B_{<\boldsymbol{c}}$ equals (for $n \geq 4$)

$$\mathsf{P}\{G_{\ell^*} \geq g_{(d,\delta,\ell^*)}\} \;=\; \mathsf{P}\big\{G \geq g_{(d,\delta,\ell^*)}/\ell^*\big\} \;=\; \mathsf{P}\big\{G \geq \ell^*/d\big\} \;=\; \frac{1}{\Psi} \int_{\ell^*/d}^{1} \big(1 - x^2\big)^{(n-3)/2} \,\mathrm{d}x,$$

where the last equality is due to Equation (3.5) on page 23. (Note that this best-case probability is an upper bound on the probability of hitting $A_{<\boldsymbol{c}}$ for *any* isotropic mutation.)

Since $\ell^* = \sqrt{\delta \cdot (2d - \delta)}$ and $0 \leq \delta \leq d$, we have $\ell^*/d = \sqrt{\xi \cdot \delta/d}$ for some function ξ (of δ) with range $[1, 2]$. As the relative size of the spherically separated gap is $s = \delta/d$, we obtain

$$\mathsf{P}\big\{\boldsymbol{c} + \boldsymbol{m} \in B_{<\boldsymbol{c}} \mid |\boldsymbol{m}| = \ell^*\big\} \;=\; \frac{1}{\Psi} \int_{\sqrt{\xi \cdot s}}^{1} \big(1 - x^2\big)^{(n-3)/2} \,\mathrm{d}x$$

as the best-case probability of hitting $B_{<\boldsymbol{c}}$, i. e., when the isotropic distribution of $\boldsymbol{m}$ is such that $\mathsf{P}\{|\boldsymbol{m}| = \ell^*\} = 1$. Analogously to the reasoning/calculation for linearly separated gaps, we get

$$1 \big/ \mathsf{P}\big\{\boldsymbol{c} + \boldsymbol{m} \in B_{<\boldsymbol{c}} \mid |\boldsymbol{m}| = \ell^*\big\} = \mathrm{poly}(n) \quad \Longleftrightarrow \quad s = O(\log n / n),$$

where the degree of the polynomial depends on the concealed constant in the O-notation. All in all, we obtain

Theorem 4.17. Let a $(1+\lambda)$ ES, $\lambda = \mathrm{poly}(n)$, optimize some function $f : \mathbb{R}^n \to \mathbb{R}$ using isotropic mutations. Assume that the current search point $\boldsymbol{c}$ faces a spherically separated gap of relative size $s > 0$ and that f is such that $\boldsymbol{c}$'s level set $\{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) = f(\boldsymbol{c})\}$ has zero Lebesgue measure or that any $\boldsymbol{x}$ in $\boldsymbol{c}$'s level set faces a spherically separated gap of relative size at least s. Then, independently of the mutation adaptation, the expected number of mutations until a better (w. r. t. the f-value) search point is generated cannot be polynomial in n unless $s = O(\log n/n)$.

If in this situation $s = \Omega(n^\varepsilon/n)$ for some positive constant ε, then—in expectation as well as with probability $1 - \exp(-\Omega(n^\varepsilon))$—the number of mutations necessary to generate a better search point is $\exp(\Omega(n^\varepsilon))$.

Proof. The first claim has just been shown in the reasoning that precedes the theorem.

For the proof of the second claim, let $s = \Omega(n^\varepsilon/n)$, so that

$$(1-\Omega(n^\varepsilon)/n)^{(n-3)/2}/\Psi \quad \leq \quad \exp\left(-\frac{(n-3)\cdot\Omega(n^\varepsilon)}{2 \quad \cdot \quad n}\right)\cdot O(\sqrt{n}) \quad = \quad \exp(-\Omega(n^\varepsilon))$$

is an upper bound on the best-case hitting probability; assume that α (as a function of n) is such that $\exp(-\alpha\cdot n^\varepsilon)$ is this upper bound, implying that $\alpha = \Omega(1)$. Then the probability of having at least one hit in $\exp(\alpha\cdot n^\varepsilon/2) = \exp(\Omega(n^\varepsilon))$ trials/mutations is upper bounded by $\exp(-\alpha\cdot n^\varepsilon/2) = \exp(-\Omega(n^\varepsilon))$ (using the union bound). □

Recall that, since $\lambda = \text{poly}(n)$, also this theorem remains valid when substituting "number of steps" for "number of mutations."

4.5.3 Exemplary Application to Concrete Functions

To demonstrate how the lower-bound result on the (expected) number of steps necessary to overcome a spherically separated gap can be applied, two example functions which yield more insight will be introduced now. In the following, "gap" means "spherically separated gap." As mentioned in the introduction, we want to investigate functions for $\mathbb{R}^n$ that correspond to the function $\textsc{Jump}_m$ for $\{0,1\}^n$. Note that $\textsc{Jump}_m$ is symmetric (i. e., any two search points with the same number of 1-bits have the same function value). We will consider symmetric functions for $\mathbb{R}^n$—spherically symmetric, of course.

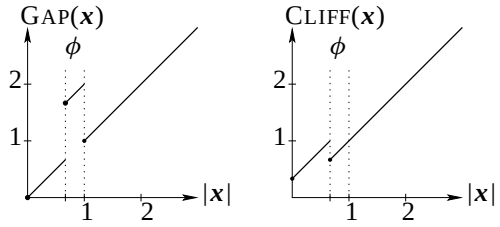


Figure 4.2: The functions GAP and CLIFF

Let $\phi\colon \mathbb{N} \to (0,1/3]$ denote a function (which determines the size of the gap). The sequence of functions $\textsc{Gap}_n^\phi\colon \mathbb{R}^n \to \mathbb{R}$, $n \in \mathbb{N}$, is defined by

$$\textsc{Gap}_n^\phi(\boldsymbol{x}) \quad := \quad \begin{cases} |\boldsymbol{x}|+1 & \text{for } |\boldsymbol{x}| \in [1-\phi(n),1) \text{ and} \\ |\boldsymbol{x}| & \text{otherwise.} \end{cases}$$

Due to ϕ's codomain, all $\boldsymbol{x}$ in the unit hyper-sphere $U = \{\boldsymbol{x} \mid |\boldsymbol{x}| = 1\}$ are locally but not globally optimal, and the origin is the unique global optimum. Note that only search points in U face a (spherically separated) gap of positive size, namely of size ϕ.

A similar class of functions is

$$\textsc{Cliff}_n^\phi(\boldsymbol{x}) := \begin{cases} |\boldsymbol{x}|+\phi(n) & \text{for } |\boldsymbol{x}| < 1-\phi(n), \\ |\boldsymbol{x}| & \text{otherwise.} \end{cases}$$

Also for CLIFF, only the local optima face a gap of positive size: A search point in the hyper-sphere $\{\boldsymbol{x} \in \mathbb{R}^n \mid |\boldsymbol{x}| = 1-\phi\}$ faces a gap of absolute size ϕ and relative size $\phi/(1-\phi)$.

So, for both functions the set of search points that face a gap (of positive size) forms a hyper-sphere and, thus, has zero Lebesgue measure. Hence, unless the initial search point is a local optimum, the evolving search point will (almost surely) never face a "spherically separated gap" —as formally defined above—since any isotropic mutation hits the hyper-sphere containing the local optima only with zero probability. It is intuitively clear, however, that the search faces some kind of gap. As we will see, a small change in our notion of when we consider a search point better than some other point will enable us to apply the lower-bound result which we obtained in Theorem 4.17 (p. 51).

Therefore, reconsider the set of points that are "better" than the current search point $\boldsymbol{c}$: We decided to consider a point $\boldsymbol{x}$ better than $\boldsymbol{c}$ iff $f(\boldsymbol{x}) < f(\boldsymbol{c})$ (for minimization), and hence, we considered the smallest ball $B_{<\boldsymbol{c}} \supseteq A_{<\boldsymbol{c}}$ containing the lower level set of $\boldsymbol{c}$ (w. r. t. f). Now, let $B^* := \{\boldsymbol{x} \mid |\boldsymbol{x}| < 1-\phi\}$ denote the open hyper-ball making up the "global optimum region" of GAP/CLIFF. Then we may consider a point $\boldsymbol{x}$ better than $\boldsymbol{c}$ iff it has a better function value *and* lies in the global-optimum region B^*. In other words, we redefine the size of the (spherically separated) gap based on the smallest ball containing $A_{<\boldsymbol{c}} \cap B^*$. Then, for GAP, any point in $R_{\text{GAP}} := \{\boldsymbol{x} \in \mathbb{R}^n \mid |\boldsymbol{x}| \geq 1\}$ faces a gap of absolute size at least ϕ, and for CLIFF, any point in $R_{\text{CLIFF}} := \{\boldsymbol{x} \in \mathbb{R}^n \mid |\boldsymbol{x}| \geq 1-\phi\}$ does so. Hence, for both functions the relative size of the gap that a search point from R faces is at least ϕ. Consequently, the best chance (under the best-case assumptions) to overcome the gap —namely to get from R into B^*—is at unit distance from the optimum/origin.

Unlike for CLIFF, for GAP we must deal separately with points $\boldsymbol{c} \in \{\boldsymbol{x} \in \mathbb{R}^n \mid |\boldsymbol{x}| \geq 2\}$: For such points, the lower-level set contains the set $M := \{\boldsymbol{x} \in \mathbb{R}^n \mid |\boldsymbol{x}| \in [1-\phi(n), 1)$ (the set of points that get the penalty of "+1"), and hence, a mutant (of such a $\boldsymbol{c}$) that hits M would get accepted by the elitist selection of a (1+λ) ES. However, since in such situations $\boldsymbol{c}$'s distance from M is at least 1 and $|\boldsymbol{c}| \geq 2$, such a mutation would have to overcome a spherically separated gap of absolute size 1 and relative size 1/2 (which is larger than the maximum ϕ-value of 1/3).

All in all, we have shown that Theorem 4.17 (p. 51) (almost) directly implies the following result:

Theorem 4.18. Let a (1+λ) ES, $\lambda = \text{poly}(n)$, optimize GAP^{ϕ} or CLIFF^{ϕ} using isotropic mutations. Assume that the initial search point lies in $R_{\text{GAP}} = \{\boldsymbol{x} \in \mathbb{R}^n \mid |\boldsymbol{x}| \geq 1\}$ resp. $R_{\text{CLIFF}} = \{\boldsymbol{x} \in \mathbb{R}^n \mid |\boldsymbol{x}| \geq 1-\phi\}$. Then, for any mutation adaptation, the expected number of mutations until the evolving search point enters the global-optimum region $B^* = \{\boldsymbol{x} \in \mathbb{R}^n \mid |\boldsymbol{x}| < 1-\phi\}$ (for the first time) cannot be polynomial in n unless $\phi = O(\log n/n)$. If $\phi = \Omega(n^{\varepsilon}/n)$ for some positive constant ε, then this number of mutations is $\exp(\Omega(n^{\varepsilon}))$—in expectation as well as with probability $1-\exp(-\Omega(n^{\varepsilon}))$.

4.5.4 Additional Notes on Overcoming Gaps

Naturally, we could easily define functions containing linearly separated gaps to demonstrate the applicability of the lower bound given in Theorem 4.16 (p. 50).

Due to the shape of the set of points that we consider better than the current search point $\boldsymbol{c}$, the size of a spherically and/or linearly separated gap which $\boldsymbol{c}$ faces might be zero in many cases when

intuition may say that $\boldsymbol{c}$ does face "some kind" of a gap. When considering isotropic mutations and the approximation error w. r. t. to the distance from a fixed point in the search space, however, the two notions of a gap that we have just considered seem somehow a natural starting point.

When $\phi = o(1)$, for instance $\phi(n) := 1/\sqrt{n}$ (so that the number of steps to overcome the cliff is exponential w. o. p.), then CLIFF_n^ϕ converges uniformly to the L^2-norm as $n \to \infty$. Since the smooth L^2-norm does not show any gaps, CLIFF can serve as a perfect example for how the assumption "in the limit of infinite dimensionality" can potentially lead to results that reveal only ill-founded conclusions for finite dimensional search space.

It is clear that the lower bounds do not only hold for (1+λ) ES as defined in Section 1.2 (p. 8), yet for any search procedure which fits the following framework (for minimization): For a given initialization of the evolving search point $\boldsymbol{c} \in \mathbb{R}^n$ the following loop is performed:

1. Depending on the complete history of the minimization so far, choose a $\lambda \in \mathbb{N}$.
2. FOR $i := 1$ TO λ DO create a new search point $\boldsymbol{y}_{[i]}$ by adding an isotropic mutation to $\boldsymbol{c}$, where the isotropic distribution of the mutation vector (in fact, the one of its length) may depend on the complete history of the optimization so far.
3. IF $S := \{\boldsymbol{y}_{[i]} \mid i \in \{1, \ldots, \lambda\},\ f(\boldsymbol{y}_i) \leq f(\boldsymbol{c})\}$ is not empty THEN decide, depending on the complete history of the optimization so far, whether a point from S replaces/becomes $\boldsymbol{c}$ and, if so, which one of them; update $\boldsymbol{c}$ accordingly.
4. IF stopping criterion met THEN output $\boldsymbol{c}$ ELSE GOTO 1.

Note that in each iteration a different λ can be chosen, and for each of the λ mutations, a different isotropic mutation may be used; respectively depending on the complete history of the search. The selection, however, is elitist, so that the sequence of function values which is induced by the evolving search point is monotonic.

4.6 Remarks on the Lower-Bound Results

As we have just seen in the preceding section, the lower bounds on the number of isotropic mutations which are necessary to overcome a (linearly/spherically separated) gap do not only hold for (1+λ) ESs that fit the framework given in Section 1.2 (p. 8), but for the generalized framework described at the end of the preceding section. Also the lower bound of $0.69n/\ln(1+3\lambda)$ on the number of steps a (1+λ) ES and/or a $(1, \lambda)$ ES necessarily needs (to halve the approximation error in the search space) is valid for a broader class of ESs/search heuristics. For instance, a "$(1 \circ \lambda)$ ES" using a "Metropolis-like" selection which accepts a worse mutant with a probability of, say, 5% would also be covered by the proof of Theorem 4.11 (p. 42). The reason for this is simple: In the modified search procedure, which is used in the analysis, all mutants that are ever generated survive and are kept in the (exponentially growing) population anyway. As a consequence, also a "simulated annealing-like" selection, where the probability of accepting a worse mutant depends on how worse the mutant is compared to its parent, would be covered.

We have to be careful, though: The modifications must be such that our modified search procedure remains independent of the function to be optimized. As we have just seen, this is no

problem for the selection mechanism. The mutation adaptation is more critical in this respect. In the proof of the lower bound we used that at the end of each step there are exactly three alternatives for the adaptation of the mutation strength σ, which may be called "increase", "keep", "decrease." We could allow more alternatives, though. If there were, say, seven alternatives for the σ-adaptation, the lower bound on the number of steps to halve the approximation error in the search space would become $0.69n/\ln(1+7\lambda)$, for instance.

Although our lower-bound results do not formally prove the following, they do strongly indicate that $(1\overset{+}{,}\lambda)$ ESs cannot achieve super-linear convergence, i. e. a convergence order of larger than one, when using isotropic mutations. This topic has recently been discussed by Teytaud and Gelly (2006) and by Teytaud, Gelly, and Mary (2006).

5 Bounds for Concrete Scenarios

So, now that we know some fairly general lower bounds on the number of steps (and mutations) which $(1^{+},\lambda)$ ESs and $(\mu+1)$ ESs need to reduce the approximation error in the search space (as long as they fit one of the frameworks given in Section 1.2 (p. 8), of course), the question arises whether a concrete ES optimizing a concrete function can achieve a runtime which asymptotically meets the lower bound, i. e., which is larger than the lower bound only by an $O(1)$-factor. It is clear that this is possible, if at all, only for very simple functions, and that this, obviously, depends on what kind of mutation adaptation is actually used.

We will consider Gaussian mutations since they are by far the most common type of isotropic mutations, and moreover, they have been used since the very first days of evolution strategies. Furthermore, we concentrate on the well-known 1/5-(success-)rule—mainly for two reasons: Firstly, it is the oldest adaptation mechanism; it was used in the very first (1+1) ES by Rechenberg and Schwefel (cf. Rechenberg (1973), Schwefel (1995)). Secondly, it is deterministic; namely, it does not introduce further randomness in the stochastic process induced by an ES. In particular, the mutation strength is not part of the evolution, but externally adapted. For this reason, it is sometimes referred to as an *exogenous* adaptation mechanism, whereas self-adaptive methods are sometimes called *endogenous.*

Usually, the 1/5-rule is used in the (1+1) ES only. Yet as we shall see, it does make sense—at least to some extent in the function scenarios to be considered—for the $(1^{+},\lambda)$ ES and also for the $(\mu+1)$ ES. Namely, for very simple functions, the 1/5-rule indeed ensures for the $(1+\lambda)$ ES and the $(\mu+1)$ ES a runtime which is of the same order as our lower bounds, and for the $(1,\lambda)$ ES, a runtime which is off by at most an $O(\sqrt{\ln\lambda})$-factor.

5.1 Gaussian Mutations and 1/5-Rule

Hereinafter, we call a mutation of a search point $\boldsymbol{c} \in \mathbb{R}^n$ with a mutation vector $\boldsymbol{m}$ which results in $f(\boldsymbol{c}+\boldsymbol{m}) \leq f(\boldsymbol{c})$ a *successful mutation,* and hence, when talking about a mutation, *success probability* means the probability that the mutant is at least as good as its parent. Based on experiments and rough calculations for two function scenarios (namely SPHERE and a corridor function), Rechenberg proposed the 1/5-rule for the adaptation of Gaussian mutations within the (1+1) ES. The idea behind this adaptation mechanism is that (in a step of the (1+1) ES) the mutation strength σ should be such that a scaled Gaussian mutation is successful with a probability of (roughly) $1/5$ since in such situations the expected gain of the step (mutation followed by selection) is maximum. Obviously, for the eletist (1+1) ES, the success probability of a step equals the probability that the mutation is accepted to become the new current search point in this step. If σ could be adapted such that every step was successful with probability $1/5$, we would observe that on average one fifth of the mutations are successful. Thus, the 1/5-rule works as follows: The

optimization process is observed without changing σ (we "keep" σ) until $5n$ mutations have been performed; if more than one fifth of the mutations in this observation period have been successful, σ is doubled ("increased"), otherwise, σ is halved ("decreased"). As a consequence, the 1/5-rule fits our $(1^+,\lambda)$ ES-framework from Section 1.2 (p. 8).

The number of mutations to be observed between two sequent σ-adaptations varies in the literature, but is almost always $\Theta(n)$. Also the choice of the constants for the adaptation of σ, here 2 resp. $1/2$, seems somehow arbitrary. In fact, one result we will obtain is that—for the function scenarios we consider—the order of the runtime (w. r. t. the dimensionality of the search space) is "robust" with respect to the concrete implementation of the 1/5-rule. Namely, any 1/5-rule that performs the σ-adaptation every $\Theta(n)$ mutations using any two constants for the scaling of σ that are greater resp. smaller than 1 results in the same asymptotic runtime; even the $1/5$ can be replaced by any positive constant smaller than $1/2$ without affecting the order of the runtime—in the function scenarios that are considered here.

5.1.1 Gaussian Mutations and 1/5-Rule for the $(1^+,\lambda)$ Evolution Strategy

The "$(1+\lambda)$ ES using scaled Gaussian mutations adapted by the 1/5-rule" works as follows: Let $\lambda\colon \mathbb{N} \to \mathbb{N}$ such that $\lambda = \text{poly}(n)$, and let "λ" abbreviate "$\lambda(n)$." We use two global counters: "g" corresponds to the number of "good" (i. e. successful) mutations, and "b" counts the "bad" ones (which have not been successful). Then, with $b := 0$ and $g := 0$ and a given initialization of the evolving search point $\boldsymbol{c} \in \mathbb{R}^n$ and the global mutation strength $\sigma \in \mathbb{R}_{>0}$, the following evolution loop is performed (the instructions that implement the 1/5-rule are marked gray):

1. FOR $i := 1$ TO λ DO BEGIN
   a) Create a new search point $\boldsymbol{y}_{[i]} := \boldsymbol{c} + \boldsymbol{m}$ with $\boldsymbol{m} := \sigma \cdot \widetilde{\boldsymbol{m}}$, where each component of $\widetilde{\boldsymbol{m}} \in \mathbb{R}^n$ is independently standard-normally distributed.
   b) IF $f(\boldsymbol{y}_{[i]}) \leq f(\boldsymbol{c})$ THEN $g := g + 1$ ELSE $b := b + 1$. END
2. IF $\min_{i\in\{1,\ldots,\lambda\}}\{f(\boldsymbol{y}_{[i]})\} \leq f(\boldsymbol{c})$ THEN $\boldsymbol{c} := \operatorname{argmin}_{i\in\{1,\ldots,\lambda\}}\{f(\boldsymbol{y}_{[i]})\}$ (when there are more than one mutant with minimum f-value, one of them is chosen uniformly at random).
3. IF $b + g \geq 5n$ THEN BEGIN
   a) IF $g < (g+b)\cdot(1/5)$ THEN $\sigma := \sigma/2$ ELSE $\sigma := \sigma \cdot 2$.
   b) $g := 0$. $b := 0$. END
4. GOTO 1.

Note that σ is adapted every $\lceil 5n/\lambda \rceil$ steps/iterations, implying that for $\lambda \geq 5n$ there is σ-adaptation after every iteration of the evolution loop.

As expected, we obtain the "$(1,\lambda)$ ES using Gaussian mutations adapted by the 1/5-rule" by dropping the IF-condition that determines whether $\boldsymbol{c}$ is replaced by (one of) the best mutants or not (Instruction 2).

5.1.2 Gaussian Mutations and 1/5-Rule for the (μ+1) Evolution Strategy

In the (μ+1) ES framework each individual consists of a search point and an associated mutation strength. As we need the counters "b" and "g" for the adaptation of the individual mutation strength, each individual is associated with its own set of counters, so that an individual $\mathcal{X} = (\boldsymbol{x}, \sigma, g, b)$ is in $\mathbb{R}^n \times \mathbb{R}_{>0} \times \mathbb{N}_0 \times \mathbb{N}_0$.

Let $\mu : \mathbb{N} \to \mathbb{N}$ such that $\mu = \text{poly}(n)$. Then the "(μ+1) ES using Gaussian mutations adapted by the 1/5-rule" works as follows (for minimization): For a given initialization of the population of μ individuals (where all g- and b-counters are zero) the following evolution loop is performed:

1. Choose one of the individuals in the (current) population uniformly at random. Let this be $\mathcal{X} = (\boldsymbol{x}, \sigma, g, b)$.
2. Create a new search point $\boldsymbol{y} := \boldsymbol{x} + \boldsymbol{m}$ with a mutation (vector) $\boldsymbol{m} := \sigma \cdot \widetilde{\boldsymbol{m}}$, where each component of $\widetilde{\boldsymbol{m}}$ is independently standard-normally distributed
3. IF $f(\boldsymbol{y}) \leq f(\boldsymbol{x})$ THEN $g := g + 1$ ELSE $b := b + 1$.
4. IF $b + g = 5n$ THEN
   a) IF $g < (b+g) \cdot (1/5)$ THEN $\sigma := \sigma/2$ ELSE $\sigma := \sigma \cdot 2$;
   b) $g := 0$; $b := 0$.
5. Create the mutant $\mathcal{Y} := (\boldsymbol{y}, \sigma, g, b)$.
   (Note that $\mathcal{Y}$ inherits the possibly updated/adapted parameters σ, b, g from its parent $\mathcal{X}$.)
6. Discard one of the $\mu + 1$ individuals by uniformly choosing one of the worst individuals (maximal f-value).
7. GOTO 1.

5.1.3 Gaussian Mutations and 1/5-Rule and the Spatial Gain

Recall Corollary 3.13 (p. 27) and, in particular, the random variable $\widetilde{G}$ which corresponds to the signed distance of the mutant $\boldsymbol{c} + \widetilde{\boldsymbol{m}}$ from a predefined hyperplane containing the search point $\boldsymbol{c}$ which is mutated. Accordingly, we now let $\widetilde{\Delta}_d$ denote the spatial gain towards a fixed search point $\boldsymbol{x}^*$ with $d = \text{dist}(\boldsymbol{c}, \boldsymbol{x}^*)$. Furthermore, when the Gaussian mutation is scaled by σ, we let $\widetilde{\Delta}_{\sigma,d}$ denote this spatial gain. Formally, for fixed $\boldsymbol{c}, \boldsymbol{x}^* \in \mathbb{R}^n$

$$\widetilde{\Delta}_{\sigma,d} \quad := \quad d - \text{dist}(\boldsymbol{c} + \sigma \cdot \widetilde{\boldsymbol{m}}\,,\, \boldsymbol{x}^*) \tag{5.1}$$

where $d = \text{dist}(\boldsymbol{c}, \boldsymbol{x}^*)$ and $\widetilde{\boldsymbol{m}}$ is a Gaussian mutation, i. e., each of the n components is independently standard-normally distributed. (Recall that we can restrict ourselves to the distance d between $\boldsymbol{c}$ and $\boldsymbol{x}^*$ because of the isotropy of a Gaussian mutation.)

As mentioned above, the idea behind the 1/5-rule is to maximize the expected gain in a step of the (1+1) ES. For instance for SPHERE, a mutation is accepted if and only if the mutant is at least as close to the optimum as its parent. In this situation, the spatial gain of a step is given by $\widetilde{\Delta}^+_{\sigma,d}$ (which abbreviates $\widetilde{\Delta}_{\sigma,d} \cdot \mathbb{1}_{\{\widetilde{\Delta}_{\sigma,d} \geq 0\}}$), and the 1/5-rule is supposed to adapt σ such that the expected one-step gain of $\mathsf{E}[\widetilde{\Delta}^+_{\sigma,d}]$ is maximum.

Yet in fact, knowing $\max_{\sigma>0} \mathsf{E}[\widetilde{\Delta}^+_{\sigma,d}]$ for a given distance d from the optimum does not help with an analysis. The 1/5-rule is obviously not able to adapt σ such that expected spatial gain is actually maximum. Besides, we already know from Lemma 4.4 (p. 34) that for $n \geq 4$

$$\max_{\sigma>0} \mathsf{E}\big[\widetilde{\Delta}^+_{\sigma,d}\big] \;<\; 0.52 \cdot d/\sqrt{n-1} \;=\; O(d/\sqrt{n})$$

anyway. So the actual questions are: For which σ does $\mathsf{E}[\widetilde{\Delta}^+_{\sigma,d}] = \Omega(d/\sqrt{n})$ hold? Is the 1/5-rule able to keep σ in the respective range? And, if so, for how many iterations of the evolution loop?

In fact, we should not restrict ourselves to $\mathsf{E}[\widetilde{\Delta}^+_{\sigma,d}]$ since this is the expected spatial gain of a (1+1) ES on SPHERE. Nevertheless, the answer to the questions will be useful not only for the SPHERE scenario. Therefore, note that "$\widetilde{\Delta}_{\sigma,d} = \Omega(d/\sqrt{n})$ with an $\Omega(1)$-probability" implies that $\mathsf{E}[\widetilde{\Delta}^+_{\sigma,d}] = \Omega(d/\sqrt{n})$ because negative gains are zeroed out by the elitist selection in this scenario.

Of course, also the total gain of a sequence of steps will be of interest. In particular, we are interested in the total gain of a number of sequent steps in all of which the same mutation strength σ is used. As we shall see in the following, it is very unlikely that such a total gain is actually larger than the double of its expectation:

Therefore, assume that the (1+λ) ES uses for a phase of k steps a fixed isotropic distribution F to generate the mutants (i. e., for each mutation the mutation vector is independently drawn according to F). This is the case for Gaussian mutations adapted by a 1/5-rule during an observation period, for instance. Let $\Delta^{[1]}, \ldots, \Delta^{[k]}$ denote the random variables which respectively correspond to the gains in the k steps of the (1+λ) ES. Optimistically assume that any mutation that yields a positive spatial gain is accepted, and that any negative gain is rejected (as it is the case for SPHERE). Then the distance from the optimum is non-increasing, and hence, we have $\Delta^{[1]} \succ \cdots \succ \Delta^{[k]}$ (cf. Proposition 4.3 (p. 33)). Let $\Delta_1, \ldots, \Delta_k$ denote k independent copies of the random variable $\Delta^{[1]}$. Then the random variable $S := \Delta_1 + \ldots \Delta_k$ stochastically dominates the total gain of the phase, namely the random variable defined as $\Delta^{[1]} + \cdots + \Delta^{[k]}$.

