Self-Adaptation in Evolution Strategies

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

von
Silja Meyer-Nieberg

Dortmund
2007

Tag der mündlichen Prüfung:	5.12.2007
Dekan:	Professor Dr. Peter Buchholz
Gutachter:	Professor Dr. Hans-Georg Beyer
	Professor Dr. Günter Rudolph

Summary

In this thesis, an analysis of self-adaptative evolution strategies (ES) is provided. Evolution strategies are population-based search heuristics usually applied in continuous search spaces which ultilize the evolutionary principles of recombination, mutation, and selection. Self-Adaptation in evolution strategies usually aims at steering the mutation process. The mutation process depends on several parameters, most notably, on the mutation strength. In a sense, this parameter controls the spread of the population due to random mutation. The mutation strength has to be varied during the optimization process: A mutation strength that was advantageous in the beginning of the run, for instance, when the ES was far away from the optimizer, may become unsuitable when the ES is close to optimizer.

Self-Adaptation is one of the means applied to this end. In short, self-adaptation means that the adaptation of the mutation strength is left to the ES itself. The mutation strength becomes a part of an individual's genome and is also subject to recombination and mutation. Provided that the resulting offspring has a sufficiently "good" fitness, it is selected into the parent population.

Two types of evolution strategies are considered in this thesis: The $(1, \lambda)$-ES with one parent and λ offspring and $(\mu/\mu_I, \lambda)$-ES with a parental population with μ parents. The latter ES-type applies intermediate recombination in the creation of the offspring. Furthermore, the analysis is restricted to two types of fitness functions: the sphere model and ridge functions. The thesis uses a dynamic systems approach, the evolution equations first introduced by Hans-Georg Beyer, and analyzes the mean value dynamics of the ES.

Acknowledgements

First of all, I would like to thank my supervisor Hans-Georg Beyer for giving me the opportunity to do research in the area of evolution strategies. I am very grateful for the discussions, his advice and his steady support.

Many thanks to Stefan Pickl for offering me the opportunity to come to Munich and his advice.

Furthermore, I would like to thank Stefan Pickl and Andreas Karcher at the Universität der Bundeswehr – München for