Let d denote the distance from the optimum at the beginning of the phase. Assume that the isotropic distribution F is such that $\mathsf{E}[S] \leq d/4$ and note that $\mathsf{E}[\Delta^{[1]}] \leq (d/4)/k$ implies this upper bound on the expected total gain of the phase. Then Hoeffding's bound, namely Theorem 2.3 (p. 13), tells us (since $\mathsf{E}[S] + d/4 \leq d/2$) that

$$\mathsf{P}\{S \geq d/2\} \;\leq\; \exp\left(\frac{-2(d/4)^2}{k \cdot (b-a)^2}\right).$$

We can chose $a := 0$ since the gain of a step cannot be negative in our scenario. Substituting for b the trivial upper bound of d on Δ_i, results in an upper bound of $\mathrm{e}^{-(1/8)/k}$ on $\mathsf{P}\{S \geq d/2\}$, which, unfortunately, tends to one as k grows. Therefore, assume that Δ_i was bounded from above by $b := d \cdot n^\varepsilon/n$. Then we have

$$\mathsf{P}\big\{S \geq d/2 \mid \Delta_1, \ldots, \Delta_k \leq d \cdot n^\varepsilon/n\big\} \;\leq\; \exp\left(\frac{-d^2/8}{k \cdot (d \cdot n^\varepsilon/n)^2}\right) \;=\; \exp\left(\frac{-n^{2-2\varepsilon}}{8k}\right).$$

If k is $O(n)$, this upper bound on the probability is $\mathrm{e}^{-\Omega(n^{1-2\varepsilon})}$. Choosing $\varepsilon := 1/3$, we obtain

$$\mathsf{P}\big\{S \geq d/2 \mid \Delta_1, \ldots, \Delta_k \leq d \cdot n^{1/3}/n\big\} \;=\; \mathrm{e}^{-\Omega(n^{1/3})}. \tag{5.2}$$

With this upper bound we can now prove the following lemma which will later be useful in the analysis of the 1/5-rule.

Lemma 5.1. Let a (1+λ) ES minimize SPHERE in $\mathbb{R}^n$. Consider a phase of $k = O(n)$ steps in which all mutation vectors are independently drawn according to the same isotropic distribution F. If F is such that the expected gain towards the optimum in the first step of the phase is at most $(d^{[0]}/4)/k$, then the probability that the total gain of the phase is at least $d^{[0]}/2$ (i. e., the approximation error in the search space is halved) is bounded above by $e^{-\Omega(n^{1/3})}$.

Proof. According to Lemma 4.5 (p. 35) an isotropic mutation yields a gain of at least $d \cdot n^{1/3}/n$ only with probability $e^{-\Omega(n^{1/3})}$. (As a consequence, the probability that the best of λ mutations in a step yields such a gain is bounded from above by $\lambda \cdot e^{-\Omega(n^{1/3})} = e^{-\Omega(n^{1/3})}$.) Thus, if F is such that the expected gain of the first step of the phase is at most $(d/4)/k$, then $\mathsf{P}\{S \geq d/2\}$ (the probability that the approximation error is halved in the considered phase of k steps) is bounded from above by $\lambda \cdot k \cdot e^{-\Omega(n^{1/3})} + e^{-\Omega(n^{1/3})}$, which is $e^{-\Omega(n^{1/3})}$ since $\lambda \cdot k = \text{poly}(n)$. □

All the facts and arguments that we used to derive this lemma do not only hold for SPHERE, but for all functions that are "like SPHERE" in the following sense.

5.2 SPHERE-like Functions

Consider unimodal functions that are monotone with respect to the distance from the minimum. More formally, a function $f : \mathbb{R}^n \to \mathbb{R}$ belongs to this class and is called "SPHERE-like" if (and only if)

1. a minimizer $\boldsymbol{x}^* \in \mathbb{R}^n$ exists, i. e., $\forall \boldsymbol{x} \in \mathbb{R}^n : \ f(\boldsymbol{x}^*) \leq f(\boldsymbol{x})$, and
2. $\forall \boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n : \text{dist}(\boldsymbol{x}^*, \boldsymbol{x}) < \text{dist}(\boldsymbol{x}^*, \boldsymbol{y}) \Rightarrow \ f(\boldsymbol{x}) < f(\boldsymbol{y})$.

The crucial property of such a function with respect to the (1+1) ES is that any mutant which is closer to the minimum is accepted, whereas any mutant which is farther away is discarded. In other words, a reduction of the approximation error in the search space is always accepted, whereas an increase is always rejected. We do not know, however, whether a mutant with the same distance from the optimum as its parent $\boldsymbol{c}$ is accepted; yet this does not make any difference as the hyper-sphere centered at $\boldsymbol{x}^*$ and containing $\boldsymbol{c}$ has zero Lebesgue measure and, hence, is hit with zero probability. All in all, when starting with the same initial approximation error, the stochastic process induced by the (1+1) ES depends on the class-defining properties, but not on the function itself.

In particular, the function $\text{SPHERE}(\boldsymbol{x}) := \sum_{i=1}^{n} x_i^2 = |\boldsymbol{x}|^2$ belongs to our class, which is presumably the most investigated and most discussed function in theory-oriented work on evolution strategies; cf. for instance Rechenberg (1973, 1994), Schwefel (1995), Rudolph (1997), Beyer (2001), Bienvenue and Francois (2003), Auger (2005). And this is also the reason for the notion "SPHERE-like."

Obviously, the L^2-norm is SPHERE-like, and it is readily seen that a function $f = g \circ L^2$ belongs to our class if $g : \mathbb{R}_{\geq 0} \to \mathbb{R}$ is monotone increasing and bounded from below. With respect to the

trajectory of the evolving search point/population in the search space, the optimization process is independent of g; the progression of the approximation with respect to the objective space, however, crucially depends on g; consider for instance $g(x) = x^2$, i. e., $f = \text{SPHERE}$, as opposed to $g(x) = 2^x$. Results with respect to g can easily be obtained from ones with respect to the search space, and hence, it makes sense to concentrate on the approximation error in the search space, which is defined as the distance from the unique minimum $\boldsymbol{x}^* \in \mathbb{R}^n$. In particular, we may assume, for notational convenience, that the minimum $\boldsymbol{x}^*$ coincides with the origin so that the approximation error (in the search space) is given by $|\boldsymbol{c}|$.

5.2.1 SPHERE-like Functions and the (1+1) ES with 1/5-Rule

As already noted above, obviously, the 1/5-rule cannot ensure that each mutation is successful with a probability of exactly 1/5. Nevertheless, the question for which σ a step succeeds with a probability of 1/5 is interesting. Formally, we are interested in the specific σ for which $\mathsf{P}\{\widetilde{\Delta}_{\sigma,d} \geq 0\} = 1/5$. By using Equation (4.2) on page 33 with $\delta := 0$, we obtain

$$\mathsf{P}\{\widetilde{\Delta}_{\sigma,d} \geq 0 \mid |\sigma \cdot \widetilde{\boldsymbol{m}}| = \ell\} = 1/5 \quad \Longleftrightarrow \quad \mathsf{P}\{G_\ell \geq \ell^2/(2d)\} = 1/5.$$

Since the equation on the right is equivalent to $\mathsf{P}\{G \geq \ell/(2d)\} = 1/5$, Lemma 3.12 (p. 25) tells us that

$$\mathsf{P}\{\widetilde{\Delta}_{\sigma,d} \geq 0 \mid |\sigma \cdot \widetilde{\boldsymbol{m}}| = \ell\} = 1/5 \quad \Longrightarrow \quad \ell = \Theta(d/\sqrt{n}).$$

Recall from the reasoning that precedes Corollary 3.13 (p. 27) that $|\widetilde{\boldsymbol{m}}| \in [\sqrt{n}/2, 2\sqrt{n}]$ with probability $1 - O(1/n)$, and hence, we obtain analogously to that reasoning

$$\mathsf{P}\{\widetilde{\Delta}_{\sigma,d} \geq 0\} = 1/5 \quad \Longrightarrow \quad \sigma = \Theta(d/n).$$

Since all arguments remain valid when substituting "1/5" by an arbitrary constant $\varepsilon \in (0, 1/2)$ (so that ε as well as $1/2 - \varepsilon$ are $\Omega(1)$, cf. Corollary 3.13 (p. 27) again), we obtain

Lemma 5.2. Fix $d \in \mathbb{R}_{>0}$ and $\varepsilon \in (0, 1/2)$. Then $\mathsf{P}\{\widetilde{\Delta}_{\sigma,d} \geq 0\} = \varepsilon$ implies $\sigma = \Theta(d/n)$.

So, we considered a gain (of a mutant $\boldsymbol{c} + \boldsymbol{m}$) towards a point $\boldsymbol{x}^*$ (at distance d from $\boldsymbol{c}$) of size $\delta = 0$, which corresponds to a "parallel gain" of $g = \ell^2/(2d)$ when $|\boldsymbol{m}| = \ell$. What about a positive gain? When choosing, say, $\delta := d/n$ rather than zero, then the corresponding $g_{\ell,\delta}$ becomes $d/n + (\ell^2 - d^2/n^2)/(2d)$ (cf. Equation (4.2) on page 33). Thus, for $\ell = \Theta(d/\sqrt{n})$ we obtain a corresponding g that is $\Theta(d/n)$, i. e. $\Theta(\ell/\sqrt{n})$. Since the arguments hold for any δ that is $\Theta(d/n)$ rather than exactly d/n, we have in fact shown that, if $\ell = \Theta(d/\sqrt{n})$, then a δ which is $\Theta(d/n)$ corresponds to some g which is $\Theta(\ell/\sqrt{n})$. Recall that $G_\ell \sim \ell \cdot G$. Thus, we can finally apply Lemma 3.12 (p. 25) (Item 4) to obtain the following result (recall Equation (4.1) on page 32 for the definition of "$\Delta_{\boldsymbol{x}^*,\ell}$").

Lemma 5.3. Let $\boldsymbol{x}^* \in \mathbb{R}^n$ be fixed and $d = \text{dist}(\boldsymbol{c}, \boldsymbol{x}^*) > 0$. Given that ℓ is $\Theta(d/\sqrt{n})$, then for any constant ε we have $\mathsf{P}\{\Delta_{\boldsymbol{x}^*,\ell} \geq \varepsilon \cdot d/n\} = \Omega(1)$.

Since $|\widetilde{\boldsymbol{m}}| \in [\sqrt{n}/2, 2\sqrt{n}]$ with probability $1 - O(1/n)$ (as utilized several times), we obtain as a direct consequence

Corollary 5.4. Let $\boldsymbol{x}^* \in \mathbb{R}^n$ be fixed and $d = \operatorname{dist}(\boldsymbol{c}, \boldsymbol{x}^*) > 0$. Given that σ is $\Theta(d/n)$, then for any constant ε we have $\mathsf{P}\{\widetilde{\Delta}_{\sigma,d} \geq \varepsilon \cdot d/n\} = \Omega(1)$.

Putting it all together with Corollary 3.13 (p. 27) we obtain the following lemma which will be very frequently used in our analyses.

Lemma 5.5. Let $\boldsymbol{x}^* \in \mathbb{R}^n$ be fixed and $d = \operatorname{dist}(\boldsymbol{c}, \boldsymbol{x}^*) > 0$. Then $\mathsf{P}\{\widetilde{\Delta}_{\sigma,d} \geq 0\} = \Omega(1)$ as well as $1/2 - \mathsf{P}\{\widetilde{\Delta}_{\sigma,d} \geq 0\} = \Omega(1)$ if and only if $\sigma = \Theta(d/n)$, and if so, then for any constant ε we have $\mathsf{P}\{\widetilde{\Delta}_{\sigma,d} \geq \varepsilon \cdot d/n\} = \Omega(1)$.

In less formal words: If the mutation strength σ is such that the probability of the mutant being closer to the optimum is "roughly" 1/5, then the distance from the optimum is reduced by an $1/n$-fraction with a constant probability.

The lower bound on the one-step gain, which we have just obtained, will enable us to show our first result for a concrete scenario—once we have the following lemma (the counterpart of Lemma 4.6 (p. 36)).

Lemma 5.6. Let $X_1, X_2, \ldots$ denote random variables with bounded range and S the random variable defined by $S = \min\{t \mid X_1 + \cdots + X_t \geq g\}$ for a given $g > 0$. Given that S is a stopping time, if $\mathsf{E}[S] < \infty$ and $\mathsf{E}[X_i \mid S \geq i] \geq \ell > 0$ for $i \in \mathbb{N}$, then $\mathsf{E}[S] \leq \mathsf{E}[X_1 + \cdots + X_S]/\ell$.

Proof. First of all note that the X_i need not be independent—making the assumption necessary that S is a stopping time, though. Note that, since the X_i are bounded, the assumption/precondition $\mathsf{E}[S] < \infty$ implies $\mathsf{E}[X_1 + \cdots + X_S] < \infty$.

The proof follows the one of Lemma 4.6 (p. 36) up to the point where the lower bound ℓ on $\mathsf{E}[X_i \mid S \geq i]$ is utilized (rather than an upper bound which is called "u" therein).

$$
\begin{aligned}
& \mathsf{E}[X_1 + \cdots + X_S] \\
\overset{\text{cf. Lemma 4.6 (p. 36)}}{=} & \sum_{i=1}^{\infty} \mathsf{P}\{S \geq i\} \cdot \mathsf{E}[X_i \mid S \geq i] \\
\geq & \sum_{i=1}^{\infty} \mathsf{P}\{S \geq i\} \cdot \ell \\
= & \mathsf{E}[S] \cdot \ell
\end{aligned}
$$

□

So, this lemma (which may sound trivial) enables us to show our first result for a concrete and well-known scenario:

Theorem 5.7. Let the (1+1) ES using scaled Gaussian mutations optimize a SPHERE-like function in $\mathbb{R}^n$ using a fixed mutation strength σ (i. e. no mutation adaptation). Given that the initialization is such that $d^{[0]} > 0$ and $\sigma = \Theta(d^{[0]}/n)$, the expected number of steps i until $d^{[i]} \leq d^{[0]}/2$ is $\Theta(n)$, i. e., the expected number of steps to halve the approximation error in the search space grows linearly in the dimensionality of the search space.

Proof. The $\Omega(n)$-bound has already been shown in Theorem 4.8 (p. 39), so that we concentrate on the $O(n)$-bound in the following.

First of all note that negative gains are zeroed out by elitist selection in this scenario. As long as the approximation error has not been halved, in each step the approximation error is reduced by an $1/n$-fraction with probability $\Omega(1)$ since the distance from $\boldsymbol{x}^*$ is in $(d^{[0]}/2, d^{[0]}]$. Thus, the expected gain towards $\boldsymbol{x}^*$ is $\Omega(d^{[0]}/n)$ in each step (recall: negative gains are zeroed out). For the application of the previous Lemma 5.6 (p. 63), we let X_i denote the spatial gain towards the optimum in the ith step, and we know that we can choose a lower bound ℓ on the single-step gain which is $\Omega(d^{[0]}/n)$. Since the total gain of the steps (until $d^{[i]} \leq d^{[0]}/2$ for the first time) is obviously at most $d^{[0]}$, Lemma 5.6 (p. 63) yields an upper bound of $d^{[0]}/\Omega(d^{[0]}/n)$, which is $O(n)$, on the expected number of steps until $d^{[i]} \leq d^{[0]}/2$—if the expectation is finite (recall the precondition "$\mathsf{E}[S] < \infty$" in Lemma 5.6 (p. 63)).

Therefore, let B denote the hyper-ball exactly containing all points with a distance of at most $d^{[0]}/2$ from $\boldsymbol{x}^*$. We are interested in the number of iterations of the (1+1) ES until the evolving search point hits B. Since the mass of B w. r. t. the measure/distribution over $\mathbb{R}^n$ induced by adding $\sigma \cdot \tilde{\boldsymbol{m}}$ to some point $\boldsymbol{x} \in \mathbb{R}^n$ is positive (say lower bounded by $p > 0$ if $\boldsymbol{x}$'s distance from the center of B is at most $d^{[0]}$), the expected number steps until B is hit is indeed finite (at most $1/p$ in our case; formally, the trials are dependent, yet we can consider Bernoulli trials to obtain the upper bound of $1/p$). □

Unfortunately, unlike the lower bound in Theorem 4.8 (p. 39), the upper bound which we have just obtained is an asymptotic one, i. e., it tells us nothing about the constant hidden in the "$O(n)$." This constant depends on the actual relation between σ and $d^{[0]}$, and we only assume that the initialization results in $\sigma = \Theta(d^{[0]}/n)$. Yet what is more, in contrast to the lower bound, the upper bound can be iterated at most a constant times. That is, for any constant $\kappa \geq 1$, the expected number of steps until $d^{[i]} \leq d^{[0]}/2^\kappa$ is $O(n)$ by the very same arguments. But what about the number of steps until, say, $d^{[i]} \leq d^{[0]}/2^n$? For this question, considering an adaptation-less (1+1) ES does not make sense. For a fixed σ, the closer $\boldsymbol{c}$ gets to $\boldsymbol{x}^*$, the smaller the expected progress. And thus —even though $\boldsymbol{c}$ would converge (namely almost surely) towards $\boldsymbol{x}^*$, which is readily seen just because σ is fixed—the progress towards $\boldsymbol{x}^*$ would become slower and slower. And moreover, we would like an upper bound which holds with an overwhelming probability rather than only in expectation.

This is the point where the 1/5-rule comes into play. We must show that it keeps $\sigma = \Theta(d/n)$ as the optimization proceeds, i. e., that the mutation strength remains in the *evolution window* (this notion, in fact the German term *Evolutionsfenster,* was coined by Rechenberg (1973, p. 139), cf. Beyer (2001, pp. 17, 69) for instance).

Interestingly, we can show that the 1/5-rule works for SPHERE-like functions using the lower-bound result from Theorem 4.11 (p. 42); namely, we will utilize that after $O(n)$ steps of the

(1+1) ES the approximation error in the search space is still (at least) a constant fraction of the initial one (at least with probability $1-\mathrm{e}^{-\Omega(n)}$).

Theorem 5.8. Let a (1+1) ES using Gaussian mutations adapted by a 1/5-rule minimize a SPHERE-like function $f\colon \mathbb{R}^n \to \mathbb{R}$. If the initialization is such that $\sigma = \Theta(d/n)$, then the 1/5-rule maintains this property for an arbitrary polynomial number of steps with probability $1-\mathrm{e}^{-\Omega(n^{1/3})}$.

Proof. The run of a (1+1) ES is virtually partitioned into *phases* in each of which σ not changed. Recall from Lemma 5.5 (p. 63) that $\sigma = \Theta(d/n)$ is equivalent to the probability of generating a better mutant being bounded by $\Omega(1)$ as well as by $1/2-\Omega(1)$. This is crucial since this enables us to switch back and forth between considering the relative mutation strength σ/d in a step, on the one hand, and the mutation's success probability in that step, on the other hand. Namely, for a given mutation strength σ, we let $p_{\boldsymbol{c}} := \mathsf{P}\{f(\boldsymbol{c}+\sigma\cdot\widetilde{\boldsymbol{m}}) \le f(\boldsymbol{c})\}$ denote the success probability (of the mutation). Then $\sigma = \Theta(|\boldsymbol{c}|/n)$ if and only if there is a constant $\varepsilon > 0$ such that $p_{\boldsymbol{c}} \in [\varepsilon, 1/2-\varepsilon]$ for n large enough; we may drop the subscript "$\boldsymbol{c}$" in unambiguous situations. Note that for two search points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$ we have $p_{\boldsymbol{x}} \ge p_{\boldsymbol{y}} \iff |\boldsymbol{x}| \ge |\boldsymbol{y}|$ (cf. Proposition 4.3 (p. 33)). Since $|\boldsymbol{c}|$ is non-increasing in our scenario, by a trivial scaling argument, doubling σ after a phase surely results in a smaller success probability compared to any of the success probabilities in that phase. Halving σ at the end of a phase, however, results in a larger success probability compared to the success probability of the first mutation in that phase only if the approximation error has not been halved within this phase. As it is harder to tackle, we start our analysis with the latter situation.

Since in our scenario the distance from the optimum is non-increasing, p is also non-increasing during a phase. Let $p_{(i)}$ denote the success probability of the *first* mutation within the ith phase. Assume that σ is large such that at the beginning of the ith phase the success probability is small, say, $p_{(i)} \le \varepsilon < 0.1$ yet still $p_{(i)} = \Omega(1)$. (The positive constant ε will be chosen appropriately small later.) To show that the 1/5-rule works, we have to show that σ will be halved after the ith phase, and that this does result in $p_{(i+1)} \ge p_{(i)}$, i. e. in an increase in the success probability. If this is the case, then the success probability of the last mutation in the ith phase is a lower bound on the success probabilities that occur. To see that this threshold, namely the success probability of the last mutation in the ith phase, is indeed $\Omega(1)$ if $p_{(i)}$ is $\Omega(1)$, recall the lower bound from Theorem 4.11 (p. 42). It tells us (by choosing b as a constant large enough) that after the ith phase, which lasts $\Theta(n)$ mutations, the distance from the optimum is a constant fraction of the one at the beginning of the phase with probability $1-\mathrm{e}^{-\Omega(n)}$. Given that this is the case, also the ratio σ/d at the end of the ith phase is of the same order as at the beginning of the phase, implying that $p = \Omega(1)$ at the end of the ith phase (given that $p_{(i)} = \Omega(1)$, of course). In the following, we assume that this is the case (and keep in mind that we err with a probability of $\mathrm{e}^{-\Omega(n)}$).

Thus, in each mutation within the ith phase $\varepsilon \ge p = \Omega(1)$, and hence, we expect at most an ε-fraction of the mutations in this phase to be successful. By Chernoff's bound, with probability $1-\mathrm{e}^{-\Omega(n)}$ (since we expect $\Omega(n)$ successful mutations) at most a 2ε-fraction of the mutations are actually successful . Again we assume that this is the case (and again we keep in mind that we err with a probability of $\mathrm{e}^{-\Omega(n)}$).

Since $2\varepsilon < 1/5$, less than 20% of the mutations are successful so that after the ith phase the scaling factor σ is halved, resulting in an increase of the success probability—when comparing $p_{(i+1)}$ with the success probability of the last mutation in the ith phase. The crucial question is, however, whether $p_{(i+1)} \ge p_{(i)}$.

Here is the point where the choice of ε comes into play. Not only the upper bound on the (expected) number of successful mutations in the phase is proportional to ε, yet also the total gain of the ith phase; in particular, we can choose ε small enough (i. e., $\sigma = O(d/n)$ large enough) such that the distance from the optimum is not halved within the ith phase with probability $1 - \mathrm{e}^{-\Omega(n^{1/3})}$ (Lemma 5.1 (p. 61)), i. e., the increase of the success probability due to the halving of σ after the ith phase overbalances the decrease of the success probability which is due to the reduction of the approximation error within the ith phase. Then, as already noted above, the success probability of the last mutation in the ith phase (which is $\Omega(1)$ under our assumptions) is the lower bound on the success probabilities which occur. This $\Omega(1)$-threshold on the mutations' success probabilities corresponds to σ being bounded by $O(d/n)$.

Since things may go wrong (i. e., our assumptions are not met) with a probability of $\mathrm{e}^{-\Omega(n^{1/3})}$, our reasoning does not show that $\sigma = O(d/n)$, i. e. $p = \Omega(1)$, "forever with probability one", yet "for any polynomial number of phases with probability $1 - \mathrm{e}^{-\Omega(n^{1/3})}$" because adding up a polynomial number of error probabilities each of which is $\mathrm{e}^{-\Omega(n^{1/3})}$ results in a total error probability which is bounded by $\mathrm{e}^{-\Omega(n^{1/3})}$ (using the union bound).

Fortunately, the upper threshold of $1/2 - \Omega(1)$ on the mutations' success probabilities, i. e., that σ remains $\Omega(d/n)$, is easier to show (as already noted at the very beginning of this proof). Therefore, assume that the mutation strength σ is small such that in the last step of the jth phase the success probability is large, say, $p \in [0.3, 0.4]$. Since during a phase p is non-increasing, we expect at least 30% of the mutations in the jth phase to be successful, i. e. $\Omega(n)$ many. By Chernoff's bound, with probability $1 - \mathrm{e}^{-\Omega(n)}$ more than 20% of the mutations in the jth phase are actually successful, so that σ is doubled. This results in a smaller $p_{(j+1)}$ compared to the last mutation of the jth phase—yet also compared to $p_{(j)}$, the success probability of the first mutation in the jth phase (cf. above). To see that also $p_{(j)}$ (our upper threshold on the mutations' success probabilities) is bounded above by $1/2 - \Omega(1)$ if the success probability in the last mutation of the jth phase is at most 0.4, recall that we have $p_{(j)} = 1/2 - \Omega(1)$ if the distance at the end of the phase is at least a constant fraction of the one at the beginning, which is the case with probability $1 - \mathrm{e}^{-\Omega(n)}$ (by choosing b as a constant large enough in Theorem 4.11 (p. 42) such that "$b \cdot 0.69n/\ln(1+3\lambda)$" is at least the number of iterations in the jth phase). Thus, for any polynomial number of phases, with probability $1 - \mathrm{e}^{-\Omega(n)}$ the success probability p remains bounded from above by $1/2 - \Omega(1)$, i. e., σ remains bounded by $\Omega(d/n)$.

Altogether we have shown that, if $\sigma^{[0]} = \Theta(d^{[0]}/n)$ after initialization, then $\sigma = \Theta(d/n)$ for an arbitrary polynomial number of steps—at least with probability $1 - \mathrm{e}^{-\Omega(n^{1/3})}$. □

Note that in this proof of that the 1/5-rule works for the (1+1) ES on a SPHERE-like function, we merely used that the observation period (a phase) lasts $\Theta(n)$ mutations, rather than exactly $5n$. Moreover, increasing σ by 10%, say, rather than by 100% (doubling) surely results in a decrease in the success probability. Moreover, reducing σ by 30%, say, rather than by 50% (halving) after a phase results in a larger success probability unless the approximation error has been reduced by at least 30% within that phase, which is also just a constant fraction. Finally, we could consider a 1/6-rule or a 1/3-rule, for instance. In the case of a 1/3-rule, in the reasoning for the upper threshold of $1/2 - \Omega(1)$ on the success probabilities, we would consider the interval $[1/3 + \varepsilon/2, 1/3 + \varepsilon]$ for some positive constant $\varepsilon < 1/2 - 1/3$, rather than "$[0.3, 0.4]$," of course.

Corollary 5.9. Theorem 5.8 (p. 65) does not only hold for the 1/5-rule that observes $5n$ mutations and doubles/halves the mutation strength, but for any 1/5-rule which observes $\Theta(n)$ mutations and up-/down-scales σ using two predefined positive constants which are larger resp. smaller than one.

Moreover, the theorem holds for analogous ε-rules, where $\varepsilon \in (0, 1/2)$ is a fixed constant.

Naturally, an observation period of n would result for $n = 1$ in a σ-adaptation that would presumably fail because after each mutation/step σ would be up-/down-scaled, depending on whether this single step has been successful or not. This is no contradiction, however, since in that case the error probability "$e^{-\Omega(n)}$" may be very very close to one.

Now that we have proved that the 1/5-rule works—in the considered scenario—, we can easily show an upper bound on the runtime:

Theorem 5.10. Let a (1+1) ES using Gaussian mutations adapted by a 1/5-rule minimize a SPHERE-like function in $\mathbb{R}^n$, and let $b\colon \mathbb{N} \to \mathbb{N}$ such that $b = \text{poly}(n)$. If the initialization is such that $\sigma^{[0]} = \Theta(d^{[0]}/n)$, then the number of steps i until $d^{[i]} \leq d^{[0]}/2^{b(n)}$ is $\Theta(b_{(n)} \cdot n)$ with probability $1 - e^{-\Omega(n^{1/3})}$.

Proof. The $\Omega(bn)$-bound has already been shown in Theorem 4.11 (p. 42), so that we concentrate on the $O(b \cdot n)$-bound here.

Recall that the 1/5-rule ensures $\sigma = \Theta(d/n)$ for any polynomial number of steps (at least with probability $1 - e^{-\Omega(n^{1/3})}$), in particular for any number of steps which is $O(b \cdot n)$.

Let κ denote a constant, which will be chosen large enough a posteriori. Within κbn steps, in each of which $\sigma = \Theta(d/n)$, each step reduces the approximation error at least by d/n with an $\Omega(1)$-probability (cf. Lemma 5.5 (p. 63)). Thus, the expected number of steps each of which reduces the approximation error by (at least) an $1/n$-fraction is $\Omega(\kappa bn)$. Since $(1-1/n)^{n \cdot \ln 2} \leq 1/2$ for $n \geq 2$ (and since the approximation error is non-increasing), after at most $0.7n$ such steps the approximation error is halved, and after $0.7bn$ such steps the approximation is less than $d^{[0]}/2^b$. Now, by choosing κ large enough, the expected number of such steps is at least bn, and by Chernoff's bound, the probability that less than $0.7bn$ such steps occur within the κbn steps is bounded by $e^{-\Omega(b \cdot n)}$.

All in all, we have shown that within $\kappa bn = O(bn)$ steps with probability $1 - e^{-\Omega(b \cdot n)}$ at least $0.7bn$ of them reduce the approximation error by at least d/n, respectively, and that this implies that the approximation error has become smaller than a 2^{-b}-fraction of the initial one—under the assumption that the 1/5-rule works (i. e., $\sigma = \Theta(d/n)$ in all κbn steps). As this is the case with probability $1 - e^{-\Omega(n^{1/3})}$ (as shown above), the total error probability is also bounded by $e^{-\Omega(n^{1/3})}$. □

The proof has been apparently simple. This is because most of the effort has gone into the proof of that the 1/5-rule works (in the considered scenario). Again we have to keep in mind the asymptotic nature of the result. For low-dimensional search spaces, fine-tuning the 1/5-rule (namely its parameters) may well make sense. Such a tuning, however, cannot change how the runtime scales with the dimension of the search space, that is the point. The concrete implementation of the 1/5-rule influences only the constant in "$O(b \cdot n)$"—it cannot do anything against that $\Omega(b \cdot n)$ mutations are necessary (with an overwhelming probability).

This can also be interpreted as some kind of robustness result: Even if the parameters of the 1/5-rule are not fine-tuned, $O(b \cdot n)$ steps suffice with an overwhelming probability.

We stick with the function scenario, namely we stick with SPHERE-like functions, yet switch to the (1+λ) ES now.

5.2.2 SPHERE-like Functions and the (1+λ) ES with 1/5-Rule

An observation period of the 1/5-rule lasts $\Theta(n)$ mutations, i. e. $\Theta(\lceil n/\lambda \rceil)$ steps. Yet the number of steps which are necessary to halve the approximation error is $\Omega(n/\ln(1+\lambda))$ with probability $1-e^{-\Omega(n)}$ as we have shown in Theorem 4.11 (p. 42). In other words, since in each step λ samples are drawn at the same location in the search space, the 1/5-rule can adapt σ more accurately, because the total number of samples between two sequent σ-adaptations is still $5n$ (or $\Theta(n)$ for the generalized 1/5-rule). In particular, the larger λ, the smaller the chance that halving σ after a phase does not result in an increase of the success probability.

As a consequence, for $\lambda = O(n)$, each and every argument within the reasoning in the proof of Theorem 5.8 (p. 65) (in which we have shown that the 1/5-rule works for the (1+1) ES on a SPHERE-like function) carries over because a phase consists of $\Theta(n)$ mutations. This fact was used in the two applications of the Chernoff bound to obtain an error probability of $e^{-\Omega(n)}$ because of an expectation that is $\Theta(n)$, respectively. Now, if λ is such that σ is adapted after every step, which implies that $\lambda = \Omega(n)$, then the two expectations[1] are of order $\Theta(\lambda)$, respectively, so that the error probabilities are of order $e^{-\Omega(\lambda)}$, i. e., they are still $e^{-\Omega(n)}$ since $\lambda = \Omega(n)$. Thus, the proof carries over not only for λ that are $O(n)$ but for any $\lambda = \text{poly}(n)$.

Corollary 5.11. Let a (1+λ) ES using Gaussian mutations adapted by a 1/5-rule minimize a SPHERE-like function $f : \mathbb{R}^n \to \mathbb{R}$. If the initialization is such that $\sigma = \Theta(d/n)$, then the 1/5-rule maintains this property for an arbitrary polynomial number of steps with probability $1-e^{-\Omega(n^{1/3})}$.

This is also true when considering the more general notion of a 1/5-rule as described in Corollary 5.9 (p. 67).

To obtain an upper bound on the runtime, however, we need to know the gain that the λ mutations in a step of the (1+λ) ES yield. For the (1+1) ES, this gain is given by the random variable $\widetilde{\Delta}_{\sigma,d}$. Since in the (1+λ) ES the λ mutants in a step are generated using the same σ, we have λ independent samples w. r. t. the same distribution. Hence, the maximum of λ independent instances of $\widetilde{\Delta}_{\sigma,d}$ corresponds to the gain of the mutants, namely to the gain of the best of them. This is commonly called the λth order statistic (of λ copies) of $\widetilde{\Delta}_{\sigma,d}$, denoted here by $\widetilde{\Delta}_{\sigma,d}^{\langle\lambda:\lambda\rangle}$.

The proof of Theorem 5.10 (p. 67) is mainly due to the observation that—given that the mutation strength σ is $\Theta(d/n)$—a mutation reduces the approximation error by d/n with probability $\Omega(1)$, i. e., we utilize that $\mathsf{P}\{\widetilde{\Delta}_{\sigma,d} \geq d/n)\} = \Omega(1)$ for $\sigma = \Theta(d/n)$. When we want to adopt this approach, we merely need to know for which (function) α we have $\mathsf{P}\{\widetilde{\Delta}_{\sigma,d}^{\langle\lambda:\lambda\rangle} \geq \alpha \cdot d/n)\} = \Omega(1)$ for $\sigma = \Theta(d/n)$. Obviously, $\alpha = \Omega(1)$ because the best of the mutants is considered. (Besides, the lower-bound result from Theorem 4.11 (p. 42) tells us that $\alpha = O(\ln(1+\lambda))$.) Let $\lambda \geq 2$ in the following. If α is such that $\mathsf{P}\{\widetilde{\Delta}_{\sigma,d} \geq \alpha \cdot d/n)\} \geq 1/\lambda$, then $\mathsf{P}\{\widetilde{\Delta}_{\sigma,d}^{\langle\lambda:\lambda\rangle} \geq \alpha \cdot d/n)\} \geq 1-(1-1/\lambda)^\lambda \geq 1-1/e > 0.63$, i. e., a step of the (1+λ) ES realizes a gain of at least $\alpha \cdot d/n$ with probability $\Omega(1)$.

[1] namely the expected number of successful steps and the expected number of steps each of which yields a gain of at least d/n

Now, recall Lemma 3.12 (p. 25) (in particular Item 1) which deals with unit isotropic mutations and their signed distance from a fixed hyper-plane. Since $\sqrt{n}\cdot g\cdot \mathrm{e}^{-\Theta(g^2\cdot n)} \geq 1/\lambda$ for some function g_1 which is $\Theta(\sqrt{\ln\lambda/n})$, we have $\mathsf{P}\{G \geq g_1\} \geq 1/\lambda$. Thus, for an isotropic mutation of length ℓ,

$$\mathsf{P}\{G_\ell \geq g_\ell\} \geq 1/\lambda \text{ for some } g_\ell \text{ which is } \Theta(\ell\cdot\sqrt{\ln\lambda/n}), \tag{5.3}$$

and consequently, for ℓ which are $\Theta(d/\sqrt{n})$, this g_ℓ is $\Theta(\sqrt{\ln\lambda}\cdot d/n)$. Since the length of a Gaussian mutation is in $[\sqrt{n}/2, 2\sqrt{n}]$ with probability $1-O(1/n)$, we obtain that for $\sigma=\Theta(d/n)$, with probability $(1/\lambda)\cdot(1-O(1/n)) \geq 0.5/\lambda$ for n large enough, $\widetilde{G}_\sigma = \Omega(\sqrt{\ln\lambda}\cdot d/n)$. Recalling the interrelation between the gain δ_ℓ towards a fixed point (at distance d) and the signed distance g_ℓ given in Equation (4.2) on page 33, we see that $\delta_\ell \geq g_\ell - \ell^2/(2d) = g_\ell - \Theta(d/n)$ for $\ell=\Theta(d/\sqrt{n})$. Hence, with a probability of at least $0.5/\lambda$ also $\widetilde{\Delta}_{\sigma,d}$ is $\Omega(\sqrt{\ln\lambda}\cdot d/n)$. Finally using $(1-0.5/\lambda)^\lambda \leq \mathrm{e}^{-0.5} < 0.61$, we have shown

Lemma 5.12. Let $\sigma=\Theta(d/n)$. Then there is a δ which is $\Omega(\sqrt{\ln\lambda}\cdot d/n)$ such that for n large enough $\mathsf{P}\left\{\widetilde{\Delta}^{\langle\lambda:\lambda\rangle}_{\sigma,d} \geq \delta\right\} \geq 0.39$.

With the help of this lemma we can now prove an upper bound on the runtime of the (1+λ) ES for the considered scenario.

Theorem 5.13. Let a (1+λ) ES, $\lambda \geq 2$, using Gaussian mutations adapted by a 1/5-rule minimize a SPHERE-like function in $\mathbb{R}^n$. Let $b\colon \mathbb{N}\to\mathbb{N}$ such that $b=\mathrm{poly}(n)$. If the initialization is such that $\sigma^{[0]}=\Theta(d^{[0]}/n)$, then the number of steps i until $d^{[i]} \leq d^{[0]}/2^{b(n)}$ is $O(b_{(n)}\cdot n/\sqrt{\ln\lambda})$ with probability $1-\mathrm{e}^{-\Omega(n^{1/3})}$.

Proof. The proof follows the one of Theorem 5.10 (p. 67).

Recall that the 1/5-rule ensures $\sigma=\Theta(d/n)$ for any polynomial number of steps (at least with probability $1-\mathrm{e}^{-\Omega(n^{1/3})}$), in particular, for any number of steps that is $O(b\cdot n/\sqrt{\ln\lambda})$.

Let κ denote a constant which will be chosen large enough later. Within $\kappa bn/\sqrt{\ln\lambda}$ steps, in each of which $\sigma=\Theta(d/n)$, each step reduces the approximation error by $\Omega(\sqrt{\ln\lambda}\cdot d/n)$ with an $\Omega(1)$-probability (cf. the preceding lemma). Thus, the expected number of steps each of which reduces the approximation error by $\Omega(\sqrt{\ln\lambda}\cdot d/n)$ is $\Omega(\kappa bn)$. Since $(1-\Omega(\sqrt{\ln\lambda})/n)^s \leq 1/2$ for n large enough for some s which is $O(n/\sqrt{\ln\lambda})$, after at most s such steps the approximation error is halved; and after $b\cdot s$ such steps the approximation is less than $d^{[0]}/2^b$. Now, by choosing κ large enough, the expected number of such steps is at least $2bs$, and by Chernoff's bound, the probability that less than bs such steps actually occur within the κbn steps is $\mathrm{e}^{-\Omega(bs)}$, i. e. $\mathrm{e}^{-\Omega(bn/\sqrt{\ln\lambda})}$ which is $\mathrm{e}^{-\Omega(bn/\sqrt{\ln n})}$ because $\lambda=\mathrm{poly}(n)$.

All in all, we have shown that within $\kappa bn/\sqrt{\ln\lambda} = O(bn/\sqrt{\ln\lambda})$ steps with a probability of at least $1-\mathrm{e}^{-\Omega(bn/\sqrt{\ln n})}$ the approximation error has become smaller than $d^{[0]}/2^{b(n)}$—under the assumption that the 1/5-rule works, i. e., that $\sigma=\Theta(d/n)$ in all $\kappa bn/\sqrt{\ln\lambda}$ steps. Since this is the case with probability $1-\mathrm{e}^{-\Omega(n^{1/3})}$, the total error probability is also bounded by $\mathrm{e}^{-\Omega(n^{1/3})}$. □

So, the proof is again simple, yet—unlike for the (1+1) ES—the result is not completely satisfying: The lower bound from Theorem 4.11 (p. 42) tells us that w. o. p. $\Omega(n/\ln\lambda)$ steps are necessary to halve the approximation error. The upper bound that we have just proved, however, says that w. o. p. $O(n/\sqrt{\ln\lambda})$ steps suffice, i. e., the bounds are not asymptotically tight, but off by a factor of order $\sqrt{\ln\lambda}$. There are three potential reasons for this:

1. The lower bound is weak.
2. The upper bound is weak.
3. The 1/5-rule just fails to make the (1+λ) ES get along with a number of steps that is at most by a constant larger than the optimal number of steps (w. r. t. our (1+λ) ES framework).

The gap between the bounds is solely due to the failure of the 1/5-rule—it will turn out that our lower bound is indeed sharp (w. r. t. the asymptotic order). So why does the 1/5-rule fail for the (1+λ) ES? The intuition behind the reason is simple: When you know that you have several trials, you should go a higher risk in a trial. Recall: The idea behind the 1/5-rule is to maximize the expected gain in a step of the (1+1) ES (on SPHERE). So, how can a simple rule maximize the expected gain of a step consisting of λ trials/mutations? Interestingly, also a 1/5-rule can—at least for the (1+λ) ES a "proper" 1/5-rule can.

5.2.3 SPHERE-like Functions and a Modified 1/5-Rule for the (1+λ) ES

We modify the 1/5-rule as follows: Rather than trying to adapt σ such that each mutation succeeds with a probability of (close to) 1/5, σ should be such that *each step* of the (1+λ) ES succeeds with a probability of (close to) 1/5. This results in the following (1+λ) ES with *modified 1/5-rule based on the steps' success probabilities* rather than on the mutations' success probabilities:

With $b := 0$ and $g := 0$ and a given initialization of the evolving search point $\boldsymbol{c} \in \mathbb{R}^n$ and the mutation strength $\sigma \in \mathbb{R}_{>0}$, the following evolution loop is performed (the instructions implementing the modified 1/5-rule are marked gray):

1. FOR $i := 1$ TO λ DO
   Create a new search point $\boldsymbol{y}_{[i]} := \boldsymbol{c} + \boldsymbol{m} \in \mathbb{R}^n$ with $\boldsymbol{m} := \sigma \cdot \widetilde{\boldsymbol{m}}$, where each of the n components of $\widetilde{\boldsymbol{m}}$ is independently standard-normally distributed.
2. IF $\min_{i\in\{1,\ldots,\lambda\}}\{f(\boldsymbol{y}_{[i]})\} \leq f(\boldsymbol{c})$ THEN BEGIN
   a) $\boldsymbol{c} := \operatorname{argmin}_{i\in\{1,\ldots,\lambda\}}\{f(\boldsymbol{y}_{[i]})\}$ (when there are more than one mutant with minimum fitness, one of them is chosen uniformly at random)
   b) $g := g + 1$ END

   ELSE $b := b + 1$.
3. IF $b + g \geq 5n/\log_2(1+\lambda)$ THEN BEGIN
   a) IF $g < (g+b)\cdot(1/5)$ THEN $\sigma := \sigma/2$ ELSE $\sigma := \sigma \cdot 2$.
   b) $g := 0$. $b := 0$. END
4. GOTO 1.

Note that σ is adapted every $\lceil 5n/\log_2(1+\lambda)\rceil$ steps (rather than $\lceil 5n/\lambda\rceil$ as in the 1/5-rule that is based on the number of successful mutations; for $\lambda = 1$ the two rules do not differ). The reason for this choice is due to the general lower bound that we have proved. $\Omega(n/\ln(1+\lambda))$ steps are

necessary w. o. p. to halve the approximation error. The observation phase—after which σ is halved (or doubled)—does not last longer (by more than a constant factor) than the number of steps necessary to halve the approximation error. Thus, halving σ after a phase should indeed result in an increase of the success probability by the same reasoning that we have followed in the proof of that the 1/5-rule works for the (1+1) ES. (As before, doubling σ surely results in a decrease in the success probability anyway.)

Interestingly, we have almost already shown that the modified 1/5-rule adapts σ such that it is $\Theta(\sqrt{\ln\lambda}\cdot d/n)$, which is by a factor of order $\sqrt{\ln\lambda}$ larger than with the original 1/5-rule. Therefore, recall that we looked at $\widetilde{\Delta}_{\sigma,d}^{\langle\lambda:\lambda\rangle}$ and, in particular, at $G_\ell^{\langle\lambda:\lambda\rangle}$. Since the best of λ independent identical trials succeeds with probability $\Omega(1)$ if one trial succeeds with probability $\Omega(1/\lambda)$, and since $\mathsf{P}\{G_\ell \geq g_\ell\} = \Omega(1/\lambda)$ for gains g_ℓ that are $O(\ell\sqrt{\ln\lambda/n})$ (cf. the reasoning that has led to Inequality (5.3) on page 69), we obtain—using Equation (4.2) on page 33 with $\delta := 0$ and solving $g_\ell = \ell^2/(2d)$ for ℓ—that $\mathsf{P}\{\Delta_{d,\ell} \geq 0\} = \Omega(1/\lambda)$, i. e., $\mathsf{P}\{\Delta_{d,\ell}^{\langle\lambda:\lambda\rangle} \geq 0\} = \Omega(1)$, for $\ell = O(d\cdot\sqrt{\ln\lambda/n})$.

Starting with the question for which g_ℓ the probability $\mathsf{P}\{G_\ell \geq g_\ell\}$ is *at most* $1/\lambda$ (instead of "at least"), by the symmetric reasoning we obtain that $\mathsf{P}\{\Delta_{d,\ell}^{\langle\lambda:\lambda\rangle} \geq 0\}$ is bounded above by $1/2-\Omega(1)$ for $\ell = \Omega(d\cdot\sqrt{\ln\lambda/n})$. Then, again utilizing that the length of a scaled Gaussian mutation deviates only very little from its expectation $\mathsf{E}[|\sigma\cdot\widetilde{\boldsymbol{m}}|] \asymp \sigma\sqrt{n}$, we obtain

$$\mathsf{P}\left\{\widetilde{\Delta}_{d,\sigma}^{\langle\lambda:\lambda\rangle} \geq 0\right\} \text{ is bounded } \begin{cases} \text{below by } \Omega(1) & \Longrightarrow \sigma = O(\sqrt{\ln\lambda}\cdot d/n) \\ \text{above by } 1/2-\Omega(1) & \Longrightarrow \sigma = \Omega(\sqrt{\ln\lambda}\cdot d/n). \end{cases}$$

Assume σ was such that $\mathsf{P}\{\widetilde{\Delta}_{d,\sigma}^{\langle\lambda:\lambda\rangle} \geq 0\} = 1/5$, implying that $\sigma = \Theta(\sqrt{\ln\lambda}\cdot d/n)$. As the length of the mutation vector is in the interval $[\sigma\sqrt{n}/2, 2\sigma\sqrt{n}]$ with probability $1-O(1/n)$, consider an ℓ that is $\Theta(d\sqrt{\ln\lambda/n})$ in the following. Then, by choosing $\delta := \ln\lambda\cdot d/n$ in Equation (4.2) on page 33, we obtain an corresponding g_δ which is $\Theta(\ln\lambda\cdot d/n)$. Thus, g_δ is of the same order as $g_0 = \ell^2/(2d)$, the signed distance (from the hyper-plane containing the parent) that corresponds to a zero gain towards the optimum. As each of the λ mutants yields a gain of that order with probability $\Omega(1/\lambda)$ (as shown above), we obtain that each mutant yields a gain of at least $\ln\lambda\cdot d/n$ with probability $\Omega(1/\lambda)$. Hence, the best of them yields a gain towards the optimum of at least $\ln\lambda\cdot d/n$ with probability $1-(1-\Omega(1/\lambda))^\lambda = \Omega(1)$. As our assumption on ℓ holds true with probability $1-O(1/n)$, we have indeed shown the following:

Lemma 5.14. Let $\lambda \geq 2$ and $\sigma = \Theta(\sqrt{\ln\lambda}\cdot d/n)$. Then $\mathsf{P}\{\widetilde{\Delta}_{\sigma,d}^{\langle\lambda:\lambda\rangle} \geq \ln\lambda\cdot d/n\} = \Omega(1)$.

Using this lemma we can show the upper bound on the runtime of the (1+λ) ES with the modified 1/5-rule—once we have shown that this rule keeps $\sigma = \Theta(\sqrt{\ln\lambda}\cdot d/n)$. Yet this can again be shown analogously to the proof of Theorem 5.8 (p. 65).

Theorem 5.15. Let a (1+λ) ES, $2 \leq \lambda = \text{poly}(n)$, using Gaussian mutations adapted by the modified 1/5-rule minimize a SPHERE-like function in $\mathbb{R}^n$. If the initialization is such that $\sigma = \Theta(\sqrt{\ln\lambda}\cdot d/n)$, then the modified 1/5-rule maintains this property for an arbitrary polynomial number of steps with probability $1-\mathrm{e}^{-\Omega(n^{1/3})}$.

Proof. The run of an $(1+\lambda)$ ES is virtually partitioned into phases of length $\Theta(n/\ln\lambda)$ (in each of which σ is not changed). Recall that $\sigma = \Theta(\sqrt{\ln\lambda}\cdot d/n)$ is equivalent to the probability of generating a better mutant in a step (which consists of λ mutations) being bounded by $\Omega(1)$ as well as by $1/2-\Omega(1)$. This is crucial since it enables us to switch back and forth between considering the relative mutation strength in a step, on the one hand, and the success probability of that step, on the other hand. So, this time, we let p denote the *step's success probability.* Then $\sigma = \Theta(\sqrt{\ln\lambda}\cdot d/n)$ if and only if there is an $\varepsilon \in \mathbb{R}_{>0}$ such that $p \in [\varepsilon, 1/2-\varepsilon]$ for n large enough.

Assume that σ is small such that in the last step of a phase $p \in [0.3, 0.4]$. Since p is non-increasing, each of the $\Theta(n/\ln\lambda)$ steps in the phase succeeds with a probability of at least 0.3. Thus, we expected at least 30% of the steps, i. e. $\Theta(n/\ln\lambda)$ many, to succeed. By Chernoff's bound, more than 20% of them are actually successful with a probability of $1-\mathrm{e}^{-\Omega(n/\ln\lambda)}$, which is $1-\mathrm{e}^{-\Omega(n/\ln n)}$ because $\lambda = \mathrm{poly}(n)$. Thus, σ is doubled, which surely results in a smaller p (since the approximation error cannot increase). As σ is such that in the last step of a phase $p \le 0.4$, then also in the first step of the phase $p = 1/2-\Omega(1)$ unless the approximation error has been reduced by more than a constant fraction in this phase—which happens only with a probability of at most $\mathrm{e}^{-\Omega(n)}$ according to the lower bound in Theorem 4.11 (p. 42). Hence, p remains upper bounded by $1/2-\Omega(1)$, i. e., σ remains $\Omega(\sqrt{\ln\lambda}\cdot d/n)$.

Now assume that σ is large such that in the first step of the ith phase $p \le \varepsilon < 0.1$ yet $p = \Omega(1)$, implying $\sigma = \Omega(\sqrt{\ln\lambda}\cdot d/n)$. Since p is non-increasing, we expect at most 10% of the steps (namely $\Theta(n/\ln\lambda)$ many) to be successful, and again by Chernoff's bound, with a probability of $1-\mathrm{e}^{-\Omega(n/\ln\lambda)}$ less than 20% are actually successful, so that σ is halved. By choosing the constant ε small enough, not only the number of successful steps can be made small enough, but the total gain of the phase can be made so small that the approximation is halved in this phase only with probability $\mathrm{e}^{-\Omega(n^{1/3})}$ (Lemma 5.1 (p. 61)). Hence, with this error probability the halving of σ results in $p_{(i+1)} \ge p_{(i)}$. Thus, the success probability of the last step in the ith phase is the lower threshold on the steps' success probabilities, and this threshold is $\Omega(1)$ since the approximation error has at most been halved in the ith phase. Finally, recall that $p = \Omega(1)$ implies $\sigma = O(\sqrt{\ln\lambda}\cdot d/n)$.

As we have a polynomial number of error probabilities which are $\mathrm{e}^{-\Omega(n^{1/3})}$ each, the total error probability is also/still bounded by $\mathrm{e}^{-\Omega(n^{1/3})}$. □

Now the upper-bound result:

Theorem 5.16. Let a $(1+\lambda)$ ES, $\lambda \ge 2$, using Gaussian mutations adapted by the modified 1/5-rule minimize a SPHERE-like function in $\mathbb{R}^n$. Let $b\colon \mathbb{N} \to \mathbb{N}$ such that $b = \mathrm{poly}(n)$. If the initialization is such that $\sigma^{[0]} = \Theta(\sqrt{\ln\lambda}\cdot d^{[0]}/n)$, then the number of iterations i until $d^{[i]} \le d^{[0]}/2^{b(n)}$ is $O(b_{(n)}\cdot n/\ln\lambda)$ with probability $1-\mathrm{e}^{-\Omega(n^{1/3})}$.

Proof. This proof follows the one of Theorem 5.10 (p. 67).

Recall that our modified version of the 1/5-rule ensures $\sigma = \Theta(\sqrt{\ln\lambda}\cdot d/n)$ for any polynomial number of steps (at least with probability $1-\mathrm{e}^{-\Omega(n^{1/3})}$), in particular, for any number of steps that is $O(b\cdot n/\ln\lambda)$.

Let κ denote a constant, which will be chosen large enough later. Within $\kappa bn/\ln\lambda$ steps, in each of which $\sigma = \Theta(\sqrt{\ln\lambda}\cdot d/n)$, each step reduces the approximation error by $\ln\lambda\cdot d/n$ with

probability $\Omega(1)$. Thus, we expect $\Omega(\kappa b n)$ steps each of which actually reduces the approximation error by at least $\ln\lambda \cdot d/n$. Since $(1-\ln\lambda/n)^s \leq 1/2$ for n large enough for some s that is $O(n/\ln\lambda)$, after at most s such steps the approximation error is halved; and after $b \cdot s$ such steps the approximation is less than $d^{[0]}/2^b$. Now, by choosing κ large enough, the expected number of such steps is at least $2bs$, and by Chernoff's bound, the probability that less than $b \cdot s$ such steps actually occur within the $\kappa b n$ steps is $\mathrm{e}^{-\Omega(b\cdot s)}$, i. e. $\mathrm{e}^{-\Omega(b\cdot n/\ln\lambda)}$ which is $\mathrm{e}^{-\Omega(b\cdot n/\ln n)}$ because $\lambda = \mathrm{poly}(n)$.

All in all, we have shown that in $\kappa b n/\ln\lambda = O(bn/\ln\lambda)$ steps with probability $1-\mathrm{e}^{-\Omega(bn/\ln n)}$ the approximation error becomes smaller than $d^{[0]}/2^b$—under the assumption that the 1/5-rule works, i. e., that $\sigma = \Theta(d/n)$ in all $\kappa \cdot b \cdot n/\ln\lambda$ steps. Since this is the case with probability $1-\mathrm{e}^{-\Omega(n^{1/3})}$, also the total error probability is bounded by $\mathrm{e}^{-\Omega(n^{1/3})}$. □

As already noted above, this upper bound on the runtime shows the following:

Conclusion 5.17. For (1+λ) ESs the general lower bound from Theorem 4.11 (p. 42) is asymptotically sharp.

As our lower bound, namely that $\Omega(n/\ln(1+\lambda))$ steps are necessary to halve the approximation error with probability $1-\mathrm{e}^{-\Omega(n)}$, holds for any $(1^+,\lambda)$ ES and any $(1,\lambda)\sigma$SA-ES (which fit our framework), the upper bound for the modified 1/5-rule tells us: When observing the reduction of the approximation error on a SPHERE-like function obtained by any other $(1^+,\lambda)$ ES or $(1,\lambda)\sigma$SA-ES within a polynomial number of steps, then the (1+λ) ES using Gaussian mutations adapted by the modified 1/5-rule realizes such a reduction within a number of steps that is at most by a constant factor larger than the number of steps of other ES (at least with probability $1-\mathrm{e}^{-\Omega(n^{1/3})}$). To put it more concise:

Conclusion 5.18. For any given λ (which may depend on the dimensionality of the search space) no $(1^+,\lambda)$ ES and no $(1,\lambda)\sigma$SA-ES can minimize a SPHERE-like function "considerably" faster than the (1+λ) ES using Gaussian mutations adapted by the modified 1/5-rule (given a proper initialization of the mutation strength).

Naturally, one might ask whether our general lower bound is also sharp for $(1,\lambda)$ ES, i. e., whether there is a σ-adaptation mechanism that makes the $(1,\lambda)$ ES get along (at least for a SPHERE-like function) with a number of steps that is of the same order as for the (1+λ) ES with the modified 1/5-rule.

5.2.4 SPHERE-like Functions and the $(1,\lambda)$ ES with 1/5-Rule

Unfortunately, the modified 1/5-rule does not make (much) sense for the $(1,\lambda)$ ES. There would be a strong drift away from the optimum, similar to the situation with the original 1/5-rule and the (1,1) ES. The original 1/5-rule, however, does make sense for the $(1,\lambda)$ ES—at least when λ is "large enough" as we shall see. The case when $\lambda = \Omega(n^\varepsilon)$ is especially simple to tackle. Let "1/5-rule" denote the original version (as described in Section 5.1.1 (p. 58)) in the following.

Recall that the 1/5-rule is supposed to keep σ such that each mutation is successful with a probability of roughly 1/5. Now, assume the initialization is such that the success probability in the first mutation of the first phase is at least $\beta \in \mathbb{R}_{>0}$. Then—for $\lambda = \Omega(n^\varepsilon)$—with probability

$1-(1-\beta)^\lambda = 1-\mathrm{e}^{-\Omega(n^\varepsilon)}$ at least one of the λ mutants is closer to the optimum than its parent. Thus, the IF-condition in the $(1+\lambda)$ ES that makes it different from the $(1,\lambda)$ ES would evaluate to "true," and hence, in such a case there is no difference between the $(1,\lambda)$ ES and $(1+\lambda)$ ES. As our lower-bound result tells us that during an observation period (which lasts $\Theta(\lceil n/\lambda \rceil)$, i. e. $O(\lceil n^{1-\varepsilon} \rceil)$, steps) the approximation error is not halved with probability $1-\mathrm{e}^{-\Omega(n)}$, with this probability the success probabilities of all mutations are $\Omega(1)$ in the first phase. As a consequence, the approximation error is *not* monotone decreasing during the phase only with a probability that is bounded above by $O(\lceil n^{1-\varepsilon} \rceil) \cdot \mathrm{e}^{-\Omega(n^\varepsilon)}$, which is $\mathrm{e}^{-\Omega(n^\varepsilon)}$. In other words, if the elitist $(1+\lambda)$ ES was run (with the same initialization) rather than the $(1,\lambda)$ ES, with probability $1-\mathrm{e}^{-\Omega(n^\varepsilon)}$ the IF-condition that implements elitist selection would never evaluate to "false" in the phase. In less formal words, w. o. p. the mutations in a phase are such that there is no difference between the $(1,\lambda)$ ES and the $(1+\lambda)$ ES—given that $\lambda = \Omega(n^\varepsilon)$.

Since the probability that there is a step in which none of the λ mutants is better than the parent is $\mathrm{e}^{-\Omega(n^\varepsilon)}$ even for any polynomial number of steps, the results that we obtained for the $(1+\lambda)$ ES carry over for the $(1,\lambda)$ ES. Namely, the 1/5-rule works (cf. Corollary 5.11 (p. 68)):

Lemma 5.19. Let a $(1,\lambda)$ ES with $\lambda = \Omega(n^\varepsilon)$ for a constant $\varepsilon > 0$ minimize a SPHERE-like function in $\mathbb{R}^n$ using Gaussian mutations adapted by a 1/5-rule. If the initialization is such that $\sigma = \Theta(d/n)$, then the 1/5-rule maintains this property for an arbitrary polynomial number of steps with probability $1-\exp(-\Omega(n^{\min\{1/3,\varepsilon\}}))$. This is also true when considering the more general notion of a 1/5-rule as described in Corollary 5.9 (p. 67).

And also the upper-bound result carries over directly (cf. Theorem 5.13 (p. 69)):

Theorem 5.20. Let a $(1,\lambda)$ ES with $\lambda = \Omega(n^\varepsilon)$ for a constant $\varepsilon > 0$ minimize a SPHERE-like function in $\mathbb{R}^n$ using Gaussian mutations adapted by a 1/5-rule. If the initialization is such that $\sigma^{[0]} = \Theta(d^{[0]}/n)$, then the number of steps i until $d^{[i]} \leq d^{[0]}/2^{b(n)}$, where $b\colon \mathbb{N} \to \mathbb{N}$ such that $b = \mathrm{poly}(n)$, is $O(b_{(n)} \cdot n/\sqrt{\ln \lambda})$ with probability $1-\exp(-\Omega(n^{\min\{1/3,\varepsilon\}}))$.

So, if λ is so large that there is w. o. p. not a single step (within a polynomial number of steps) which results in an increase of the approximation error, then we can simply recycle the proofs for the $(1+\lambda)$ ES. Yet what about smaller λ? In fact, we can show that for any fixed implementation of the 1/5-rule there is a *constant* λ^* such that the $(1,\lambda^*)$ ES using Gaussian mutations adapted by this specific 1/5-rule results in an asymptotically optimal runtime (for a SPHERE-like function and given that initially $\sigma^{[0]} = \Theta(d^{[0]}/n)$, of course). To show this, we have to deal with the situation that steps do occur in which the approximation error increases. The first step in our analysis is to bound the maximum loss in approximation quality which a single step may cause.

Therefore consider an isotropic mutation $\boldsymbol{m}$ of length ℓ and recall the so-called signed distance $g \in [-\ell,\ell]$ of the mutant from the hyperplane that contains $\boldsymbol{c}$ and lies perpendicular to the line passing through $\boldsymbol{c}$ and $\boldsymbol{x}^*$. Note: We consider the case $g < 0$. Then (given that the length of the mutation vector $\boldsymbol{m}$ is ℓ) the mutant's distance from $\boldsymbol{x}^*$ is at most $\sqrt{(d-g)^2+\ell^2}$ (by applying Pythagoras using that the mutant's distance from the line passing through $\boldsymbol{c}$ and $\boldsymbol{x}^*$ is at most ℓ). Item 1 of Lemma 3.12 (p. 25) tells us that $\mathsf{P}\{G_\ell \leq -\ell/n^{1/3}\} = \mathrm{e}^{-\Omega(1/3)}$ (because of the symmetry of the random variable G_ℓ). Hence, with probability $1-\ \mathrm{e}^{-\Omega(n^{1/3})}$

$$\begin{aligned}\operatorname{dist}(\boldsymbol{c}+\boldsymbol{m},\boldsymbol{x}^*) &\leq \sqrt{(d+\ell/n^{1/3})^2+\ell^2} \\ &= \sqrt{d^2+2d\ell/n^{1/3}+\ell^2/n^{2/3}+\ell^2}.\end{aligned}$$

For $\ell := d/n^{1/3}$, we obtain

$$\begin{aligned}\operatorname{dist}(\boldsymbol{c}+\boldsymbol{m},\boldsymbol{x}^*) &\leq \sqrt{d^2+2d^2/n^{2/3}+d^2/n^{4/3}+d^2/n^{2/3}}\\ &= d\cdot\sqrt{1+3/n^{2/3}+1/n^{4/3}}\\ &\leq d\cdot\sqrt{(1+2/n^{2/3})^2}\\ &= d+2d/n^{2/3},\end{aligned}$$

and it is readily seen that for any ℓ which is $O(d/n^{1/3})$ (rather than exactly $d/n^{1/3}$) we obtain $d+O(d/n^{2/3})$ as an upper bound on the mutant's distance from $\boldsymbol{x}^*$. Thus, for any constant $\kappa_1 > 0$ there is a constant κ_2 such that

$$\mathsf{P}\big\{\operatorname{dist}(\boldsymbol{c}+\boldsymbol{m},\boldsymbol{x}^*) \geq d+\kappa_2\cdot d/n^{2/3} \mid |\boldsymbol{m}| \leq \kappa_1\cdot d/n^{1/3}\big)\big\} \;=\; \mathrm{e}^{-\Omega(n^{1/3})}.$$

In particular, for Gaussian mutations, if σ is such that $\mathsf{P}\big\{|\sigma\cdot\widetilde{\boldsymbol{m}}| = O(d/n^{1/3})\big\} = 1-\mathrm{e}^{-\Omega(n^{1/3})}$, then the absolute loss in approximation quality (the absolute increase in distance from $\boldsymbol{x}^*$) of a mutation is $O(d/n^{2/3})$ with probability $1-\mathrm{e}^{-\Omega(n^{1/3})}$.

Thus, when the mutation strength σ is $\Theta(d/n)$, we need $\mathsf{P}\big\{|\widetilde{\boldsymbol{m}}| \geq \varepsilon\cdot n^{2/3})\big\} = \mathrm{e}^{-\Omega(n^{1/3})}$ for any constant $\varepsilon > 0$ for our line of reasoning to work. Therefore, recall from Section 3.2 (p. 19) that $|\widetilde{\boldsymbol{m}}|$ is χ-distributed so that the density for a length of x equals $x^{n-1}\cdot\mathrm{e}^{-x^2/2}\cdot 2^{1-n/2}/\Gamma(n/2)$. The interesting part (namely the factors that depend on x) is $x^{n-1}\cdot\mathrm{e}^{-x^2/2} = \mathrm{e}^{(n-1)\ln x - x^2/2}$. When $x := \varepsilon\cdot n^{2/3}$ for some constant $\varepsilon > 0$, this is bounded above by $\mathrm{e}^{-\Omega(n^{4/3})}$, and so is the integral over the interval $[x,\infty)$. Altogether, we have shown the following: Given that $\sigma = \Theta(d/n)$, then $\operatorname{dist}(\boldsymbol{c}+\sigma\cdot\widetilde{\boldsymbol{m}},\boldsymbol{x}^*) = d+O(d/n^{2/3})$ with probability $1-\mathrm{e}^{-\Omega(n^{1/3})}$, i. e., there is a constant $\kappa > 0$ such that $\mathsf{P}\big\{\widetilde{\Delta}_{d,\sigma} \leq -\kappa\cdot d/n^{2/3})\big\} = \mathrm{e}^{-\Omega(n^{1/3})}$.

This upper bound on the loss which a single mutation (and, consequently, also the best of λ mutations) may yield, can now be used in an application of Hoeffding's bound to obtain the following result:

Lemma 5.21. Let the $(1,\lambda)$ ES using Gaussian mutations minimize a SPHERE-like function in $\mathbb{R}^n$. Consider a phase of $\Theta(n)$ steps in which σ is not changed. Let d denote the distance from $\boldsymbol{x}^*$ at the beginning of this phase and assume that $\sigma = \Theta(d/n)$. Then, if λ is large enough such that $\mathsf{E}[\widetilde{\Delta}_{\sigma,d}^{\langle\lambda:\lambda\rangle}] = \Omega(d/n)$, the total gain of this phase is $\Omega(d)$ with probability $1-\mathrm{e}^{-\Omega(n^{1/3})}$.

Proof. Assume that the total gain is smaller than $d/2$ (otherwise there is nothing to show). Let k denote the number of steps, i. e., $k = \Theta(n)$, and let $\Delta^{[1]},\ldots,\Delta^{[k]}$ denote the random variables which correspond to the spatial gains in the k steps. Due to our assumption, each of them stochastically dominates the random variable $\widetilde{\Delta}_{\sigma,d/2}$. So we let $\Delta_1,\ldots,\Delta_k$ denote k independent instances of $\widetilde{\Delta}_{\sigma,d/2}$ and define $S := \Delta_1+\cdots+\Delta_k$. Then the total gain $\Delta^{[1]}+\cdots+\Delta^{[k]}$ of the phase stochastically dominates the random variable S. Since $\mathsf{E}[\widetilde{\Delta}_{\sigma,d}^{\langle\lambda:\lambda\rangle}] = \Omega(d/n)$ by precondition, we have $\mathsf{E}[\widetilde{\Delta}_{\sigma,d/2}^{\langle\lambda:\lambda\rangle}] = \Omega(d/n)$, and hence, $\mathsf{E}[S] = \Omega(d)$. Using Hoeffding's bound, namely Theorem 2.3 (p. 13), we obtain

$$\mathsf{P}\{S \leq \mathsf{E}[S]/2\} \;\leq\; \exp\left(\frac{-2\cdot(\mathsf{E}[S]/2)^2}{k\cdot(b-a)^2}\right) \;=\; \mathrm{e}^{-\Omega(d^2/n)/(b-a)^2},$$

where $[a,b]$ is the range of the random variables Δ_i. We already know (from Lemma 4.5 (p. 35)) that we can choose $b := d/n^{2/3}$ because with probability $1-\mathrm{e}^{-\Omega(n^{1/3})}$ none of the $k\cdot\lambda$ mutations

yields a spatial gain towards $\boldsymbol{x}^*$ of more than $d/n^{2/3}$. For the (1+λ) ES, we could choose $a := 0$ because a step's gain is non-negative when using elitist selection. In the reasoning preceding this lemma, we have shown for the $(1,\lambda)$ ES that with probability $1-\mathrm{e}^{-\Omega(n^{1/3})}$ the gains are such that we can choose $b := -\kappa \cdot d/n^{2/3}$ with some constant $\kappa > 0$. Thus, $b-a = O(d/n^{2/3})$, and hence, the exponent $-\Omega(d^2/n)/(b-a)^2$ becomes $-\Omega(d^2/n)/O(d^2/n^{4/3})$, which is $-\Omega(n^{1/3})$. In other words, for a constant $\kappa > 0$

$$\mathsf{P}\left\{S \leq \mathsf{E}[S]/2 \mid -\kappa \cdot d/n^{2/3} \leq \Delta_1, \ldots, \Delta_k \leq d/n^{2/3}\right\} = \mathrm{e}^{-\Omega(n^{1/3})},$$

and moreover, we already know that the condition on the range of the Δ_is is met with probability $1-\mathrm{e}^{-\Omega(n^{1/3})}$. All in all, with probability $1-\mathrm{e}^{-\Omega(n^{1/3})}$ the total gain is at least $\mathsf{E}[S]/2 = \Omega(d)$. □

As one may already guess, if we can show that "λ is large enough such that in the first step $\mathsf{E}[\widetilde{\Delta}_{\sigma,d}^{\langle\lambda:\lambda\rangle}] = \Omega(d/n)$" can be replaced by "$\lambda \geq \lambda^*$ for some constant λ^* (which depends on the relative mutation strength σ/d)," then obtaining a bound on the runtime is straight forward (in the same way as we did for the (1+1) ES).

Therefore, recall the random variable $\widetilde{G}$ which corresponds to the signed distance of a Gaussian mutation from a fixed hyperplane. Due to the isotropy of a Gaussian mutation, $\widetilde{G}$ is symmetric, i. e., $-\widetilde{G} \sim \widetilde{G}$. Symmetric random variables bear the following property:

Proposition 5.22. Let the random variable X be symmetric, i. e., $\mathsf{P}\{X \geq g\} = \mathsf{P}\{X \leq -g\}$ for any $g \in \mathbb{R}$. Then $\mathsf{E}[X^{\langle 2:2\rangle}] \geq \mathsf{E}[X \cdot \mathbb{1}_{\{X\geq 0\}}]$ $(= \mathsf{E}[X \mid X \geq 0]/2)$.

Proof. Note that $\mathsf{P}\{X \geq 0\} = \mathsf{P}\{X \leq 0\} \geq 1/2$ due to the symmetry. As $X^{\langle 2:2\rangle} = \max\{X_1, X_2\}$, where X_1, X_2 are independent copies of X,

$$\begin{aligned}\mathsf{E}\left[X^{\langle 2:2\rangle}\right] \;=\; & \mathsf{E}\left[X^{\langle 2:2\rangle} \cdot \mathbb{1}_{\{X_1,X_2\geq 0\}}\right] + \mathsf{E}\left[X^{\langle 2:2\rangle} \cdot \mathbb{1}_{\{X_1\geq 0\geq X_2\}}\right] \\ & + \mathsf{E}\left[X^{\langle 2:2\rangle} \cdot \mathbb{1}_{\{X_1,X_2\leq 0\}}\right] + \mathsf{E}\left[X^{\langle 2:2\rangle} \cdot \mathbb{1}_{\{X_1\leq 0\leq X_2\}}\right].\end{aligned}$$

The first summand can be bounded from below as follows:

$$\begin{aligned}\mathsf{E}\left[X^{\langle 2:2\rangle} \cdot \mathbb{1}_{\{X_1,X_2\geq 0\}}\right] &\geq \mathsf{E}[X_1 \cdot \mathbb{1}_{\{X_1,X_2\geq 0\}}] \\ &= \mathsf{E}[X_1 \cdot \mathbb{1}_{\{X_1\geq 0\}}] \cdot \mathsf{P}\{X_2 \geq 0\} \\ &\geq \mathsf{E}[X_1 \cdot \mathbb{1}_{\{X_1\geq 0\}}] \cdot 1/2.\end{aligned}$$

Analogously, we obtain $\mathsf{E}\left[X^{\langle 2:2\rangle} \cdot \mathbb{1}_{\{X_1,X_2\leq 0\}}\right] \geq \mathsf{E}[X_1 \cdot \mathbb{1}_{\{X_1\leq 0\}}]/2$ and $\mathsf{E}\left[X^{\langle 2:2\rangle} \cdot \mathbb{1}_{\{X_i\geq 0\geq X_{3-i}\}}\right] \geq \mathsf{E}[X_i \cdot \mathbb{1}_{\{X_i\geq 0\}}]/2$ for $i \in \{1,2\}$. Altogether,

$$\mathsf{E}\left[X^{\langle 2:2\rangle}\right] \;\geq\; 3 \cdot \mathsf{E}[X \cdot \mathbb{1}_{\{X\geq 0\}}]/2 + \mathsf{E}[X \cdot \mathbb{1}_{\{X\leq 0\}}]/2 \;=\; \mathsf{E}[X \cdot \mathbb{1}_{\{X\geq 0\}}]$$

since $\mathsf{E}[X \cdot \mathbb{1}_{\{X\leq 0\}}] = -\mathsf{E}[X \cdot \mathbb{1}_{\{X\geq 0\}}]$ because of the symmetry $-X \sim X$. □

This implies the following: When the (1+1) ES and the (1,2) ES minimize the linear function $\textsc{Sum}_n$ (defined in Equation (3.1) on page 20) using plain Gaussian mutations (no σ-adaptation, i. e., σ fixed to one), then after i steps the expected distance of the evolving search point from the hyperplane given by the level set of the initial search point is at least as large for the (1,2) ES as it is for the (1+1) ES—for any number of steps i. Clearly, when we increase λ, the drift away from the hyperplane becomes stronger and stronger.

Therefore, recall that for any $g \in [0,\ell]$ (due to the symmetry) $\mathsf{P}\{G_\ell \geq g\} = \mathsf{P}\{G_\ell \leq -g\}$; let p denote this probability. For the random variable $G_\ell^{\langle\lambda:\lambda\rangle}$, however, $\mathsf{P}\{G_\ell^{\langle\lambda:\lambda\rangle} \geq g\} = 1-(1-p)^\lambda$ as apposed to $\mathsf{P}\{G_\ell^{\langle\lambda:\lambda\rangle} \leq -g\} = p^\lambda$.

In the following we prove "$1-(1-p)^\lambda \geq 3\cdot p^\lambda$ for $\lambda \geq 2$ and $p \in [0,1/2]$" and start off with $\lambda = 2$. Then

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

which holds since $2p \in [0,1]$.

> This shows that for $\lambda = 2$ a positive spatial gain of at least $g \geq 0$ is at least thrice as probable as a negative gain of at most $-g \leq 0$, for *any* $g \geq 0$. Interestingly, this implies the preceding proposition, so that we have found an alternative proof: $\mathsf{E}[G_\ell^{\langle 2:2\rangle}] = \mathsf{E}[G_\ell^{\langle 2:2\rangle^+}] + \mathsf{E}[G_\ell^{\langle 2:2\rangle^-}]$ and $\mathsf{E}[G_\ell^{\langle 2:2\rangle^+}] \geq -3\cdot\mathsf{E}[G_\ell^{\langle 2:2\rangle^-}]$, so that $\mathsf{E}[G_\ell^{\langle 2:2\rangle}] \geq \mathsf{E}[G_\ell^{\langle 2:2\rangle^+}]\cdot(3-1)/4 \geq \mathsf{E}[G_\ell^+]/2$.

For $\lambda \geq 3$, on the one hand $3p^\lambda = p^{\lambda-2}\cdot 3p^2$, and on the other hand

$$\begin{array}{rcl} 1-(1-p)^\lambda & = & (1-p)^{\lambda-2}\big((1-p)^{2-\lambda} - (1-p)^2\big) \\ & \geq & (1-p)^{\lambda-2}\big(\quad 1 \quad - (1-p)^2\big). \end{array}$$

Hence, we merely have to show that $(1-p)^{\lambda-2} \geq p^{\lambda-2}$, which in fact holds since $p \in [0,1/2]$ (so that $1-p \geq p$). Moreover, if $0 \leq p \leq 1/2-\varepsilon$ for a constant $\varepsilon > 0$, then for any constant κ, we can choose λ large enough such that

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

and consequently, $1-(1-p)^\lambda \geq 3\kappa\cdot p^\lambda$) in such a case.

Thus, if g is such that $\mathsf{P}\{G_\ell \geq g\} \leq 1/2-\varepsilon$, namely $g = \Omega(\ell/\sqrt{n})$, then $\mathsf{P}\{G_\ell^{\langle\lambda:\lambda\rangle} \geq g\} \geq 3\cdot\kappa\cdot\mathsf{P}\{G_\ell^{\langle\lambda:\lambda\rangle} \leq -g\}$ for λ large enough, where λ grows when the constant κ is increased as well as when the constant ε is decreased).

Now, note that the random variable $\Delta_{d,\ell}^{\langle\lambda:\lambda\rangle} + \ell^2/(2d)$ stochastically dominates $G_\ell^{\langle\lambda:\lambda\rangle}$ because Equation (4.2) on page 33 implies $\delta \geq g - \ell^2/(2d)$. Thus, if we choose λ large enough such that $\mathsf{E}[G_\ell^{\langle\lambda:\lambda\rangle}] \geq \ell^2/d$, then $\mathsf{E}[\Delta_{d,\ell}^{\langle\lambda:\lambda\rangle}] \geq \ell^2/(2d)$.

Recall that the 1/5-rule tries to adapt the length ℓ of the mutations such that $\mathsf{P}\{\Delta_{d,\ell} \geq 0\} = \mathsf{P}\{G_\ell \geq \ell^2/(2d)\} \approx 1/5$, which implies $\ell = \Theta(d/\sqrt{n})$, so that $\ell^2/(2d) = \Theta(d/n)$. By choosing λ a constant large enough, we can ensure—for ℓ such that $\ell^2/d = \Theta(d/n)$—that $\mathsf{P}\{G_\ell^{\langle\lambda:\lambda\rangle} \geq \ell^2/d\} = (1-\Omega(1))^\lambda \geq 1-\varepsilon$ for any constant $\varepsilon > 0$. As a consequence, $\mathsf{P}\{-\ell^2/d < G_\ell^{\langle\lambda:\lambda\rangle} < \ell^2/d\} \leq \varepsilon$, and thus, we have $\mathsf{E}[G_\ell^{\langle\lambda:\lambda\rangle}\cdot\mathbb{1}_{\{-\ell^2/d<G_\ell^{\langle\lambda:\lambda\rangle}<\ell^2/d\}}] \geq -\varepsilon\cdot\ell^2/d$. Since, as we have proved above, $\mathsf{E}[G_\ell^{\langle\lambda:\lambda\rangle}\cdot\mathbb{1}_{\{G_\ell^{\langle\lambda:\lambda\rangle}\geq\ell^2/d\}}] \geq -3\cdot\mathsf{E}[G_\ell^{\langle\lambda:\lambda\rangle}\cdot\mathbb{1}_{\{G_\ell^{\langle\lambda:\lambda\rangle}\leq-\ell^2/d\}}]$, we can indeed choose λ^* as a large enough constant such that $\mathsf{E}[G_\ell^{\langle\lambda^*:\lambda^*\rangle}] \geq \ell^2/d$. As we have seen, this implies $\mathsf{E}[\Delta_{\ell,d}^{\langle\lambda^*:\lambda^*\rangle}] \geq \ell^2/(2d) = \Theta(d/n)$. Summing up, we have obtained the following result:

Lemma 5.23. Let $\ell = \Theta(d/\sqrt{n})$. There is a constant λ^* such that $\mathsf{E}\big[\Delta_{\ell,d}^{\langle\lambda^*:\lambda^*\rangle}\big] = \Omega(d/n)$.

Recall that for a Gaussian mutation $\sqrt{n}/2 \leq |\widetilde{\boldsymbol{m}}| \leq 2\sqrt{n}$ with probability $1 - O(1/n)$ and that the tail of the underlying χ-distribution drops exponentially (cf. the reasoning preceding Lemma 5.21 (p. 75)). With this it is readily checked that the above lemma also holds for scaled Gaussian mutations:

Corollary 5.24. Let $\sigma = \Theta(d/n)$. Then there is a constant λ^* such that $\mathsf{E}\left[\widetilde{\Delta}_{\sigma,d}^{\langle\lambda^*:\lambda^*\rangle}\right] = \Omega(d/n)$.

The next step in our way to the analysis of the runtime of the $(1,\lambda)$ ES: We have to check that the 1/5-rule works. Therefore recall the proof of Theorem 5.8 (p. 65) in which we showed that the 1/5-rule works for the (1+1) ES. In particular, we have shown that σ remains bounded from below by $\Omega(d/n)$, where the actual constant hidden by the Ω-notation depends on the choice of the parameters of the 1/5-rule. As we have just shown, we can choose a constant λ^* large enough such that the expected one-step gain of the $(1,\lambda^*)$ ES is at least as large as the one of the (1+1) ES —just given that σ is (and remains) bounded by $\Theta(d/n)$. In particular, since λ^* is a constant, an observation period lasts $\Theta(n)$ steps—just as for the (1+1) ES with the more general 1/5-rule from Corollary 5.9 (p. 67). Finally, it is readily checked that all arguments carry over so that we obtain the following result:

Lemma 5.25. Given an implementation of a 1/5-rule according to Corollary 5.9 (p. 67), there exists a constant λ^* such that, when the $(1,\lambda^*)$ ES using Gaussian mutations adapted by this 1/5-rule minimizes a SPHERE-like function in $\mathbb{R}^n$, the following holds: Given that the initialization is such that $\sigma = \Theta(d/n)$, then the 1/5-rule maintains this property for an arbitrary polynomial number of steps with probability $1 - \mathrm{e}^{-\Omega(n^{1/3})}$.

Finally, also the proof of the runtime bound carries over and we obtain the following result.

Theorem 5.26. Let a $(1,\lambda^*)$ ES using Gaussian mutations adapted by a 1/5-rule minimize a SPHERE-like function in $\mathbb{R}^n$. Let $b\colon \mathbb{N} \to \mathbb{N}$ such that $b = \mathrm{poly}(n)$. Given that the constant λ^* is chosen large enough, if the initialization is such that $\sigma^{[0]} = \Theta(d^{[0]}/n)$, then the number of steps i until $d^{[i]} \leq d^{[0]}/2^{b(n)}$ is $O(b(n) \cdot n)$ with probability $1 - \mathrm{e}^{-\Omega(n^{1/3})}$.

Since λ^* is a constant, the number of f-evaluations is of the same order as the number steps. As a consequence, this upper bound asymptotically meets our lower bound from Theorem 4.11 (p. 42), i. e., the runtime is off by a factor which is bounded above by a constant. Here we see again the limits of asymptotic results: In practice, one would like to chose λ^* as small as possible, and thus, we are again at the point where fine-tuning the 1/5-rule does well make sense. Although it is possible (in principle) to calculate the smallest λ^* in dependence on the implementation of the 1/5-rule, we refrain from this calculation as it would not yield any new insights. Beyer (2001, p. 73) claims (based on the model-based progress-rate results) that "the largest progress rate per descendant can be attained at $\lambda = 5$." This means that—given perfect σ-adaptation—the expected one-step gain divided by λ is maximum for $\lambda = 5$, which indicates—yet does not directly imply—that the expected number of function evaluations to halve the distance from $\boldsymbol{x}^*$ is minimum for the (1,5) ES—under the assumption of perfect σ-adaptation. Experiments seem to show that a (1,8) ES seems to work even when using an implementation of the 1/5-rule which is not fine-tuned (like the one that doubles/halves σ).

$(1,\lambda)$ ES with λ a constant as small as possible are especially interesting with respect to the optimization in fitness landscapes with "cliffs" or "gaps" which must be overcome to enable a progression towards the optimum search point. This has been discussed by Jägersküpper and Storch (2006), yet it will not be discussed in this dissertation.

5.2.5 SPHERE-like Functions and the $(\mu+1)$ ES with 1/5-Rule

Recall the $(\mu+1)$ ES with 1/5-rule as described in Section 5.1.2 (p. 59). Our general lower bound from Theorem 4.14 (p. 46) tells us that the number of steps to halve the approximation error must grow linearly in the dimension of the search space (n) as well as in the size of the population (μ). Namely, less than $0.115\mu n$ steps suffice only with probability $\mathrm{e}^{-\Omega(n)}$, and hence, the question is whether the 1/5-rule makes the $(\mu+1)$ ES get along with $O(\mu n)$ steps. Naturally, one would guess that this should be the case. However, a proof seems to be non-trivial.

$P^{[i]}$ denotes the population *after* the ith step. Recall that a σ-adaptation takes place in step i in which an individual $\mathcal{X} = (\boldsymbol{x}, \sigma, b, g) \in P^{[i-1]}$ is selected for reproduction for which $b+g = 5n-1$, i. e., the *1/5-rule-count* of the chosen individual $\mathcal{X}$ must equal $5n-1$. Consider the first $\mu\sqrt{n}$ steps in a run of the $(\mu+1)$ ES with 1/5-rule. In these $\mu\sqrt{n}$ steps, the depth of each family tree (rooted at an initial individual) induced by the $(\mu+1)$ ES is at most $3\sqrt{n}$ with probability $1-\mathrm{e}^{-\Omega(\sqrt{n})}$ according to Theorem 4.13 (p. 44). Assume that there is an individual $\mathcal{X}$ in $P^{[\mu\sqrt{n}]}$ whose 1/5-rule-count is at least $4n$. Since $\mathcal{X}$'s lineage has a length of at most $3\sqrt{n}$ ancestors, from one ancestor to the next, the 1/5-rule-count increases on average by at least $4n/(3\sqrt{n}) = \sqrt{n}4/3$, respectively. Thus, there is at least one ancestor $\mathcal{Y}$ in $\mathcal{X}$'s lineage that has at least $\sqrt{n}4/3$ offspring. This implies that $\mathcal{Y}$ was selected for reproduction at least $\sqrt{n}4/3$ times. Since an individual is selected for reproduction only with probability $1/\mu$, the probability that $\mathcal{Y}$ is selected for reproduction at least $\sqrt{n}4/3$ times within $\mu\sqrt{n}$ steps is $\mathrm{e}^{-\Omega(\sqrt{n})}$ by Chernoff's bound (even despite the chance that $\mathcal{Y}$ may be removed from the population before it is mutated $\sqrt{n}4/3$ times at all). Thus, we have proved the following.

Lemma 5.27. Let the $(\mu+1)$ ES with 1/5-rule (observation period of $5n$ steps) optimize some function in $\mathbb{R}^n$. Then, with probability $1-\mathrm{e}^{-\Omega(\sqrt{n})}$, in the first $\mu\sqrt{n}$ steps there is no σ-adaptation.

In fact, our reasoning shows that between any two adaptations in a fixed lineage at least $\mu\sqrt{n}$ steps take place with probability $1-\mathrm{e}^{-\Omega(\sqrt{n})}$. Since a polynomial number of error probabilities each of which is $\mathrm{e}^{-\Omega(\sqrt{n})}$ results in a total error probability which is also bounded by $\mathrm{e}^{-\Omega(\sqrt{n})}$, we directly obtain

Corollary 5.28. Let the $(\mu+1)$ ES with 1/5-rule (observation period of $5n$ steps) optimize some function in $\mathbb{R}^n$. Consider the population after a polynomial number of steps. Then, with probability $1-\mathrm{e}^{-\Omega(\sqrt{n})}$, for each individual $\mathcal{X}$ in the population, between any two sequent σ-adaptations in the history of $\mathcal{X}$ at least $\mu\sqrt{n}$ steps take place.

(This does *not* imply that between *any* two sequent adaptations in a run of the $(\mu+1)$ ES at least $\mu\sqrt{n}$ steps take place—the two adaptations may affect different lineages.)

We concentrate on SPHERE-like functions in the following. Note—and keep in mind—this trivial but crucial observation: When the $(\mu+1)$ ES minimizes a SPHERE-like function, not only

the distance of the respectively best individual from $\boldsymbol{x}^*$ is non-increasing, also the distance of the respectively *worst* individual from $\boldsymbol{x}^*$ cannot increase.

For an individual $\mathcal{X} = (\boldsymbol{x}, \sigma, b, g)$ we let $|\mathcal{X}| := \operatorname{dist}(\boldsymbol{x}, \boldsymbol{x}^*)$. Given that there is no optimum in the population P, the *bandwidth of the population* is given by $\max_{\mathcal{X},\mathcal{Y}\in P} |\mathcal{X}| / |\mathcal{Y}|$. The population's bandwidth can be considered as one measure of the diversity of a population (in particular for SPHERE-like functions). We will now show that the population's diversity w. r. t. this measure collapses within a few steps of the $(\mu{+}1)$ ES.

Let $d^{[i]} := \min_{\mathcal{X}\in P^{[i]}} |\mathcal{X}|$ denote the population's distance from $\boldsymbol{x}^*$ after the ith step. Depending on their location in the search space w. r. t. the initial approximation error $d^{[0]} > 0$, the set of individuals is partitioned into four regions, where ε is an arbitrary small but positive constant:

$$\begin{aligned}
R_0 &:= \{\mathcal{X} \mid & & |\mathcal{X}| & < & (1-\varepsilon)d^{[0]}\}\\
R_1 &:= \{\mathcal{X} \mid (1-\varepsilon)d^{[0]} & \leq & |\mathcal{X}| & < & d^{[0]}\}\\
R_2 &:= \{\mathcal{X} \mid d^{[0]} & \leq & |\mathcal{X}| & < & (1-\varepsilon)^{-1}d^{[0]}\}\\
R_3 &:= \{\mathcal{X} \mid (1-\varepsilon)^{-1}d^{[0]} & \leq & |\mathcal{X}| & & \}
\end{aligned}$$

Hence, for the initial population, $P^{[0]} \subset R_2 \cup R_3$, i. e., there is neither a R_0-individual nor a R_1-individual in the initial population. We do not put any assumption on the bandwidth of the initial population. We know, however, that $P^{[0]} \cap R_2$ contains at least one individual, namely the individual that determines $d^{[0]}$, the population's initial distance from the optimum $\boldsymbol{x}^*$.

Lemma 5.29. Let a $(\mu{+}1)$ ES using Gaussian mutations adapted by the 1/5-rule minimize a SPHERE-like function. Assume that the initialization is such that $d^{[0]} > 0$ as well as for any $\mathcal{X} = (\boldsymbol{x}, \sigma, \ldots) \in P^{[0]}$: $\sigma = \Theta(|\mathcal{X}|/n)$.

Then, for any constant $\alpha \in (0, 1/2)$, after a number of steps i that is $O(n^\alpha \mu \log \mu)$, with probability $\mathrm{e}^{-\Omega(n^\alpha)}$, for all $\mathcal{X} \in P^{[i]}$: $(1-\varepsilon)d^{[0]} \leq |\mathcal{X}| < d^{[0]}$ for an arbitrary small constant $\varepsilon > 0$, i. e., the population's bandwidth has dropped below $1+\varepsilon$ $(> (1-\varepsilon)^{-1})$.

Proof. First of all note that the lower-bound result in Corollary 4.15 (p. 47) tells us that, for any number of steps i that is $o(\mu n)$, the probability that $P^{[\leq i]} := P^{[0]} \cup \cdots \cup P^{[i]}$ contains an individual from R_0 is $\mathrm{e}^{-\Omega(n)}$ and that, for the same reason, the probability that there is an individual in $P^{[\leq i]}$ that descends from an R_3-individual is also $\mathrm{e}^{-\Omega(n)}$.

Let $S^{[i]} := P^{[i]} \cap (R_0 \cup R_1)$ denote the subpopulation (after i steps) which contains exactly those individuals from $P^{[i]}$ with a distance of less than $d^{[0]}$ from $\boldsymbol{x}^*$. As already discussed above, with probability $1 - \mathrm{e}^{-\Omega(n)}$ for any number of steps that is $o(\mu n)$, the subpopulation S does never contain an R_0-individual nor an individual that descends from an R_3-individual. In the following we assume that this is the case—and keep in mind the error probability $\mathrm{e}^{-\Omega(n)}$ and that the number of steps must be bounded from above by $o(\mu n)$.

Assume for a moment that no σ-adaptation takes place. By definition of S, initially $S^{[0]}$ is empty, i. e., $\#S^{[0]} = 0$. Then we are interested in the number of steps i until $\#S^{[i]} = \mu$. Note that $\#S$ is non-decreasing because of the elitist selection. The expected number of steps until $\#S = 1$ is $O(\mu)$ since a mutation results with probability $\Omega(1)$ in a search point which is closer to $\boldsymbol{x}^*$ than its parent, and we pessimistically assume that the best individual (namely the one at distance $d^{[0]}$ from $\boldsymbol{x}^*$) must be selected for reproduction. (The other $\mu - 1$ individuals may be arbitrarily far away from $\boldsymbol{x}^*$.) Subsequently, whenever an individual from S is selected for reproduction,

then the mutation is successful with probability $\Omega(1)$ (since for any $\mathcal{X} \in S$ we have $\mathcal{X} \in R_1$ and moreover $\sigma = \Theta(d^{[0]}/n)$ because $\mathcal{X}$ cannot descend from an R_3-individual). In case of a success, $\#S$ increases (unless $\#S$ is already μ, of course). Thus, the expected number of steps until $\#S$ increases is $(\#S/\mu)/\Omega(1)$, i. e. $O(\#S/\mu)$. As a consequence, the expected number of steps i until $\#S^{[i]} = \mu$ is $O(\mu \log \mu)$. Assume that κ is a constant such that $\kappa\mu\log\mu$ is an upper bound on this expected number of steps (for n large enough). Then (using Markov's inequality) more than $2\kappa\mu\log\mu$ steps are necessary only with a probability of at most $1/2$. Thus, $n^\alpha 2\kappa\mu\log\mu$ steps are necessary only with a probability which is bounded from above by $2^{-n^\alpha} = \mathrm{e}^{-\Omega(n^\alpha)}$. Finally, since this number of steps is $o(\mu\sqrt{n})$, by the time when $\#S = \mu$ there has actually been no σ-adaptation with probability $1 - \mathrm{e}^{-\Omega(\sqrt{n})}$, cf. Lemma 5.27 (p. 79).

As all error probabilities are bounded by $\mathrm{e}^{-\Omega(n^\alpha)}$, respectively, the total error probability is also bounded by $\mathrm{e}^{-\Omega(n^\alpha)}$. □

In this proof we have implicitly shown a bound on the *takeover time*, which we define as the number of steps i until *all* individuals in $P^{[i]}$ are better (i. e. closer to $\boldsymbol{x}^*$) than *the best* individual in the initial population (where "initial" may also refer to a point in time when we (re)start our observation of the $(\mu{+}1)$ ES).[2] Namely, we have shown that this takeover time is $O(\mu\log\mu)$ in expectation, and that it is $O(n^\varepsilon\mu\log\mu)$ with probability $1 - \mathrm{e}^{-\Omega(n^\varepsilon)}$. In other words, after this time all initial individuals have been removed from the population. (Besides, this implies that, in our scenario, the number of offspring that an individual produces before it is removed from the population is $O(\log\mu)$ in expectation, and $O(n^\varepsilon\log\mu)$ w. o. p.)

Our assumption that no σ-adaptation takes place until (the first) takeover is sufficient for the proof. Taking a closer look at the proof, we see that it is merely necessary that each mutation results with probability $\Omega(1)$ in a search point that is closer to $\boldsymbol{x}^*$. And this is just what the 1/5-rule is supposed to do. Before we come to this point, however, we consider a hypothetically situation to become acquainted with what is going on in a run of the $(\mu{+}1)$ ES on a SPHERE-like function.

Proposition. Let a $(\mu{+}1)$ ES using Gaussian mutations minimize a SPHERE-like function in $\mathbb{R}^n$. Hypothetically assume that in each step the mutant is generated using the mutation strength d/n, where d denotes the parents distance from $\boldsymbol{x}^*$, implying that each mutation succeeds with an $\Omega(1)$-probability. Let ε denote an arbitrary small but positive constant.

Then, for any constant $\alpha \in (0, 1/2)$, after a number of steps that is $O(n^\alpha\mu\log\mu)$, with probability $1 - \mathrm{e}^{-\Omega(n^\alpha)}$, the population's bandwidth has dropped below $1+\varepsilon$.

Subsequently, the population's bandwidth remains bounded by $1+\varepsilon$ for any polynomial number of steps with probability $1 - \mathrm{e}^{-\Omega(n^\alpha)}$.

Proof. So, let us consider the situation after the first takeover and assume that this takeover happens in step t_1, i. e., $P^{[t_1]}$ is the first population containing none of the initial individuals. Recall that with probability $1 - \mathrm{e}^{-\Omega(n^\alpha)}$ we have $P^{[t_1]} \subset R_1$ (which we assume as a fact in the following), and that $d^{[t_1]}$ is $P^{[t_1]}$'s distance from $\boldsymbol{x}^*$. Now we redefine our four regions (our

[2] Our notion of "takeover time" differs slightly from the original one. Originally, the takeover time denotes the number of iterations of a loop in which solely selection is performed until the complete population consists of copies of the best individual, cf. Goldberg and Deb (1990).

partition of the set of individuals) as follows:

$$\begin{aligned}
R'_0 &:= \{\mathcal{X} \mid & & |\mathcal{X}| & < & (1-\varepsilon)d^{[t_1]}\} \\
R'_1 &:= \{\mathcal{X} \mid (1-\varepsilon)d^{[t_1]} & \leq & |\mathcal{X}| & < & d^{[t_1]}\} \\
R'_2 &:= \{\mathcal{X} \mid d^{[t_1]} & \leq & |\mathcal{X}| & < & (1-\varepsilon)^{-1}d^{[t_1]}\} \\
R'_3 &:= \{\mathcal{X} \mid (1-\varepsilon)^{-1}d^{[t_1]} & \leq & |\mathcal{X}| & & \}
\end{aligned}$$

(Note that $R'_2 \subseteq R_1 \cup R_2$.) Then $P^{[t_1]} \subset R'_2$, and by the very same reasoning as for the first takeover, the second takeover happens in step t_2 after another $O(\mu \log \mu)$ steps in expectation, and after another $O(n^\alpha \mu \log \mu)$ steps with probability $1 - \mathrm{e}^{-\Omega(n^\alpha)}$. Then $P^{[t_2]} \subset (R'_0 \cap R'_1)$, and in particular, $P^{[t_2]} \subset R'_1$ with probability $1 - \mathrm{e}^{-\Omega(n^\alpha)}$ as before. Now we could again redefine our four regions w. r. t. $d^{[t_2]}$ to investigate the third takeover, and so on.

Since the sum of a polynomial number of error probabilities each of which is $\mathrm{e}^{-\Omega(n^\alpha)}$ is bounded by $\mathrm{e}^{-\Omega(n^\alpha)}$, and since the bandwidth of $R_1 \cup R_2$ is $1/(1-\varepsilon)^2$, with probability $1 - \mathrm{e}^{-\Omega(n^\alpha)}$ the population's bandwidth does not exceed $(1-\varepsilon)^{-2}$ (which is smaller than $1+\varepsilon'$ for some constant $\varepsilon' > 0$) for any polynomial number of steps (subsequent to the very first takeover, of course). □

Consequently, the population's diversity, which collapses in the very first few steps, becomes steady-state w. r. t. a bandwidth very close to one. This means that the population moves somewhat homogeneously towards the optimum. And this is the reason why the runtime must grow linearly in the population size. In particular, just because it takes the $(\mu{+}1)$ ES $\Omega(\mu n)$ steps to halve the approximation error, the 1/5-rule should be able to update the mutation strengths frequently enough to keep them in the range that ensures success probabilities of $\Omega(1)$. It will be even easier to show the upper bound of $1/2 - \Omega(1)$ on the mutations' success probabilities, i. e., that the mutation strengths do not get too small, but remain bounded by $\Omega(|\mathcal{X}|/n)$.

Theorem 5.30. Let a $(\mu{+}1)$ ES using Gaussian mutations adapted by the 1/5-rule minimize a SPHERE-like function. Assume that the initialization is such that $d^{[0]} > 0$ as well as for any $\mathcal{X} = (\boldsymbol{x}, \sigma, \ldots) \in P^{[0]}$: $\sigma = \Theta(|\mathcal{X}|/n)$. Then, with probability $1 - \mathrm{e}^{-\Omega(n^{1/3})}$, for any $i = \mathrm{poly}(n)$, for any $\mathcal{X} = (\boldsymbol{x}, \sigma, \ldots) \in P^{[i]}$: $\sigma = \Theta(|\mathcal{X}|/n)$, i. e., the 1/5-rule keeps the mutation strengths such that any mutation is successful with a probability that is $\Omega(1)$ as well as bounded from above by $1/2 - \Omega(1)$.

Proof. We choose $\alpha := 0.4$ in the lemmas above. With probability $1 - \mathrm{e}^{-\Omega(n^{0.4})}$, by the time when the first adaptation happens there has already been the first takeover (i. e., the population's bandwidth collapsed already), so that all individuals have a distance of less than $d^{[0]}$ from $\boldsymbol{x}^*$ and no individual has an ancestor with a distance of $d^{[0]}/(1-\varepsilon)$ or more from $\boldsymbol{x}^*$. We assume this as a fact in the following (and keep in mind the error probability).

Assume that $\mathcal{X} = (\boldsymbol{x}, \sigma, g, b) \in P^{[i]}$ is chosen for reproduction with $b + g = 5n - 1$ (so that adaptation takes place) and that σ is doubled after $\mathcal{X}$ is mutated. Then the number of steps of the $(\mu{+}1)$ ES between this adaptation and the previous one (in $\mathcal{X}$'s lineage, of course) is larger than $\mu\sqrt{n}$ with probability $1 - \mathrm{e}^{-\Omega(\sqrt{n})}$ (Corollary 5.28 (p. 79)). Since a number of steps which is $O(n^{0.4}\mu \log \mu)$ is $o(\mu\sqrt{n})$, with probability $1 - \mathrm{e}^{-\Omega(n^{0.4})}$ there is a takeover between the two adaptations. Thus, with probability $1 - \mathrm{e}^{-\Omega(n^{0.4})}$, $\mathcal{X}$'s distance from $\boldsymbol{x}^*$ has become smaller between

the two sequent adaptations. Consequently, assuming that this is in fact the case (and again keeping in mind the error probability) the doubling of the mutation strength does result in a smaller success probability. Analogously to the proof of that the 1/5-rule works for the (1+1) ES (at least for SPHERE-like functions; Theorem 5.8 (p. 65)), this implies that there is a lower threshold of $\Omega(|\mathcal{X}|/n)$ on the mutation strengths, i. e., that the success probabilities remain bounded from above by $1/2 - \Omega(1)$—for any polynomial number of steps with probability $1 - \mathrm{e}^{-\Omega(n^{0.4})}$ (by the union bound).

It remains to show the $\Omega(1)$-threshold on the mutations' success probabilities, i. e., that the mutation strengths remain bounded by $O(|\mathcal{X}|/n)$. Therefore, assume that in step j the individual $\mathcal{X} = (\boldsymbol{x}, \sigma^*, g, b)$ with $g + b = 5n - 1$ is selected for reproduction. Assume that the previous adaptation of $\mathcal{X}$'s mutation strength took place in the ith step in the run of the (μ+1) ES. Note that during the "phase" from step i to step j, in all mutations of $\mathcal{X}$ the mutation strength σ^* is used. $\mathcal{X}$ has been chosen for reproduction overall $5n$ times in this phase, so that the number of mutations in the considered lineage is at most $5n$; let k denote the number of mutations in $\mathcal{X}$'s lineage in this phase. Then we have to show that if σ^* is large enough but $O(\left|\boldsymbol{x}^{[i]}\right|/n)$ (such that the first mutation in the phase succeeds with a small probability which is yet $\Omega(1)$), then not only σ^* is halved w. o. p., yet also $\left|\boldsymbol{x}^{[j]}\right| > \left|\boldsymbol{x}^{[i]}\right|/2$ w. o. p., so that the halving of the mutation strength actually results in an increase in the success probability of the mutations. This can be shown analogously to the proof of that the 1/5-rule works for the (1+1) ES (the proof of Theorem 5.8 (p. 65))—if we can deal with the following issue: $|\mathcal{X}|$ need not necessarily be non-increasing during the phase. However, because of the bound on the population's bandwidth, we know that after the ith step $|\mathcal{X}|$ can never rise above $\left|\boldsymbol{x}^{[i]}\right|/(1-\varepsilon)$. Let d^* denote the largest distance between $\mathcal{X}$ and $\boldsymbol{x}^*$ during the phase. Recall the reasoning (using Hoeffding's bound) that has finally lead to Equation (5.2) on page 60. Also here the total gain (along $\mathcal{X}$'s linage) is stochastically dominated by a random variable S, namely by the random variable S defined as the sum of k independent instances of the random variable $\widetilde{\Delta}^+_{d^*,\sigma^*}$; let those be denoted by $\Delta_1, \ldots, \Delta_k$. Then analogously to the derivation of Equation (5.2) on page 60 (using $k \leq 5n$), we obtain that, if σ^* is large such that $\mathsf{E}\left[\widetilde{\Delta}^+_{d^*,\sigma^*}\right] \leq d^*/(30n)$ ($\leq (d^*/6)/k$), then $\mathsf{P}\left\{S \geq d^*/3 \mid \Delta_1, \ldots, \Delta_k \leq d^*/n^{2/3}\right\} = \mathrm{e}^{-\Omega(n^{1/3})}$. Moreover, we already know that the condition "$\Delta_1, \ldots, \Delta_k \leq d^*/n^{2/3}$" is met with probability $1 - \mathrm{e}^{-\Omega(n^{1/3})}$. Since $d^* < \left|\boldsymbol{x}^{[i]}\right|/(1-\varepsilon)$, we obtain for ε small enough $d^*/3 < \left|\boldsymbol{x}^{[i]}\right|/2$, so that the probability that $\mathcal{X}$'s distance from $\boldsymbol{x}^*$ is halved in the phase (i. e., $\left|\boldsymbol{x}^{[j]}\right| \leq \left|\boldsymbol{x}^{[i]}\right|/2$) is bounded above by $\mathrm{e}^{-\Omega(n^{1/3})}$. In other words, with probability $1 - \mathrm{e}^{-\Omega(n^{1/3})}$ the halving of σ^* after the phase results in an increase in the success probability of a mutation of $\mathcal{X}$. Finally, $\sigma^* = \Theta(d^*/n)$ because an expected gain which is small enough but of order $\Theta(d^*/n)$ is used in the reasoning, and thus, the success probabilities of all $\mathcal{X}$-mutations within the phase are $\Omega(1)$. In particular, the smallest success probability in this phase—which determines the lower threshold we are aiming at—is $\Omega(1)$. □

And again, once we have shown that the 1/5-rule works, the upper-bound result is easy to obtain.

Theorem 5.31. Let a (μ+1) ES using Gaussian mutations adapted by the 1/5-rule minimize a SPHERE-like function. Assume that the initialization is such that $d^{[0]} > 0$ as well as for each initial individual $\mathcal{X} = (\boldsymbol{x}, \sigma, \ldots) \in P^{[0]}$: $\sigma = \Theta(|\mathcal{X}|/n)$. Then, with probability $1 - \mathrm{e}^{-\Omega(n^{1/3})}$, the number of steps i until $d^{[i]} \leq d^{[0]}/2^{b(n)}$ is $O(\mu \cdot b(n) \cdot n)$, where $b\colon \mathbb{N} \to \mathbb{N}$ such that $b = \mathrm{poly}(n)$.

Proof. In each step i the best individual in the population is selected for reproduction with probability $1/\mu$. Since the mutation strength is $\Theta(d^{[i]}/n)$—given that the 1/5-rule works—, the mutant is by at least $d^{[i]}/n$ closer to $\boldsymbol{x}^*$ than its parent with an $\Omega(1)$-probability, cf. Lemma 5.5 (p. 63). Thus, within $\kappa\mu n$ steps we expect $\Omega(\kappa n)$ steps each of which reduces the approximation error by an $1/n$-fraction. By choosing κ a constant large enough, using Chernoff's bound, with probability $1-\mathrm{e}^{-\Omega(n)}$ at least $0.7n$ of the $\kappa\mu n$ steps actually reduce the approximation error by an $1/n$-fraction, respectively. Finally, $(1-1/n)^{0.7n} < \mathrm{e}^{-\ln 2} = 1/2$, i. e., with probability $1-\mathrm{e}^{-\Omega(n)}$ the $\kappa\mu n$ steps suffice to halve the approximation error.

Since our bound on the error probability that the 1/5-rule works is $\mathrm{e}^{-\Omega(n^{1/3})}$, however, the total error probability (for any polynomial number of steps) is $\mathrm{e}^{-\Omega(n^{1/3})}$. □

The upper bound asymptotically meets the lower bound from Theorem 4.14 (p. 46). This tells us, on the one hand, that the 1/5-rule indeed makes the $(\mu$+1) ES get along with a number steps which is only by an $O(1)$-factor larger than the optimum number of steps (w. r. t. isotropic mutations, of course). On the other hand, this shows the following:

Conclusion 5.32. The general lower bound for $(\mu$+1) ESs from Theorem 4.14 (p. 46) is asymptotically sharp.

5.3 The (1+1) ES on Positive Definite Quadratic Forms

The 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) belongs to the class of positive definite quadratic forms which consists of all $f\colon \mathbb{R}^n \to \mathbb{R}$ with $\boldsymbol{x} \mapsto \boldsymbol{x}^\top \boldsymbol{Q}\boldsymbol{x}$, where the matrix $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ is positive definite, i. e., $f(\boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in \mathbb{R}^n \setminus \{\boldsymbol{o}\}$. Such a positive definite quadratic form (PDQF) induces an ellipsoidal fitness landscape and the minimum is located at the origin. Since the optimum function value is 0, the current approximation error is defined as $f(\boldsymbol{c})$, the f-value of the current individual. It will shortly become clear why this makes sense in this scenario. Even though we consider the approximation error w. r. t. the f-value from now on, the spatial gain of a mutation/step in the search space will still be of great importance to the analysis.

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 by 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 PDQF equals $\boldsymbol{x}^\top \boldsymbol{R}\boldsymbol{D}\boldsymbol{R}^{-1}\boldsymbol{x}$, and since $\boldsymbol{x}^\top \boldsymbol{R} = (\boldsymbol{R}^\top \boldsymbol{x})^\top$, the PDQF actually equals $(\boldsymbol{R}^\top \boldsymbol{x})^\top \boldsymbol{D}(\boldsymbol{R}^{-1}\boldsymbol{x})$. As $\boldsymbol{R}^\top = \boldsymbol{R}^{-1}$ for an orthogonal matrix, the PDQF 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, 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, cf. Lanczos (1956, p. 95).

[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."

Consequently, in the following we assume that $\boldsymbol{Q}$ is a diagonal matrix each entry of which is positive ($\boldsymbol{Q}$'s canonical 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.

Recall that for a given f-value of ϕ, the corresponding level set is $\{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) = \phi\}$, The *lower level set* is given by $\{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) < \phi\}$. The level set induced by SPHERE $= \phi^2$, for instance, 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 $M \subseteq \mathbb{R}^n \setminus \{\boldsymbol{0}\}$, the bandwidth of the set M equals $\sup_{\boldsymbol{x},\boldsymbol{y} \in M} \{|\boldsymbol{x}| / |\boldsymbol{y}|\}$. 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 $\xi_1 \geq \cdots \geq \xi_n$, we have $\min\{|\boldsymbol{x}| \mid \boldsymbol{x} \in E_{\phi^2}\} = \phi/\sqrt{\xi_1}$ 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 $\sqrt{\xi_1/\xi_n}$. (All level sets but the 0-level set, of course.) Note the relationship between this bandwidth and $\boldsymbol{Q}$'s condition number: The condition number equals the square of the bandwidth.

We may call the fitness landscape induced by a sequence $f_n \colon \mathbb{R}^n \to \mathbb{R}$ of PDQFs *close to being spherically symmetric* if the bandwidth (and with it the condition number) is $O(1)$ as n grows, more precisely, 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.

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 $\xi : \mathbb{N} \to \mathbb{R}_{\geq 1}$ such that $\xi = \text{poly}(n)$ as well as $\xi = \omega(1)$ as n grows:

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

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.

Before we look at this specific subclass of PDQFs with unbounded condition number, however, we investigate the complete class of PDQFs with bounded condition number.

5.3.1 Positive Definite Quadratic Forms with Bounded Condition Number

In this section we will formally prove that "slightly deforming" SPHERE does not affect the order of the algorithmic runtime of the (1+1) ES using Gaussian mutations adapted by the 1/5-rule. More important than this (maybe unsurprising) 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 which we will focus on later.

Therefore, let $f : \mathbb{R}^n \to \mathbb{R}$ denote a PDQF. Then, as we have already seen above, the level set $E_{\phi^2} = \{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) = \phi^2\}$ (with $\phi > 0$) forms a hyper-ellipsoid and has bandwidth $\sqrt{\xi_1/\xi_n}$. As we want to utilize our results for SPHERE, we need to know the maximum and the minimum curvature at points in E_{ϕ^2}. 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 (in the x_1-x_n-plane). 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

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

the maximum curvature of the plane curve I equals $\xi_1/(\sqrt{\xi_n} \cdot \phi)$ at the point $(0, \phi/\sqrt{\xi_n})$ in the x_1-x_n-plane, which has maximum distance from 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 $(\phi/\sqrt{\xi_1}, 0)$ in the x_1-x_n-plane, 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* $\boldsymbol{c} \in E_{\phi^2}$, there is a hyper-sphere $S^+ \ni \boldsymbol{c}$ with radius $r^+ = \phi \cdot \sqrt{\xi_1}/\xi_n$ such that the lower level set $E_{<\phi^2}$ lies completely inside this hyper-sphere 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^+. Moreover, it tells us that there is another hyper-sphere $S^- \ni \boldsymbol{c}$ with radius $r^- = \phi \cdot \sqrt{\xi_n}/\xi_1$ such that the open ball B^- whose missing boundary is S^- is a subset of the lower level set $E_{<\phi^2}$.

For PDQFs with level sets of bounded bandwidth, the radii of S^+ and S^- are of the same order, namely of order $\Theta(|\boldsymbol{c}|)$. This will be crucial in the following.

Now consider a mutant $\boldsymbol{c}' := \boldsymbol{c} + \boldsymbol{m}$. This mutant $\boldsymbol{c}'$ 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 $\boldsymbol{c}' \in E_{\leq\phi^2} := E_{\phi^2} \cup E_{<\phi^2}$. Recall that so far "$\Delta$" has denoted the random variable corresponding to a mutation's spatial gain towards a fixed point $\boldsymbol{x}^*$. Equivalently, Δ corresponds to the mutant's (random) signed distance from the hyper-sphere which is centered at $\boldsymbol{x}^*$ and contains the parent $\boldsymbol{c}$. As here the level-sets are no longer spherically symmetric (but ellipsoidal), these two perspectives are no longer consistent. Hence, in the following we let Δ denote the mutant's signed distance from its parent's level set (rather than the gain towards the center of the ellipsoid).

As we have just seen, $\boldsymbol{c}' \in E_{\leq\phi^2} \Rightarrow \boldsymbol{c}' \in B^+ \cup S^+$, so that we obtain

$$\mathsf{E}[\Delta_F \cdot \mathbb{1}_{\{f(\boldsymbol{c}') \leq f(\boldsymbol{c})\}}] = \mathsf{E}\left[\Delta_F \cdot \mathbb{1}_{\{\boldsymbol{c}' \in E_{\leq\phi^2}\}}\right] \leq \mathsf{E}[\Delta_F \cdot \mathbb{1}_{\{\boldsymbol{c}' \in B^+ \cup S^+\}}]$$

for the expected distance from $E_{>\phi^2} := \mathbb{R}^n \setminus E_{\leq\phi^2}$ after a step—for any isotropic distribution F over $\mathbb{R}^n$ according to which the mutation vector is sampled in a step of the (1+1) ES. In particular, we obtain that for a scaled Gaussian mutation, $\mathsf{E}[\widetilde{\Delta}^+_{\sigma,r^+}]$ is an upper bound on this expected spatial gain away from $E_{>\phi^2}$.

However, here we are interested in how fast the f-value reduces during a run of the (1+1) ES. We obtain an upper bound on the f-gain if we assume that the spatial gain is realized completely along the component with the heaviest weight ξ_1. Therefore, for an f-value of ϕ^2, we optimistically assume that the search were located at $\boldsymbol{c} = (\phi/\sqrt{\xi_1}, 0, \ldots, 0) \in \mathbb{R}^n$ and that the mutant were located at $\boldsymbol{c}' = (\phi/\sqrt{\xi_1} - \alpha \cdot r^+, 0, \ldots, 0) \in \mathbb{R}^n$ for some $\alpha \colon \mathbb{N} \to \mathbb{R}_{\geq 0}$; "$\alpha$" abbreviates "$\alpha(n)$."

Then, as $\alpha \cdot r^+ = \alpha \cdot \phi\sqrt{\xi_1}/\xi_n$, we obtain

$$\begin{aligned} f(\boldsymbol{c}') &= \xi_1 \cdot \left(\frac{\phi}{\sqrt{\xi_1}} - \alpha \cdot r^+\right)^2 = \xi_1 \cdot \phi^2 \cdot \left(\frac{1}{\xi_1} - \frac{2\alpha}{\xi_n} + \frac{\alpha^2 \cdot \xi_1}{{\xi_n}^2}\right) \\ &\geq \xi_1 \cdot \phi^2 \cdot \left(\frac{1}{\xi_1} - \frac{2\alpha}{\xi_n}\right) = \phi^2 \cdot \left(1 - \frac{2\alpha \cdot \xi_1}{\xi_n}\right) \\ &= f(\boldsymbol{c}) \cdot \left(1 - \frac{2\alpha\xi_1}{\xi_n}\right). \end{aligned} \tag{5.4}$$

Obviously, this upper bound on the f-gain of $f(\boldsymbol{c}) \cdot 2\alpha\xi_1/\xi_n$ is not very useful unless it is $o(f(\boldsymbol{c}))$, i. e., unless $\alpha \cdot \xi_1/\xi_n = o(1)$. One reason for this is that the maximum radius of curvature, which we have just used for the upper bound, is $r^+ = \phi \cdot \sqrt{\xi_1}/\xi_n$, whereas $\max\{|\boldsymbol{x}| \mid \boldsymbol{x} \in E_{\phi^2}\}$ is only $\phi/\sqrt{\xi_n}$, i. e., the diameter of S^+ is by a factor of $\sqrt{\xi_1/\xi_n}$ larger that the diameter of E_{ϕ^2}. (This factor equals the bandwidth of the level set E_{ϕ^2}.)

As we have seen before in Lemma 4.4 (p. 34), $\mathsf{E}[\widetilde{\Delta}^+_{\sigma,r^+}] \leq 0.52 \cdot r^+/(n-1)$ for any mutation strength σ. In fact, the lemma tells us that the expected distance from $E_{>\phi^2}$ is bounded above by $0.52 \cdot r^+/(n-1) = 0.52 \cdot (\phi\sqrt{\xi_1}/\xi_n)/(n-1)$ anyhow the distribution of $|\boldsymbol{m}|$ is chosen in a step of the (1+1) ES.

For PDQFs with bounded bandwidth/condition number we have (by definition) $\xi_1/\xi_n = O(1)$, so that substituting $\alpha := 0.52/(n-1)$ in Inequality (5.4) on page 87 results in an upper bound on a step's expected f-gain of $f(\boldsymbol{c}) \cdot (\xi_1/\xi_n) \cdot 1.04/(n-1) = O(f(\boldsymbol{c})/n)$—which is the same order as for SPHERE. Consequently, we obtain the same asymptotic lower bound on the expected runtime. This maybe rather unsurprising. Nevertheless, it is interesting that our lower bound is inversely proportional to the condition number—and not to the bandwidth, which intuition might tell us. (This might indicate that our lower bound is not necessarily as sharp as possible. For a bounded condition number, however, this does not make much of a difference.)

Theorem 5.33. Let a (1+1) ES using isotropic mutations minimize a positive definite quadratic form $f_n \colon \mathbb{R}^n \to \mathbb{R}$, $n \geq 4$, with bounded condition number $\mathfrak{C}_n$. Then the expected number of steps t until $f(\boldsymbol{c}^{[t]}) \leq f(\boldsymbol{c}^{[0]})/2$ is larger than $(n-1) \cdot 0.48/\mathfrak{C}_n = \Omega(n)$; the expected number of steps until $f(\boldsymbol{c}^{[t]}) \leq f(\boldsymbol{c}^{[0]})/2^{b(n)}$ is larger than $b(n) \cdot (n-1) \cdot 0.48/\mathfrak{C}_n - b(n) + 1 = \Omega(b(n) \cdot n)$, where $b \colon \mathbb{N} \to \mathbb{N}$.

Proof. Let "$f^{[i]}$" abbreviate "$f(\boldsymbol{c}^{[i]})$" and recall that $\mathfrak{C}_n = \xi_1/\xi_n$ for the condition number, where ξ_1 and ξ_n are the largest resp. the smallest eigenvalue associated with the PDQF to be minimized. For the application of Lemma 4.6 (p. 36), this time we let X_i denote the random variable corresponding to the f-gain in ith step. Due to the elitist selection, negative gains are always discarded. Consequently, the f-value will never exceed $f^{[0]}$ (the initial approximation error). As a further consequence, the X_i are bounded, namely $0 \leq X_i \leq f^{[0]}$.

Naturally, for the application of Lemma 4.6 (p. 36) we choose $g := f^{[0]}/2$ and note that the random variable S (as defined in the lemma) is a stopping time in our case. As we have just seen, $\mathsf{E}[X_i] \leq f^{[0]} \cdot \mathfrak{C} \cdot 1.02/(n-1) =: u$. Then the lower bound g/u (from Lemma 4.6 (p. 36)) on the expected number of steps which are necessary to halve the approximation error finally solves to $(f^{[0]}/2)\big/(f^{[0]} \cdot \mathfrak{C} \cdot 1.02/(n-1) \geq 0.48 \cdot (n-1)/\mathfrak{C}$.

Due to the linearity of expectation, the expected number of steps to halve the approximation error b times is lower bounded by $0.48 \cdot (n-1)/\mathfrak{C} + (b-1) \cdot (0.48 \cdot (n-1)/\mathfrak{C} - 1)$, where the rightmost "$-1$" emerges because the last step within a halving-phase coincides with the first step of the subsequent halving-phase. □

Now that we know that a (1+1) ES needs $\Omega(n)$ steps in expectation to halve the approximation error, naturally, we would like to obtain a lower bound which holds with an overwhelming probability. Before we come to this, however, note that there is an alternative, simpler way of proving an $\Omega(n)$-bound:

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 $f^{[0]}/5$ but smaller than $f^{[0]}/4$. By Markov's inequality, with a probability of at least $1/2$, the total gain of these k steps is smaller than $2 \cdot f^{[0]}/4$. 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 lower bound of $\Omega(b \cdot n)$ on the expected number of steps to halve the approximation error b times.

This proof is apparently simple. It results in worse lower bound, though. If we did an estimation for the constant $\kappa/2$, we would end up with a constant that is much smaller than 0.48. Nevertheless, not the bound, but its proof is useful: If we can show that the total gain of the k steps exceeds the double of its expectation not only with a probability which is bounded above by 1/2, but which is exponentially small, then we end up with a lower bound on the number of steps which holds with an overwhelming probability.

Therefore, the next step is to apply Hoeffding's bound to the total gain which a sequence of steps yields. Unfortunately, the random variables which correspond to the single-step gains are not independent—which has not been an issue above because of the linearity of expectation. However, also part of our best-case assumption is that in each step $\boldsymbol{c}$ is located at a point (in the respective level set) where the curvature is minimum (so that the radius of the hyper-sphere S^+ which we use in the estimate is maximum, which again results in maximum expected best-case gain). As the f-value is non-increasing, we thus obtain an upper bound (in the sense of stochastic dominance) on the total gain of k sequent steps by adding up the gain of k independent instances of the first step. Therefore, let $X_1, \ldots, X_k$ denote k independent instances of the random variable which corresponds to the best-case f-gain in the first step, and let $S := X_1 + \cdots + X_k$.

Now, if $0 \leq X_i \leq z > 0$, then Hoeffding (1963, Theorem 2) (cf. Theorem 2.3 (p. 13)) tells us that $\mathsf{P}\{S \geq \mathsf{E}[S] + x\} \leq \exp(-2 \cdot (x/z)^2/k)$ for $x > 0$. With $x := \mathsf{E}[S]$ this inequality becomes

$$\mathsf{P}\{S \geq 2 \cdot \mathsf{E}[S]\} \quad \leq \quad \exp\left(-2 \cdot (\mathsf{E}[S]/z)^2/k\right) \quad =: \quad p\,,$$

and hence, the probability that k steps suffice to halve the approximation error is not only bounded by $1/2$ (as Markov's inequality tells us) but also by p. Now, if we can show that $(\mathsf{E}[X]/z)^2 = \Omega(n^{1+\varepsilon})$ for some constant $\varepsilon > 0$, then p is bounded above by $\mathrm{e}^{-\Omega(n^\varepsilon)}$ since $k = \Theta(n)$, so that the reasoning used above (for the simple bound on the expected number of steps) yields that $b \cdot k = \Omega(b \cdot n)$ steps are necessary (to halve the approximation error $b = \mathrm{poly}(n)$ times) not only in expectation but also with probability $1 - \mathrm{e}^{-\Omega(n^\varepsilon)}$.

As shown in Lemma 4.5 (p. 35), $\mathsf{P}\{\Delta_{r^+,\ell} \geq r^+ \cdot n^{\varepsilon-1}\} = \mathrm{e}^{-\Omega(n^\varepsilon)}$ for any constant $\varepsilon \in (0,1)$ whatever the length ℓ. Thus, substituting $\alpha := n^{\varepsilon-1}$ in the estimation of $f(\boldsymbol{c}')$ in Inequality (5.4) on page 87 yields that a step's f-gain is smaller than $2 \cdot f(\boldsymbol{c}) \cdot \mathfrak{C} \cdot n^{\varepsilon-1} = O(f(\boldsymbol{c}) \cdot n^{\varepsilon-1})$ with probability $1 - \mathrm{e}^{-\Omega(n^\varepsilon)}$. Thus, when considering a polynomial number of steps, with probability $1 - \mathrm{e}^{-\Omega(n^\varepsilon)}$ in all these steps the f-gain is $O(f(\boldsymbol{c}) \cdot n^{\varepsilon-1})$, respectively. We assume that this is the case (and keep in mind the error probability of $\mathrm{e}^{-\Omega(n^\varepsilon)}$). Then we obtain

$$\left(\frac{\mathsf{E}[S]}{z}\right)^2 \geq \left(\frac{f^{[0]}/5}{2 \cdot f^{[0]} \cdot \mathfrak{C} \cdot n^{\varepsilon-1}}\right)^2 = \Omega(n^{2-2\varepsilon}),$$

so that $p = \mathrm{e}^{-\Omega(n^{2-2\varepsilon}/k)}$, i. e., $p = \mathrm{e}^{-\Omega(n^{1-2\varepsilon})}$ since $k = \Theta(n)$. Choosing $\varepsilon := 1/3$, we obtain $p = \mathrm{e}^{-\Omega(n^{1/3})}$. Since for this choice also our upper bound of $r^+ \cdot n^{-2/3}$ on the maximum single-step gain holds with probability $1 - \mathrm{e}^{-\Omega(n^{1/3})}$ (even for any polynomial number of steps), all in all the probability to halve the approximation error within the k steps is bounded above by $\mathrm{e}^{-\Omega(n^{1/3})}$. So we have proved the following:

Theorem 5.34. Let a (1+1) ES using isotropic mutations minimize a positive definite quadratic form $f_n : \mathbb{R}^n \to \mathbb{R}$ with bounded condition number $\mathfrak{C}_n$. Let $\boldsymbol{c}^{[t]}$ denote the evolving search point after t steps. Then the number of steps t until $f(\boldsymbol{c}^{[t]}) \leq f(\boldsymbol{c}^{[0]})/2^{b(n)}$ is $\Omega(b_{(n)} \cdot n)$ with probability $1 - \mathrm{e}^{-\Omega(n^{1/3})}$, where $b : \mathbb{N} \to \mathbb{N}$ such that $b = \mathrm{poly}(n)$.

In the preceding lower-bound proofs we implicitly assume optimal adaptation of the length of the isotropic mutations. Consequently, the concrete adaptation mechanism is irrelevant, and moreover, the arguments for halving the approximation error can simply be iterated[4] to obtain a lower bound on the number of steps which are necessary to reduce the approximation error to a predefined fraction. For an upper bound on the runtime, however, precisely these two aspects are the crucial points in an analysis.

We consider Gaussian mutations adapted by the 1/5-rule for the upper bound. Firstly, we have to check that the 1/5-rule still works.

Theorem 5.35. Let a (1+1) ES using Gaussian mutations adapted by a 1/5-rule minimize a PDQF with bounded bandwidth/condition number in $\mathbb{R}^n$. If the initialization is such that $\sigma = \Theta(|\boldsymbol{c}|/n)$, i. e., the success probability of the mutation in the first step is $\Omega(1)$ as well as $1/2 - \Omega(1)$, then with probability $1 - \mathrm{e}^{-\Omega(n^{1/3})}$ 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, of course. It implies that, for a given f-value of ϕ^2, either σ is $\Theta(|\boldsymbol{c}|/n)$ or it is not, independently 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 minimum/origin within its level set. In other words, for a given f-value of ϕ^2, either the mutation strength σ 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}.

[4] because of the linearity of expectation/the exponentially small error probability

For a fixed mutation strength σ, we let $p_{\boldsymbol{c}} := \mathsf{P}\{f(\boldsymbol{c}+\sigma\cdot\widetilde{\boldsymbol{m}}) \leq f(\boldsymbol{c})\}$ denote the success probability (of the step), and we let

$$p_{\boldsymbol{c}}^{\max} := \max_{\boldsymbol{x}\in E_{f(\boldsymbol{c})}} \mathsf{P}\{f(\boldsymbol{x}+\sigma\cdot\widetilde{\boldsymbol{m}}) \leq f(\boldsymbol{x})\} \quad \text{and} \quad p_{\boldsymbol{c}}^{\min} := \min_{\boldsymbol{x}\in E_{f(\boldsymbol{c})}} \mathsf{P}\{f(\boldsymbol{x}+\sigma\cdot\widetilde{\boldsymbol{m}}) \leq f(\boldsymbol{x})\}$$

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

During a phase in a run of the (1+1) ES the mutation strength σ 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 $p^{\max}$ as well as to $p^{\min}$ that was applied to the success probability "p" in the analysis for SPHERE-like functions. This reasoning from the proof of Theorem 5.8 (p. 65) will be shortly recapitulated in the following.

> We will show that (w. o. p. for an arbitrary polynomial number of steps) $p^{\min} = \Omega(1)$ on the one hand, 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 ith phase. Assume that the mutation strength σ is large such that $\varepsilon \geq p_{(i)}^{\max} = \Omega(1)$ for a constant ε (which we will choose appropriately small later) and n large enough. Since $p^{\max}$ is non-increasing and $p \leq p^{\max}$ 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's bound, w. o. p. less than a 2ε-fraction of the steps are successful so that the mutation strength σ is halved (we choose $2\varepsilon \leq 1/5$). This results 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 probabilities 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 σ (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 which we have shown. Namely, the spatial gain of a phase (of $O(n)$ steps) is w. o. p. 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.
>
> The upper threshold of $1/2-\Omega(1)$ on the steps' success probabilities is easer to show. Assume that the mutation strength σ 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 $p \geq p^{\min} \geq 0.3$ and since during a phase (in which σ 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's bound, w. o. p. more than 20% of the steps are actually successful so that σ 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 $p_{(j)}^{\min}$, the success probability in the first step of jth phase (because $p^{\min}$ is non-increasing during a phase). Then $p_{(j)}^{\max}$ is the upper threshold we are aiming at. To see that $p_{(j)}^{\max}$ 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 $p_{(j)}^{\min} = 1/2-\Omega(1)$ if in the last step of the jth 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).

Since all the error probabilities in our "w. o. p."-statements are bounded by $e^{-\Omega(n^{1/3})}$, altogether we have shown that with probability $1-e^{-\Omega(n^{1/3})}$ in each of an arbitrary polynomial number of steps σ is such that the success probability is $\Omega(1)$ as well as $1/2-\Omega(1)$. □

Now, having checked that the 1/5-rule also works for PDQFs with bounded condition number, we can show that the gain of a phase is large enough to obtain an upper bound on the runtime which asymptotically matches the more general (w. r. t. the adaptation of the mutation vectors' lengths) lower bound given in Theorem 5.34 (p. 89).

Theorem 5.36. Let a (1+1) ES using Gaussian mutations adapted by a 1/5-rule minimize a PDQF $f: \mathbb{R}^n \to \mathbb{R}$ with bounded bandwidth, i. e., the corresponding condition number is $O(1)$. If the initialization is such that $\sigma^{[0]} = \Theta(\left|\boldsymbol{c}^{[0]}\right|/n)$, then with probability $1-e^{-\Omega(n^{1/3})}$ the number of steps t until $f(\boldsymbol{c}^{[t]}) \leq f(\boldsymbol{c}^{[0]})/2^{b(n)}$ is $O(b_{(n)} \cdot n)$, where $b: \mathbb{N} \to \mathbb{N}$ such that $b = \text{poly}(n)$.

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 5.35 (p. 89) tells us that this also holds (at least w. o. p.) for an arbitrary polynomial number of steps. Thus, $\sigma = \Theta(|\boldsymbol{c}|/n)$ in all these steps, and we assume this as a fact (and keep in mind the error probability of $e^{-\Omega(n^{1/3})}$).

Analogously to the reasoning that precedes (and has led to) Inequality (5.4) on page 87, we have for the mutant $\boldsymbol{c}'$

$$f(\boldsymbol{c}') \leq f(\boldsymbol{c}) \iff \boldsymbol{c}' \in E_{\leq \phi^2} \Leftarrow \boldsymbol{c}' \in B^- \cup S^-,$$

so that we obtain

$$\mathsf{E}[\Delta_F \cdot \mathbb{1}_{\{f(\boldsymbol{c}') \leq f(\boldsymbol{c})\}}] = \mathsf{E}\left[\Delta_F \cdot \mathbb{1}_{\{\boldsymbol{c}' \in E_{\leq \phi^2}\}}\right] \geq \mathsf{E}[\Delta_F \cdot \mathbb{1}_{\{\boldsymbol{c}' \in B^- \cup S^-\}}]$$

for the expected distance from $E_{>\phi^2}$ for any isotropic distribution F over $\mathbb{R}^n$ according to which the mutation vector is sampled in a step of the (1+1) ES. In particular, for a scaled Gaussian mutation, $\mathsf{P}\left\{\widetilde{\Delta}_{\sigma,r^-} \geq r^-/n\right\} = \Omega(1)$ when $\sigma = \Theta(r^-/n)$ by Lemma 5.5 (p. 63). Since, for $f(\boldsymbol{c}) = \phi^2$, we have $r^- = \phi\sqrt{\xi_n}/\xi_1 = \Theta(|\boldsymbol{c}|)$, each step yields a spatial gain of at least $r^-/n = (\phi/n)\sqrt{\xi_n}/\xi_1$ with probability $\Omega(1)$.

Now, 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. Therefore, for an f-value of ϕ^2, we assume that the search were located at $\boldsymbol{c} = (0,\ldots,0,\phi/\sqrt{\xi_n}) \in \mathbb{R}^n$ and that the mutant were located at $\boldsymbol{c}' = (0,\ldots,0,\phi/\sqrt{\xi_n} - r^-/n) \in \mathbb{R}^n$. Then, as $r^- = \phi\sqrt{\xi_n}/\xi_1$,

$$\begin{aligned} f(\boldsymbol{c}') &= \xi_n \cdot \left(\frac{\phi}{\sqrt{\xi_n}} - \frac{\phi \cdot \sqrt{\xi_n}}{n \cdot \xi_1}\right)^2 = \xi_n \cdot \phi^2 \cdot \left(\frac{1}{\xi_n} - \frac{2}{n \cdot \xi_1} + \frac{\xi_n}{n^2 \cdot {\xi_1}^2}\right) \qquad (5.5) \\ &\leq \xi_n \cdot \phi^2 \cdot \left(\frac{1}{\xi_n} - \frac{1}{n \cdot \xi_1}\left(2 - \frac{\xi_n}{n \cdot \xi_1}\right)\right) \qquad \text{(note that } \xi_n/\xi_1 \leq 1 \text{ by definition)} \\ &\leq \phi^2 \cdot \left(1 - \frac{\xi_n}{n \cdot \xi_1}\right) = f(\boldsymbol{c}) \cdot \left(1 - \frac{1}{n \cdot \mathfrak{C}}\right) \end{aligned}$$

where $\mathfrak{C} = \xi_1/\xi_n$ is the condition number associated with the PDQF.

Thus, each step reduces the approximation error by an $\Omega(1/n)$-fraction with probability $\Omega(1)$. By Chernoff's bound, 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 $\Omega(n)$ with probability $1-\mathrm{e}^{-\Omega(n)}$. Consequently, with this probability, the f-value is reduced by a constant fraction within a phase (an observation period of the 1/5-rule). In particular, a constant number (which is nevertheless proportional to the condition number) of such 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 of the initial f-value.

As the error probability that the 1/5-rule fails is the (asymptotically) largest one, our reasoning holds for any polynomial number of steps with probability $1-\mathrm{e}^{-\Omega(n^{1/3})}$. □

Now that we have seen how and why the 1/5-rule works for PDQFs with bounded bandwidth, we are ready to consider PDQFs which result in ellipsoidal level sets with unbounded bandwidth. Up to now it has not been necessary to care about the actual location of the search point in it's respective level set. Note, however, that our lower bound is inversely proportional to the condition number, whereas our upper bound grows proportional to the condition number. And precisely the answer to the question where the trajectory of the evolving search point 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 1/5-rule minimizes an "ill-conditioned" PDQF with an unbounded condition number.

5.3.2 Positive Definite Quadratic Forms with Unbounded Condition Number

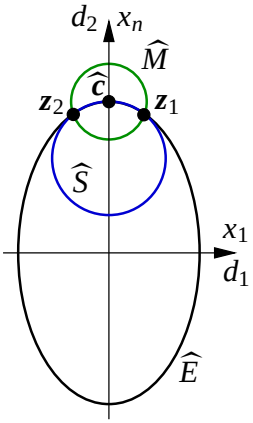
In this section, we concentrate on the (1+1) ES using Gaussian mutations adapted by the 1/5-rule, and we consider the following class of (sequences of) PDQFs, where $n \in 2\mathbb{N}$ and $\xi : \mathbb{N} \to \mathbb{R}_{>1}$ such that $\xi = \omega(1)$ as n grows ("ξ" abbreviates "$\xi(n)$" for better legibility):

$$f_n(\boldsymbol{x}) \quad := \quad \xi \cdot \left({x_1}^2 + \cdots + {x_{n/2}}^2\right) + {x_{n/2+1}}^2 + \cdots + {x_n}^2 \tag{5.6}$$

All results in this section will be obtained w. r. t. this scenario.

$f_n(\boldsymbol{x}) = \xi \cdot \textsc{Sphere}_{n/2}(\boldsymbol{y}) + \textsc{Sphere}_{n/2}(\boldsymbol{z})$ where $\boldsymbol{y} := (x_1, \ldots, x_{n/2})$ and $\boldsymbol{z} := (x_{n/2+1}, \ldots, x_n)$, and hence, the aim is to minimize the sum of two separate SPHERE-functions, one in $S_1 = \mathbb{R}^{n/2}$ and one in $S_2 = \mathbb{R}^{n/2}$, one of which has weight $\xi > 1$. For short: $f(\boldsymbol{x}) = \xi \cdot |\boldsymbol{y}|^2 + |\boldsymbol{z}|^2$.

Recall that for a scaled Gaussian mutation vector $\boldsymbol{m} \sim \sigma \cdot \widetilde{\boldsymbol{m}}$ each component of $\widetilde{\boldsymbol{m}}$ is independently standard-normally distributed. Thus, $\boldsymbol{m_1} := (m_1, \ldots, m_{n/2})$ and $\boldsymbol{m_2} := (m_{n/2+1}, \ldots, m_n)$ are two independent $(n/2)$-dimensional Gaussian mutations which are scaled by the same mutation strength σ. As $\boldsymbol{m_1}$ only affects $\boldsymbol{y}$, whereas $\boldsymbol{m_2}$ only affects $\boldsymbol{z}$, the f-value of the mutant is given by $\xi \cdot |\boldsymbol{y} + \boldsymbol{m_1}|^2 + |\boldsymbol{z} + \boldsymbol{m_2}|^2$. Though $\boldsymbol{m_1} \sim \boldsymbol{m_2}$, the changes caused by $\boldsymbol{m_1}$ are in a sense "more important" than the ones caused by $\boldsymbol{m_2}$ because of the weighting.



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 the search point is located at $(d_1, 0, \ldots, 0) \in S_1$ resp. $(0, \ldots, 0, d_2) \in S_2$. Or as Lanczos (1956, p. 95) puts it: "If in the general n-dimensional case m eigenvalues [which are associated with the PDQF] collapse into one, this means that in a certain m-dimensional 'subspace' spherical conditions prevail. Any m mutually orthogonal axes can be chosen within that subspace as principal axes of the quadratic surface." In other words, we may assume w. l. o. g. that the current search point is located at $(d_1, 0, \ldots, 0, d_2) \in \mathbb{R}^n$, i. e., that it lies in the x_1-x_n-plane. In fact, we have just described a projection $\widehat{\ } : \mathbb{R}^n \to \mathbb{R}^2$. Note that, due to the properties of the function class f_n and the isotropy of Gaussian mutations, this projection only conceals irrelevant information—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 be re-rotated).

Gain in a Single Step

In this section we have a closer look at the properties of a single mutation in our ellipsoidal fitness landscape. "f" will 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 mutation strength σ—and with it $\mathsf{E}[|\boldsymbol{m}|]$—interrelate when the success probability of a step (i. e. the probability that the mutant is accepted) is about 1/5. In other words: How does the length of the mutation vector depend on d_1 and on d_2, and how do d_1 and d_2 interrelate?

"Obviously," the heavier weighted $\text{SPHERE}_{n/2}$ in S_1 is minimized "first." Once the distance from the origin in S_1 becomes smaller and smaller, however, the changes in S_2 become more and more important. Finally, we "expect" some kind of equilibrium w. r. t. the interrelation of d_1 and d_2. Since $\nabla \widehat{f}(d_1, d_2) = (\xi\, 2 d_1, 2 d_2)^\top$, we know that for a search point which satisfies $d_1/d_2 = 1/\xi$ an infinitesimal change of d_1 has the same effect on the f-value as an infinitesimal change of d_2. Though the length of a mutation is not infinitesimal, this is an indicator that the ratio d_1/d_2 will stabilize when using isotropic mutations. And indeed, it will turn out that the process stabilizes w. r. t. $d_1/d_2 = \Theta(1/\xi)$.

In this section we shall see that in the region near the gentlest descent in our ellipsoidal fitness landscape, namely for $d_1/d_2 = O(1/\xi)$, a mutation succeeds with a probability that is $\Omega(1)$ as well as $1/2 - \Omega(1)$ if and only if $\sigma = \Theta((\sqrt{f}/n)/\xi)$, i. e., the mutation strength is inversely proportional to ξ. Furthermore, asymptotically tight bounds on the expected f-gain of a single step in such a situation will be obtained. Therefore, we will show that a mutation of a search point $\boldsymbol{c}$ for which $d_1/d_2 = O(1/\xi)$ with a mutation strength $\sigma = \Theta((\sqrt{f}/n)/\xi)$ in the ellipsoidal

fitness landscape is "similar" to the mutation of a search point $\boldsymbol{x}$ in the SPHERE scenario with $\text{SPHERE}(\boldsymbol{x}) = \Theta(f/\xi^2)$ (when using the same mutation strength).

We start our analysis at a point $\boldsymbol{c} \in \mathbb{R}^n$ with $\widehat{\boldsymbol{c}} = (0, \phi)$, i. e., $d_1 = 0$ and $d_2 = \phi$, so that $f = \phi^2$. That is, $\widehat{\boldsymbol{c}}$ is located at a point with gentlest descent (w. r. t. its level set, of course), and as a consequence, the curvature of the 2D-curve which is given by the projection $\widehat{E}$ of the n-ellipsoid $E_{\phi^2} = \{\boldsymbol{x} \in \mathbb{R}^n \mid f(\boldsymbol{x}) = \phi^2\}$, is maximum at $\widehat{\boldsymbol{c}}$.

> We show by a simple application of differential geometry as in Section 5.3.1 (p. 85) that the curvature at $\widehat{\boldsymbol{c}} \in \widehat{E}_{\phi^2}$ is $\Omega(\xi/\phi)$ if $\boldsymbol{c}$ lies in its level set such that $d_1 = O(d_2/\xi)$. Therefore, consider the ellipse given by $\xi \cdot d_1^2 + d_2^2 = \phi^2$. Thus, $d_2 = \sqrt{\phi^2 - \xi \cdot d_1^2}$ as a function of d_1, and furthermore,
>
> $$\frac{\mathrm{d}d_2}{\mathrm{d}d_1} = \frac{-\xi \cdot d_1}{\sqrt{\phi^2 - \xi \cdot d_1^2}} \quad \text{and} \quad = \frac{\mathrm{d}^2 d_2}{(\mathrm{d}d_1)^2} = \frac{-\xi^2 \cdot d_1^2}{\left(\phi^2 - \xi \cdot d_1^2\right)^{3/2}} + \frac{-\xi}{\sqrt{\phi^2 - \xi \cdot d_1^2}}.$$
>
> As the curvature (of a plane curve given by d_2 as a function of d_1) equals
>
> $$\frac{\dfrac{\mathrm{d}^2 d_2}{(\mathrm{d}d_1)^2}}{\left(1 + \left(\dfrac{\mathrm{d}d_2}{\mathrm{d}d_1}\right)^2\right)^{3/2}} = \frac{\phi^2 \xi}{\left(\phi^2 + (\xi^2 - \xi) \cdot d_1^2\right)^{3/2}},$$
>
> for $d_1 = \alpha \cdot \phi/\xi$ the curvature equals $\dfrac{\xi}{\phi \cdot \left(1 + (1 - 1/\xi) \cdot \alpha^2\right)^{3/2}}$.
>
> Finally, $\left(1 + (1 - 1/\xi) \cdot \alpha^2\right)^{3/2} = O(1)$ for $\alpha = O(1)$, namely for $d_1 = O(\phi/\xi)$. Furthermore, for $\alpha = 0$, i. e. for $d_1 = 0$, the curvature equals ξ/ϕ.

The curvature of the 2D-curve $\widehat{E}_{\phi^2}$ at $\widehat{\boldsymbol{c}} = (0, \phi)$ equals ξ/ϕ, and consequently, the radius of the osculating circle ($\widehat{S}$ in the figure on the preceding page) equals ϕ/ξ. As this circle $\widehat{S}$ actually lies in the x_1-x_n-plane, it is an 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 $\boldsymbol{c}$). In particular, $S \subset E_{\leq \phi^2}$ such that $S \cap E_{\phi^2} = \{\boldsymbol{c}\}$. Thus, the probability that a mutation hits inside S is a lower bound on $\mathsf{P}\{f' \leq f\}$. For the success probability of a scaled Gaussian mutation $\boldsymbol{m} \sim \sigma \cdot \widetilde{\boldsymbol{m}}$ we have [5]

$$\begin{aligned}
& \mathsf{P}\{f' \leq f\} \\
= \;& \mathsf{P}\{\boldsymbol{c} + \boldsymbol{m} \text{ lies inside } E\} \\
\geq \;& \mathsf{P}\{\boldsymbol{c} + \boldsymbol{m} \text{ lies inside } S\} \\
= \;& \mathsf{P}\{|\boldsymbol{x} + \boldsymbol{m}| \leq |\boldsymbol{x}| \text{ for some } \boldsymbol{x} \text{ with } |\boldsymbol{x}| = \text{radius of } S = \phi/\xi\} \\
= \;& \mathsf{P}\{\widetilde{\Delta}_{\phi/\xi,\sigma} \geq 0\}.
\end{aligned}$$

For an upper bound on the probability that a mutation hits inside E, consider an isotropic mutation with a length of $\ell < 2\phi$ (since for $\ell \geq 2\phi$, E lies inside M, so that the mutant is rejected by the elitist selection anyway). Let $M = \{\boldsymbol{x} \in \mathbb{R}^n \mid \text{dist}(\boldsymbol{c}, \boldsymbol{x}) = \ell\}$ denote the mutation sphere

[5] In fact, the (in)equalities hold for any isotropic mutation vector of a fixed length ℓ, i. e., if each of the probabilities is conditioned on the event $\{|\boldsymbol{m}| = \ell\}$. Since ℓ is arbitrary here and the radius of S is independent of ℓ, they are valid not only for scaled Gaussian mutations but for any isotropic mutation.

consisting of all potential mutants (at distance ℓ from $\boldsymbol{c}$). Then $\widehat{M}$ is a circle (cf. the figure on page 93) with radius ℓ centered at $\widehat{\boldsymbol{c}}$. Now consider the curvature at a point in $\widehat{E} \cap \widehat{M} = \{\boldsymbol{z}_1, \boldsymbol{z}_2\}$ (there are exactly two points of intersection since $0 < \ell < 2\phi$). As we have seen on page 94, the curvature at $\boldsymbol{z}_i$ is $\kappa_\ell = \Theta(\xi/\phi)$ if $\ell = O(\phi/\xi)$. Since the curvature at any point of the 2D-curve $\widehat{E}$ that lies inside $\widehat{M}$ is greater than κ_ℓ (since $\xi > 1$), $\widehat{\boldsymbol{c}}$ as well as $\boldsymbol{z}_i$ lie inside the osculating circle at $\boldsymbol{z}_{3-i}$ for $i \in \{1,2\}$. This osculating circle has radius $r_\ell := 1/\kappa_\ell$, and hence, we have $r_\ell = \Theta(\phi/\xi)$ for $\ell = O(\phi/\xi)$. Thus, there is also a circle with radius r_ℓ passing through $\widehat{\boldsymbol{c}}$ such that $\boldsymbol{z}_1$ and $\boldsymbol{z}_2$ lie inside this circle. (Consequently, the radius of the circle passing through $\boldsymbol{z}_1$, $\boldsymbol{z}_2$, and $\hat{\boldsymbol{c}}$ is 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,

$$\mathsf{P}\big\{ f' \le f \mid |\boldsymbol{m}| = \ell \big\} \quad \le \quad \mathsf{P}\big\{ \Delta_{r_\ell,\ell} \ge 0 \big\}$$

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

Recall that in the above reasoning we have assumed the current search point $\boldsymbol{c}$ to lie in the search space $\mathbb{R}^n$ such that $\widehat{\boldsymbol{c}} = (0,\phi) \in \mathbb{R}^2$, i. e., $d_1 = 0$ and $d_2 = \phi$. The estimates we have made to bound the probability that a mutation hits inside the n-ellipsoid E, however, remain valid as long as $d_1/d_2 = O(1/\xi)$ as we shall see: Since ξ/ϕ is the maximum curvature of $\widehat{E}$, there is always a circle $\widehat{S}$ with radius ϕ/ξ lying inside $\widehat{E}$ such that $\widehat{S} \cap \widehat{E} = \{\widehat{\boldsymbol{c}}\}$. And since $\widehat{S}$ is in fact an equator of an n-sphere S, we have $S \subset E_{\le \phi^2}$ 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. The result on the curvature (obtained on page 94) shows that as long as $d_1/d_2 = O(1/\xi)$ and $\ell = O(\phi/\xi)$, the curvature κ_ℓ is $O(\xi/\phi)$ (and $\kappa_\ell \ge \xi/\phi$ anyway).

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

$$\begin{aligned}
& \mathsf{P}\big\{ f' \le f \big\} \text{ is } \Omega(1) \text{ as well as } 1/2 - \Omega(1) \\
\iff \quad & \sigma = \Theta((\phi/\xi)/n) \\
\iff \quad & \mathsf{E}[|\sigma \cdot \widetilde{\boldsymbol{m}}|] = \Theta((\phi/\xi)/\sqrt{n}).
\end{aligned}$$

Thus, we are now going to investigate the gain of a step when $f = \phi^2$ and $\sigma = \Theta((\phi/\xi)/n)$. As we have seen above, there exists an n-sphere S with radius $r := \phi/\xi$ lying completely in E such that $S \cap E = \{\boldsymbol{c}\}$. Since in such a situation $\mathsf{P}\big\{\widetilde{\Delta}_{r,\sigma} \ge r/n\big\} = \Omega(1)$, with probability $\Omega(1)$ the mutant lies in $E_{\le \phi^2}$ such that its distance from $E_{>\phi^2}$ is at least r/n. If we pessimistically assume that this spatial gain were realized along the gentlest descent of f, namely that $d_1 = 0$ as well as $d_1' = 0$, so that $d_2' = d_2 - r/n = d_2 - (\phi/\xi)/n$, we obtain that with probability $\Omega(1)$

$$\begin{aligned}
f' &= \xi \cdot {d_1'}^2 + {d_2'}^2 \\
&\le \quad 0 \quad + (\phi - (\phi/\xi)/n)^2 \\
&= \phi^2 - 2\phi^2/(\xi n) + \phi^2/(\xi n)^2 \\
&= \phi^2 - \underbrace{(2 - 1/(\xi n))}\phi^2/(\xi n) \\
&\le \phi^2 - \qquad 1 \qquad \phi^2/(\xi n) \\
&= f - f/(\xi n).
\end{aligned}$$

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

$$\mathsf{E}\big[\, f - f'' \mid \sigma = \Theta((\sqrt{f}/n)/\xi)\big] \quad = \quad \Omega(f/(\xi n)).$$

Due to the pessimistic assumption on where in the fitness landscape the spatial gain is realized, this lower bound on the f-gain is valid only for $\sigma = \Theta((\sqrt{f}/n)/\xi)$), 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 $r/n = (\phi/\xi)/n$ could result in a much larger f-gain, of course. If $d_1/d_2 = O(1/\xi)$, however, the expected f-gain is $O(f/(\xi n))$ in the best case (w. r. t. the length of the mutation) as we shall see.

Therefore, let $d_1 = \varepsilon \cdot \phi/\xi$ with $\varepsilon = O(1)$ and still $f = \xi \cdot {d_1}^2 + {d_2}^2 = \phi^2$. Owing to the reasoning for the upper bound on $\mathsf{P}\big\{ f' \leq f \big\}$, we know that there is an n-sphere S with radius $r = \Theta(\phi/\xi)$ which contains $\boldsymbol{c}$ as well as $I := M \cap E_{\phi^2}$. The set I consists of all potential mutants that have the same f-value as $\boldsymbol{c}$ (namely ϕ^2), and I is the boundary of the hyper-spherical cap $C := M \cap E_{\leq\phi^2}$. Owing to the results for SPHERE-like functions, we know from Lemma 4.4 (p. 34) that $\mathsf{E}\big[\mathrm{dist}(\boldsymbol{c}', I) \cdot \mathbb{1}_{\{\boldsymbol{c}' \in C\}}\big] \leq 0.52\, r/(n-1) = O((\phi/\xi)/n)$ even for an isotropic mutation of optimum length. In other words, we know that, if an isotropic mutation hits $E_{\leq\phi^2}$, then its expected distance from $E_{>\phi^2}$ is $O((\phi/\xi)/n)$ whatever the length of this mutation. Thus, if we optimistically assume that the spatial gain were realized completely in S_1, i. e. completely on the ξ-weighted $\mathrm{SPHERE}_{n/2}$, (so that $d_2' = d_2$, implying that $d_2'' = d_2$), we obtain

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

This upper bound on the expected f-gain of a step holds for $d_1/d_2 = O(1/\xi)$ only, yet it holds 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 evolving search point is located in the region of the search space $\mathbb{R}^n$ which consists of all search points for which $d_1/d_2 = O(1/\xi)$. (Recall the initial guess that the search stabilizes in this region.)

Lemma 5.37. Consider the scenario that is described at the beginning of this Section 5.3.2 (p. 92).

If the current search point is located in the search space such that $d_1/d_2 = O(1/\xi)$, then $\mathsf{P}\{f' \leq f\}$ is $\Omega(1)$ as well as $1/2 - \Omega(1)$ if and only if $\sigma = \Theta((\sqrt{f}/n)/\xi)$.

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

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 5.3.1 for PDQFs with bounded bandwidth. Namely, if $d_1/d_2 = O(1/\xi)$ during a phase of $5n$ steps (an observation phase of the 1/5-rule) and $\sigma = \Theta((\sqrt{f}/n)/\xi)$, i. e., $\mathsf{P}\{f' \leq f\}$ is $\Omega(1)$ as well as $1/2 - \Omega(1)$, 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's bound, 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 a $\Theta(1/\xi)$-fraction in this phase. Then w. o. p. after $\Theta(\xi)$ consecutive phases the approximation error is halved—*if during all these phases the evolving search point is such that* $d_1/d_2 = O(1/\xi)$. Since, up to now, the arguments follow the ones for PDQFs with bounded condition number in Section 5.3.1 (p. 85), in particular the reasoning on the 1/5-rule can be adopted, and we directly obtain the following result:

Proposition 5.38. Consider the scenario as described at the beginning of Section 5.3.2 (p. 92).

Assume that $d_1^{[0]}/d_2^{[0]} = O(1/\xi)$ and $\sigma^{[0]} = \Theta((\left|\boldsymbol{c}^{[0]}\right|/n)/\xi)$ after initialization. If the course of the optimization is such that $d_1/d_2 = O(1/\xi)$ during the complete optimization process, then w. o. p. the number of steps to reduce the initial f-value/approximation error to a $2^{-b(n)}$-fraction is $\Theta(b(n) \cdot \xi \cdot n)$, where $b : \mathbb{N} \to \mathbb{N}$ such that $b = \text{poly}(n)$.

Obviously, the assumption/condition "$d_1/d_2 = O(1/\xi)$ during the complete optimization process" lacks any justification and is, therefore, objectionable. Rather we have to show that the stochastic process bears this property. Thus, the crucial point in the analysis is the question why should the ratio d_1/d_2 remain $O(1/\xi)$ (once this is the case). This crucial question will be tackled by a rigorous analysis in the remainder of this section.

In the following let $\Delta_1 := d_1 - d_1'$ and $\Delta_2 := d_2 - d_2'$ denote the spatial gain of the mutant towards the origin in S_1 resp. in S_2. Then d_1'/d_2' for the mutant is smaller than d_1/d_2 for its parent if and only if $\Delta_1/d_1 > \Delta_2/d_2$. Unfortunately, Δ_1 and Δ_2 correlate because $\boldsymbol{m_1}$ and $\boldsymbol{m_2}$ are adapted using the same mutation strength σ. Moreover, we must take selection into account because only certain combinations of Δ_1 and Δ_2 are accepted. To see which combinations are actually accepted, note that

$$f' \;=\; \xi \cdot (d_1 - \Delta_1)^2 + (d_2 - \Delta_2)^2 \;=\; \xi d_1^2 \underbrace{-\xi 2 d_1 \Delta_1 + \xi \Delta_1^2}_{} + d_2^2 \underbrace{-2 d_2 \Delta_2 + \Delta_2^2}_{} ,$$

and hence,

$$f' \leq f \quad\Longleftrightarrow\quad f' - f \;\leq\; 0 \quad\Longleftrightarrow\quad \overbrace{-\xi 2 d_1 \Delta_1 + \xi \Delta_1^2}\; \overbrace{-2 d_2 \Delta_2 + \Delta_2^2} \;\leq\; 0.$$

We assume $d_1, d_2 > 0$ in the following. Let α be defined by $\alpha/\xi = d_1/d_2$, i. e., α changes with the current search point $\boldsymbol{c}$ just like d_1 and d_2. Then the latter inequality is equivalent to

$$\begin{aligned}
& -2\alpha d_2 \Delta_1 + \xi \Delta_1^2 - 2 d_2 \Delta_2 + \Delta_2^2 \;\leq\; 0 \\
\Longleftrightarrow\quad & -\alpha \Delta_1 + \frac{\xi \Delta_1^2}{2 d_2} \;\leq\; \Delta_2 - \frac{\Delta_2^2}{2 d_2} \\
\Longleftrightarrow\quad & -\alpha \Delta_1 \left(1 - \frac{\Delta_1}{2 d_1}\right) \;\leq\; \Delta_2 \left(1 - \frac{\Delta_2}{2 d_2}\right) \quad \text{(using } d_2 = \xi \cdot d_1/\alpha\text{)}.
\end{aligned}$$

Thus, when using elitist selection, the mutant is accepted if and only if the last inequality holds. Whenever a mutation satisfying $-\alpha\Delta_1 > \Delta_2$ is accepted, then necessarily

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

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

> We assume for a moment that the selection mechanism was such that the mutant replaces (and becomes) the current individual if and only if $-\alpha\Delta_1 \leq \Delta_2$.

Let $i \in \{1,2\}$. As Δ_{3-i} is random, $\mathsf{E}[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leq \Delta_2\}}]$ is a random variable. For instance, the random variable $\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leq \Delta_2\}}]$ takes the value $\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leq x\}}]$ whenever the random variable Δ_2 happens to take the value x. We are interested in $\mathsf{E}\big[\mathsf{E}[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leq \Delta_2\}}]\big] = d_i - \mathsf{E}[d_i'']$, the expected reduction of the distance from the optimum in S_i in a step of the (1+1) ES. In particular, $\mathsf{E}[d_1'']\big/\mathsf{E}[d_2''] \leq d_1\big/d_2$ if and only if the expected relative gain in S_1 is at least as large as the one in S_2, i. e., if and only if

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

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

Therefore, note that

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

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

$$\begin{aligned} & \mathsf{E}\big[\mathsf{E}[\Delta_i \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leq \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_i \geq 0\}}] \cdot \mathbb{1}_{\{\Delta_{3-i} \geq 0\}}\big] \\ =\; & \mathsf{E}\big[\mathsf{E}[\Delta_i \qquad\qquad\quad \cdot \mathbb{1}_{\{\Delta_i \geq 0\}}] \cdot \mathbb{1}_{\{\Delta_{3-i} \geq 0\}}\big] \\ =\; & \mathsf{E}[\Delta_i \cdot \mathbb{1}_{\{\Delta_i \geq 0\}}] \cdot \mathsf{P}\{\Delta_{3-i} \geq 0\}. \end{aligned}$$

Thus, for the expected gain of a step in S_i we obtain

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

Since we are aiming at a lower bound on $\mathsf{E}\big[\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}}]\big]$, we may ignore the summand $\mathsf{E}\big[\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_1 \ge 0\}}] \cdot \mathbb{1}_{\{\Delta_2 < 0\}}\big]$ because it is non-negative anyway. Moreover, we may pessimistically assume that $\Delta_1 = -x/\alpha$ whenever Δ_2 happens to equal $x \ge 0$, which implies that

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

Since furthermore

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

we obtain the following lower bound for the expected gain of a step in S_1:

$$\begin{aligned} \mathsf{E}\big[\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}}]\big] \ \ge\ & \mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \ge 0\}}] \cdot \mathsf{P}\{\Delta_2 \ge 0\} \\ & -\mathsf{E}[\Delta_2 \cdot \mathbb{1}_{\{\Delta_2 \ge 0\}}] \cdot \mathsf{P}\{\Delta_1 < 0\}/\alpha\,. \end{aligned} \tag{5.7}$$

For the expected gain of a step in S_2, however, we will use the trivial upper bound

$$\mathsf{E}\big[\mathsf{E}[\Delta_2 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \le \Delta_2\}}]\big] \ \le\ \mathsf{E}[\Delta_2 \cdot \mathbb{1}_{\{\Delta_2 \ge 0\}}]. \tag{5.8}$$

With the help of these two bounds we can now prove 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 constant α^*.

Lemma 5.39. In the considered scenario, given that σ is such that $\mathsf{P}\{\Delta_1 \ge 0\}$ and $\mathsf{P}\{\Delta_2 \ge 0\}$ are $\Omega(1)$, there exists a constant α^* such that for $d_1/d_2 \ge \alpha^*/\xi$ yet $d_1/d_2 = o(1)$

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

for any constant κ for n large enough.

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

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

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

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

Using the lower/upper bound on the expected gain of a step in S_1 resp. S_2, namely Inequality (5.7) on page 99 and Inequality (5.8) on page 99, it is sufficient to show that

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

in such situations. Since $\mathsf{P}\{\Delta_1 \geq 0\}$ and $\mathsf{P}\{\Delta_2 \geq 0\}$ are $\Omega(1)$ (by precondition), $\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geq 0\}}]$ and $\mathsf{E}[\Delta_2 \cdot \mathbb{1}_{\{\Delta_2 \geq 0\}}]$ are of the same order, namely $\Theta(\mathsf{E}[|\boldsymbol{m}|]/\sqrt{n})$. Thus, we can choose a constant α^* such that the LHS of the inequality above—and with it $\mathsf{E}[\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{f' \leq f\}}]]$—is at least $\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geq 0\}}] \cdot \mathsf{P}\{\Delta_2 \geq 0\}/2$ for $\alpha \geq \alpha^*$ (and n large enough). Thus, for $\alpha \geq \alpha^*$ the LHS is $\Omega(\mathsf{E}[|\boldsymbol{m}|]/\sqrt{n})$, whereas the RHS is $o(\mathsf{E}[|\boldsymbol{m}|]/\sqrt{n})$ since $\kappa \cdot \alpha/\xi = o(1)$ due to the precondition that $\alpha = o(\xi)$. 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 $\alpha \geq \alpha^*$, then the expected relative gain (of the next step) towards the optimum in S_1 (on the ξ-weighted $\textsc{Sphere}_{n/2}$) 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 5.40. Let the mutation strength σ be fixed in the considered scenario. If in the ith step $\alpha^{[i]} \geq \alpha^*$ yet $\alpha^{[i]} = o(\xi)$ and $\mathsf{P}\{\Delta_1 \geq 0\}$ as well as $\mathsf{P}\{\Delta_2 \geq 0\}$ 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. $\alpha \leq \alpha^{[i]} + O(\alpha^{[i]}/n^{0.6})$ in all intermediate steps.

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

Concerning an upper bound on d_1, we have

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

and hence, during the k steps

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

Since $\psi(2-\psi)/\alpha^{[i]}$ is bounded by $O(n^{-0.6})$ just like ψ, we finally obtain that in all k steps

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

Now we are ready for the proof of the lemma's first claim. Therefore, assume that $\alpha \geq \alpha^{[i]} \geq \alpha^*$ for $n^{0.3} + 1$ steps. We will show that the probability of observing such a sequence of steps is

exponentially small. Therefore, note that, since w. o. p. $d_2 \geq d_2^{[i]}(1-\psi)$ as we have seen, this assumption implies that also w. o. p. $d_1 \geq d_1^{[i]}(1-\psi)$, i. e., w. o. p. $d_1 = d_1^{[i]} - O(d_1^{[i]}/n^{0.6})$ in all $n^{0.3}$ steps.

Let $X_j^{[k]}$, $j \in \{1,2\}$, denote the random variable $\Delta_j \cdot \mathbb{1}_{\{f' \leq f\}}$ in the $(i-1+k)$th step. (In particular, we have $\mathsf{E}[X_j] = \mathsf{E}[\mathsf{E}[\Delta_j \cdot \mathbb{1}_{\{f' \leq f\}}]]$.) Then, by choosing $\kappa = 2$ in Lemma 5.39 (p. 99), $\mathsf{E}[X_1^{[k]}]/d_1^{[k]} \geq 2 \cdot \mathsf{E}[X_2^{[k]}]/d_2^{[k]}$ for $1 \leq k \leq n^{0.3}$, i. e.,

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

For $j \in \{1,2\}$ let $T_j^{[k]} := X_j^{[1]} + \cdots + X_j^{[k]}$ denote the total gain of the k steps w. r. t. d_j. By linearity of expectation, $\mathsf{E}[T_1^{[k]}]/d_1^{[i]} \geq 2 \cdot \mathsf{E}[T_2^{[k]}]/d_2^{[i]}$ for $1 \leq k \leq n^{0.3}$; however, the goal is to show that $\mathsf{P}\{\, T_1^{[k]}/d_1^{[i]} \leq T_2^{[k]}/d_2^{[i]} \text{ for } 1 \leq k \leq n^{0.3} \,\}$ 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 $\alpha > \alpha^{[i]}$, 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 another search point $\check{\boldsymbol{x}}$ with $f(\check{\boldsymbol{x}}) = f(\boldsymbol{x})$ but $\check{\alpha} < \alpha$. Since this implies that $\check{d}_1 < d_1$ and $\check{d}_2 > d_2$, Proposition 4.3 (p. 33) tells us that $\check{\Delta}_1$ is stochastically dominated by Δ_1, whereas $\check{\Delta}_2$ stochastically dominates Δ_2. This implies that X_1 dominates $\check{X}_1$, whereas X_2 is dominated by $\check{X}_2$, and in particular, we have $\mathsf{E}[X_1] \leq \mathsf{E}[\check{X}_1]$ and $\mathsf{E}[X_2] \geq \mathsf{E}[\check{X}_2]$.

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, $\mathsf{E}[T_1^{[k]}]/d_1^{[i]} = 2 \cdot \mathsf{E}[T_2^{[k]}]/d_2^{[i]}$. Let T_j abbreviate $T_j^{[n^{0.3}]}$ for $j \in \{1,2\}$. Since $1.2/0.8 = 1.5 < 2$, it is sufficient to show that w. o. p. $T_1 \geq 0.8 \cdot \mathsf{E}[T_1]$ and that also w. o. p. $T_2 \leq 1.2 \cdot \mathsf{E}[T_2]$.

By Hoeffding's bound (cf. Theorem 2.3 (p. 13)), for $X_j^{[k]} \in [a_j, b_j]$ and $t_j > 0$,

$$\mathsf{P}\{T_1 - \mathsf{E}[T_1] \leq -t_1\} \;\leq\; \exp\left(\frac{-2 \cdot {t_1}^2}{n^{0.3} \cdot (b_1 - a_1)^2}\right) \quad \text{and}$$

$$\mathsf{P}\{T_2 - \mathsf{E}[T_2] \geq \;\; t_2\} \;\leq\; \exp\left(\frac{-2 \cdot {t_2}^2}{n^{0.3} \cdot (b_2 - a_2)^2}\right).$$

Choosing $t_j := 0.2 \cdot \mathsf{E}[T_j]$ for $j \in \{1,2\}$, each of the two exponents solves to

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

Thus, it remains to show that $\mathsf{E}[T_j]/(b_j - a_j) = \Omega(n^{0.2})$ because this would result in an exponent of $-\Omega(n^{-0.3} \cdot (n^{0.2})^2)$, which is $-\Omega(n^{0.1})$.

First we concentrate on $\mathsf{E}[T_1]/(b_1 - a_1)$. Since T_1 is the sum of $n^{0.3}$ random variables $X_1^{[k]}$, it suffices to show that $\mathsf{E}[X_1^{[k]}]/(b_1 - a_1) = \Omega(n^{-0.1})$ for $1 \leq k \leq n^{0.3}$. In the following we assume as a fact that $d_1 = d_1^{[i]} \pm O(d_1^{[i]}/n^{0.6})$ and $d_2 \in [\, d_2^{[i]} - O(d_2^{[i]}/n^{0.6}), d_2^{[i]}\,]$ since this happens w. o. p. (as we have already seen above in the proof of the lemma's second claim).

Recall that the mutation vector is split into two independent $(n/2)$-dimensional Gaussian mutations (one for S_1 and one for S_2) which are scaled by the same mutation strength σ. In particular, both mutation vectors have the same expected length; let $\bar{\ell}$ denote this expected length

and recall from Lemma 3.10 (p. 19) that $\bar{\ell} \asymp \sigma \cdot \sqrt{n/2}$. Owing to the results for SPHERE-like functions, we know that $\mathsf{P}\{\Delta_j \geq 0\} = \Omega(1)$ implies that $\sigma = O(d_j/n)$, i. e., $\bar{\ell} = O(d_j/\sqrt{n})$, and that, under these conditions, w. o. p. $|\Delta_j| = O(\bar{\ell}_j/n^{0.4})$. Also recall that $\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{f' \leq f\}}]$ is at least $\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geq 0\}}] \cdot \mathsf{P}\{\Delta_2 \geq 0\}/2$. Since $\mathsf{P}\{\Delta_2 \geq 0\} = \Omega(1)$ in ith step and $d_2 \geq d_2^{[i]}(1 - O(n^{-0.6}))$ in all $n^{0.3}$ steps, in each of these steps $\mathsf{P}\{\Delta_2 \geq 0\} = \Omega(1)$. Hence, $\mathsf{E}[X_1] = \Omega(\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geq 0\}}])$ in each of the $n^{0.3}$ steps. Owing to the results for SPHERE-like functions, we know (since $\bar{\ell} = O(d_1/\sqrt{n})$ as we have seen) that $\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geq 0\}}] = \Theta(\bar{\ell}/\sqrt{n})$ so that $\mathsf{E}[X_1] = \Omega(\bar{\ell}/\sqrt{n})$.

Altogether, we have shown that $\mathsf{E}[T_1] = n^{0.3} \cdot \Omega(\bar{\ell}/\sqrt{n}) = \Omega(\bar{\ell}/n^{0.2})$ and $b_1 - a_1 = O(\bar{\ell}/n^{0.4})$, implying $\mathsf{E}[T_1]/(b_1 - a_1) = \Omega(n^{0.2})$.

Concerning a lower bound on $\mathsf{E}[T_2]$, recall that $\mathsf{E}[T_1]/d_1^{[i]} = 2 \cdot \mathsf{E}[T_2]/d_2^{[i]}$. As a consequence, $\mathsf{E}[T_2] = \mathsf{E}[T_1] \cdot d_2^{[i]}/(2 \cdot d_1^{[i]}) = \Omega(n^{0.3} \cdot \bar{\ell}/\sqrt{n}) \cdot \Omega(\xi/\alpha^{[i]})$. Since $\alpha^{[i]} = O(\xi)$ (by precondition), we have $\mathsf{E}[T_2] = \Omega(\bar{\ell}/n^{0.2})$, and since $b_2 - a_2 = O(\bar{\ell}/n^{0.4})$ (cf. the reasoning for $b_1 - a_1$ above), $\mathsf{E}[T_2]/(b_2 - a_2) = \Omega(\bar{\ell}/n^{0.2})/O(\bar{\ell}/n^{0.4})$, which is also $\Omega(n^{0.2})$.

All in all, we have shown that $\mathsf{P}\{T_1 \leq 0.8 \cdot \mathsf{E}[T_1]\}$ as well as $\mathsf{P}\{T_2 \geq 1.2 \cdot \mathsf{E}[T_2]\}$ are bounded above by $\mathrm{e}^{-\Omega(n^{0.1})}$. Thus, our initial assumption that $\alpha \geq \alpha^{[i]} \geq \alpha^*$ for $n^{0.3} + 1$ steps implies that w. o. p. for the first $n^{0.3}$ steps $T_1/T_2 > \alpha^{[i]}/\xi$ (cf. above), i. e., that w. o. p. after at most $n^{0.3}$ steps α does drop below $\alpha^{[i]}$—a contradiction to our initial assumption. Thus, the sequence of steps we assumed to be observed happens only with an exponentially small probability. □

Since the 1/5-rule keeps the mutation strength unchanged for $5n$ steps, we can virtually partition each such observation phase in $5n/n^{0.3} = 5n^{0.7}$ sub-phases to each of which this lemma applies. Since $O(\alpha^{[i]}/n^{0.6}) \leq \alpha^{[i]}$ for n large enough, the preceding lemma shows the following:

When starting at a point $\boldsymbol{c}^{[0]}$ for which $\alpha^{[0]} = O(1)$, i. e., $d_1^{[0]}/d_2^{[0]} = O(1/\xi)$, then α remains smaller than $2 \cdot \max\{\alpha^{[0]}, \alpha^*\} = O(1)$ w. o. p. for any polynomial number of steps.

Incorporating these new insights into the reasoning for the 1/5-rule known from our analysis for SPHERE-like functions finally enables us to drop the objectionable assumption/condition "$d_1/d_2 = O(1/\xi)$ in the complete optimization process" in Proposition 5.38 (p. 97), so that we obtain the following result:

Theorem 5.41. Let the (1+1) ES using Gaussian mutations adapted by the 1/5-rule minimize the PDQF $f_n \colon \mathbb{R}^n \to \mathbb{R}$ given in Equation (5.6) on page 92.

Given that the initialization is such that $\sigma^{[0]} = \Theta(|\boldsymbol{c}^{[0]}|/(n\xi))$ and $d_1^{[0]}/d_2^{[0]} = O(1/\xi)$, then w. o. p. the number of steps to reduce the initial approximation error/f-value to a $2^{-b(n)}$-fraction is $\Theta(b_{(n)} \cdot \xi \cdot n)$, where $b \colon \mathbb{N} \to \mathbb{N}$ such that $b = \mathrm{poly}(n)$.

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 the initialization is such that $\alpha^{[0]}$ is considerably larger than α^*, however, we would like to know that there is a drift towards smaller α. And in fact, a closer look at the arguments in the proof of Lemma 5.40 (p. 100) reveals that the same arguments show that the drift towards smaller α is so strong when $\alpha \geq 2 \cdot \alpha^*$ that α drops w. o. p. by a constant fraction within at most n steps:

Proposition 5.42. Let the mutation strength σ be fixed in the considered scenario. If $\mathsf{P}\{\Delta_1 \geq 0\}$, $1/2 - \mathsf{P}\{\Delta_1 \geq 0\}$, $\mathsf{P}\{\Delta_2 \geq 0\}$ are $\Omega(1)$, then for n large enough: If in the ith step $\alpha^{[i]} \geq 2 \cdot \alpha^*$ yet $\alpha^{[i]} = o(\xi)$, then w. o. p. after at most n steps the search is located at a point with $\alpha \leq \alpha^{[i]} - \Omega(\alpha^{[i]})$.

Proof. Choosing $\kappa = 3$ in Lemma 5.39 (p. 99), we obtain that (at least for n large enough) $\xi \cdot \mathsf{E}\big[\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{f' \le f\}}]\big] \ge 3 \cdot \alpha \cdot \mathsf{E}\big[\mathsf{E}[\Delta_2 \cdot \mathbb{1}_{\{f' \le f\}}]\big]$. Assume that $\alpha^{[i]} \ge 2\alpha^*$ and $\alpha \ge \alpha^*$ 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 5.40 (p. 100)—except for T_j now being the sum of n instead of $n^{0.3}$ random variables—we obtain that w. o. p. $T_1/T_2 > 2 \cdot \alpha^{[i]}/\xi$, and hence, after these n steps w. o. p.

$$\begin{aligned}
\frac{d_1^{[i+n]}}{d_2^{[i+n]}} = \frac{d_1^{[i]} - T_1}{d_2^{[i]} - T_2} &< \frac{d_1^{[i]} - T_1}{d_2^{[i]} - T_1 \cdot \xi/(2 \cdot \alpha^{[i]})} \\
&= \frac{d_1^{[i]} - T_1}{d_1^{[i]} \cdot \xi/\alpha^{[i]} - T_1 \cdot \xi/(2 \cdot \alpha^{[i]})} \\
&= \frac{d_1^{[i]} - T_1}{d_1^{[i]} - T_1/2} \cdot \frac{\alpha^{[i]}}{\xi} \\
&= \left(1 - \frac{T_1/2}{d_1^{[i]} - T_1/2}\right) \cdot \frac{d_1^{[i]}}{d_2^{[i]}}.
\end{aligned}$$

Thus, we must finally show that T_1, the total gain of the n steps in S_1, is $\Omega(d_1^{[i]})$ w. o. p. Therefore, recall that T_1 is the sum of n random variables $X_1^{[k]}$ (namely $\Delta_1 \cdot \mathbb{1}_{\{f' \le f\}}$ in the $(i-1+k)$th step, respectively). In the following we consider a single step.

As shown in the proof of Lemma 5.40 (p. 100), $\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{f' \le f\}}] = \Omega(\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \ge 0\}}])$ due to the lemma's preconditions. Since $\mathsf{P}\{\Delta_1 \ge 0\}$ is $\Omega(1)$ as well as $1/2 - \Omega(1)$ (also by precondition), the mutation strength σ is such that $\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \ge 0\}}] = \Theta(d_1/n)$. All in all, the lemma's preconditions ensure that $\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{f' \le f\}}] = \Omega(d_1/n)$ in a step.

Hence, $\mathsf{E}[T_1] = n \cdot \Omega(d_1/n) = \Omega(d_1)$, and by applying Hoeffding's bound just like in the proof of Lemma 5.40 (p. 100), we finally obtain that T_1 is $\Omega(\mathsf{E}[T_1])$, i. e. $\Omega(d_1^{[i]})$, also w. o. p. □

This lemma shows that α drops very quickly—if the lemma's conditions are met. Utilizing the results that we obtained for SPHERE-like functions just as we have done in Section 5.3.1 (p. 85) for PDQFs with bounded bandwidth, it is readily checked that the condition "$\mathsf{P}\{\Delta_1 \ge 0\}$ and $1/2 - \mathsf{P}\{\Delta_1 \ge 0\}$ are $\Omega(1)$" is in fact ensured by the 1/5-rule for $d_1/d_2 \ge \alpha^*/\xi$ (recall that the case $d_1/d_2 = O(1/\xi)$ is covered by the arguments and proofs for PDQFs with bounded bandwidth in Section 5.3.1 (p. 85)). The two conditions "$\alpha = o(\xi)$" and "$\mathsf{P}\{\Delta_2 \ge 0\} = \Omega(1)$", however, originate from Lemma 5.39 (p. 99) 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 our proof rely on them? The answer is simple: In the very beginning of the reasoning we decided to focus on small α, namely on α that are $O(1)$. As a consequence, we decided on page 99 to disregard "$\Delta_2 < 0$." It appears neither in the lower bound on the expected gain in S_1 (namely Inequality (5.7) on page 99), nor in the upper bound on the expected gain in S_2 (namely Inequality (5.8) on page 99); neither in an indicator variable, nor in a probability. Yet in fact, for a fixed positive f-value and a fixed positive mutation strength, $\mathsf{P}\{\Delta_2 < 0\} \to 1$ as $\alpha \to \infty$, since the mutation of a search point with $d_2 = 0$ results in $d_2' = |\boldsymbol{m_2}| > 0$ with probability one.

Formally, we would show that $\mathsf{E}[\Delta_2 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leq \Delta_2\}}]$ actually becomes negative when α exceeds a particular α^{**}. For the lower bound on a step's expected gain in S_1, we would show that the term $\mathsf{E}\big[\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{-\alpha\Delta_1 \leq \Delta_2\}} \cdot \mathbb{1}_{\{\Delta_1 \geq 0\}}] \cdot \mathbb{1}_{\{\Delta_2 < 0\}}\big]$, which we decided to ignore on page 99, is actually $\Omega(\mathsf{E}[\Delta_1 \cdot \mathbb{1}_{\{\Delta_1 \geq 0\}}])$ for large α. However, since it is evident that the drift towards smaller α becomes larger and larger as α grows, we refrain from a full formal treatment.

5.3.3 Remarks

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 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 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 (because of a small mutation strength).

Naturally, the results carry over to functions that are translations (w. r. t. the search space $\mathbb{R}^n$) of a considered PDQF 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, we merely must consider the distance from the optimum point $\boldsymbol{x}^*$ in all arguments. 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 straightforward. 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.

Just like all other results in this chapter, also the result obtained for the (1+1) ES in the previous section is valid not only for Gaussian mutations (which are scaled by the mutation strength σ, which is deterministically adapted). We merely utilized that for a Gaussian mutation vector $\widetilde{\boldsymbol{m}}$ over $\mathbb{R}^n$ we have $\mathsf{P}\big\{|\widetilde{\boldsymbol{m}}| \in [\sqrt{n}/2, 2\sqrt{n}\big\} = 1 - O(1/n)$, cf. Lemma 3.10 (p. 19). In fact, all proofs carry over when substituting any isotropically distributed vector $\widetilde{\boldsymbol{m}}^*$ for $\widetilde{\boldsymbol{m}}$ that satisfies $\mathsf{P}\big\{|\widetilde{\boldsymbol{m}}^*| \in [a\sqrt{n}, b\sqrt{n}\big\} = 1 - o(1)$ (as n grows) for two positive constants a and b. (Note that under these conditions $\mathsf{E}[|\widetilde{\boldsymbol{m}}^*|]$ might not be finite.)

6 Conclusion and Outlook

Kenneth A. De Jong once asked me the question (w. r. t. a result that is not part of this dissertation) "So you proved the obvious?" And this question does make sense. There are at least two different answers: "Yes, I proved *the obvious*." and "Yes, I *proved* the obvious." The difference is—as it is often the case—the point of view. As discussed in the introduction, the dynamical-system approach has borne a bunch of results on the so-called progress rate for the SPHERE scenario, the expected spatial gain towards a fixed point in the search space. Despite the fact that they were obtained using the central/lateral component decomposition of the mutation vector (which we discussed in Section 3.4 (p. 28)) and the assumption that the lateral component would not deviate from its expectation, those results can be taken as a (more or less) strong indicator that the *expected* number of steps that a $(1\dot{+}\lambda)$ ES needs to halve the distance from the optimum is $\Omega(n/\ln(1+\lambda))$. Yet as we have seen in Section 4.2 (p. 35), formally concluding a lower bound on the expected number of steps from an upper bound on the expected one-step gain is anything but trivial. The aim of this work, however, was to prove lower (and upper) bounds on the number of steps/mutations. And in fact, we did prove a lower bound of $\Omega(n/\ln(1+\lambda))$—and this bound holds with an overwhelming probability of $1-\mathrm{e}^{-\Omega(n)}$. Such types of results can definitely be considered as not obvious—as they provide much deeper insight. Nevertheless, one may feel comfortable with strong indications, of course. The indications of the progress-rate results on the runtime of concrete ES, however, are not at all as strong as for the general lower bound. The reason is that they usually aim at the maximum possible progress. And obviously, an adaptation mechanism cannot ensure the optimal adaptation of the mutation strength in each step. Nevertheless, the result that the (1+1) ES using Gaussian mutations adapted by the 1/5-rule gets along with a linear number (in n) of steps to halve the approximation error when minimizing SPHERE may appear obvious—since each of thousands of simulations of this scenario has shown this behavior. Yet in fact, here we have proved why: The results presented in this dissertation prove that the parameters of the 1/5-rule can be varied in a large range without changing the order of steps, $O(n)$. Moreover, failures of the 1/5-rule in this scenario are virtually not observed because the stochastic process is such that the $O(n)$-bound holds with an overwhelming probability of $1-\mathrm{e}^{-\Omega(n^{1/3})}$. And again, this result can well be considered as not obvious.

Clearly, the 1/5-rule is not used in today's practical optimization with evolution strategies. Thus, the results obtained here are just a first starting point. On the other hand, we have proved why the 1/5-rule is not used in practice (anymore): For the very simple fitness landscapes induced by positive definite quadratic forms, the 1/5-rule makes the evolving search point move into the region close to the gentlest descent, which results in a small mutation strength and, finally, in a slow progression of the optimization. This has already been noted in experimental research, of course. With the *covariance matrix adaptation (CMA)*, Hansen and Ostermeier (1996) came up with an adaptation mechanism which is able to cope with ill-conditioned quadratic functions.

In fact, CMA with *cumulative step-length adaptation (CSA)* can be considered *the* state-of-the-art adaptation in evolution strategies. As progress-rate results for CMA/CSA-ES indicate, we can still not hope to analyze this very sophisticated adaptation in the same way as we did here for the 1/5-rule. However, a first step within reach may be the analysis of a $(1,\lambda)$ ES using a simplified version of CMA (and no CSA, but a 1/5-rule-like adaptation of the mutation strength). Another interesting extension would be to consider a $(\mu/\mu_I,\lambda)$ ES where in each step λ mutants are independently generated by adding a scaled Gaussian mutation to the centroid of the μ parent individuals. This allows for larger mutation strengths as well as for larger progress rates because of the so-called *genetic repair*, cf. Beyer (2001, Section 6.1.3.2). A runtime analysis of such an algorithm seems possible with the methods developed in this work. Various other modifications of the ESs that are covered by the results of this dissertation should also be analyzable.

Though we have to accept that (in evolutionary optimization) theory will not catch up with practice soon, we see that in this field there are lot of challenges and interesting questions to tackle with a probabilistic analysis. So let's catch up.

Bibliography

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

Arnold, D. (2002): *Noisy Optimization with Evolution Strategies*. Springer.

Auger, A. (2005): *Convergence results for the (1,λ)-SA-ES using the theory of ϕ-irreducible Markov chains*. Theoretical Computer Science, 334(1–3):35–69.

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

Beyer, H.-G., Meyer-Nieberg, S. (2005): *On the prediction of the solution quality in noisy optimization*. In *Foundations of Genetic Algorithms: 8th Int'l Workshop, Revised Selected Papers (FOGA)*, vol. 3469 of *LNCS*, 238–259, Springer.

Beyer, H.-G., Schwefel, H.-P. (2002): *Evolution strategies – a comprehensive introduction*. Natural Computing, 1:3–52.

Beyer, H.-G., Schwefel, H.-P., Wegener, I. (2002): *How to analyse evolutionary algorithms*. Theoretical Computer Science, 287(1):101–130.

Bienvenue, A., Francois, O. (2003): *Global convergence for evolution strategies in spherical problems: Some simple proofs and difficulties*. Theoretical Computer Science, 306(1–3):269–289.

de Bruijn, N. G. (1970): *Asymptotic Methods in Analysis*. North-Holland Publishing Company, Amsterdam, 3rd edn.

Droste, S., Jansen, T., Tinnefeld, K., Wegener, I. (2002a): *A new framework for the valuation of algorithms for black-box optimization*. In *Foundations of Genetic Algorithms 7 (FOGA 2002)*, 253–270, Morgan Kaufmann, San Francisco.

Droste, S., Jansen, T., Wegener, I. (1998): *On the optimization of unimodal functions with the (1+1) evolutionary algorithm*. In *Parallel Problem Solving from Nature – PPSN V*, vol. 1498 of *LNCS*, 13–22.

Droste, S., Jansen, T., Wegener, I. (2001): *Dynamic parameter control in simple evolutionary algorithms*. In *Foundations of Genetic Algorithms 6 (FOGA 2000)*, 275–294, Morgan Kaufmann, San Francisco.

Droste, S., Jansen, T., Wegener, I. (2002b): *On the analysis of the (1+1) evolutionary algorithm*. Theoretical Computer Science, 276(1–2):51–82.

Droste, S., Jansen, T., Wegener, I. (2006): *Upper and lower bounds for randomized search heuristics in black-box optimization*. Theory of Computing Systems, 39(4):525–544.

Fang, K.-T., Kotz, S., Ng, K.-W. (1990): *Symmetric multivariate and related distributions*, vol. 36 of *Monographs on statistics and applied probability*. Chapman & Hall, London.

Feller, W. (1971): *An Introduction to Probability Theory and Its Applications*, vol. 2. Wiley, 2nd edn.

Fogel, D. B. (editor) (1998): *Evolutionary Computation: The Fossil Record*. Wiley-IEEE Press.

Garnier, J., Kallel, L., Schoenauer, M. (1999): *Rigorous hitting times for binary mutations*. Evolutionary Computation, 7(2):173–203.

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

Goldberg, D. E., Deb, K. (1990): *A comparative analysis of selection schemes used in genetic algorithms*. In *Proc. 1st Workshop on Foundations of Genetic Algorithms (FOGA)*, 69–93.

Gradshteyn, I. S., Ryzhik, I. M. (1994): *Table of Integrals, Series, and Products*. Academic Press, San Diego, 5th edn.

Grinstead, C. M., Snell, J. L. (1997): *Introduction to Probability*. American Mathematical Society, 2nd edn.

Haagerup, U. (1982): *The best constants in the Khintchine inequality*. Studia Mathematika, 70:231–283.

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

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

Hofri, M. (1987): *Probabilistic Analysis of Algorithms*. Springer.

Jägersküpper, J., Storch, T. (2006): *How comma selection helps with the escape from local optima*. In *Proc. 9th Int'l Conference on Parallel Problem Solving From Nature (PPSN IX)*, vol. 4193 of *LNCS*, 52–61, Springer.

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

Lanczos, C. (1956): *Applied Analysis*. Dover Publications, New York, republication.

Mitzenmacher, M., Upfal, E. (2005): *Probability and Computing: Randomized Algorithms and Probabilistic Analysis*. Cambridge University Press.

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

Mühlenbein, H. (1992): *How genetic algorithmis really work: Mutation and hillclimbing*. In *Parallel Problem Solving from Nature 2 (PPSN)*, 15–25, North-Holland, Amsterdam.

Nemirovsky, A. S., Yudin, D. B. (1983): *Problem Complexity and Method Efficiency in Optimization*. Wiley, New York.

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

Rappl, G. (1989): *On linear convergence of a class of random search algorithms*. Zeitschrift für angewandte Mathematik und Mechanik (ZAMM), 69(1):37–45.

Rechenberg, I. (1965): *Cybernetic solution path of an experimental problem*. Royal Aircraft Establishment, in Fogel (1998).

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

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

Rudolph, G. (1997): *Convergence Properties of Evolutionary Algorithms*. Verlag Dr. Kovač, Hamburg.

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

Schwefel, H.-P. (1981): *Numerical Optimization of Computer Models*. Wiley, New York.

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

Teytaud, O., Gelly, S. (2006): *General lower bounds for evolutionary algorithms*. In *Proc. 9th Int'l Conference on Parallel Problem Solving From Nature (PPSN IX)*, vol. 4193 of *LNCS*, 21–31, Springer.

Teytaud, O., Gelly, S., Mary, J. (2006): *On the ultimate convergence rates for isotropic algorithms and the best choices among various forms of isotropy*. In *Proc. 9th Int'l Conference on Parallel Problem Solving From Nature (PPSN IX)*, vol. 4193 of *LNCS*, 32–41, Springer.

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

Wegener, I. (2003): *Towards a theory of randomized search heuristics*. In *Proc. 28th Int'l Symposium on Mathematical Foundations of Computer Science (MFCS)*, vol. 2747 of *LNCS*, 125–141, Springer.

Witt, C. (2005a): *Runtime analysis of the (μ+1) EA on simple pseudo-Boolean functions*. Evolutionary Computation, 14(1):65–86.

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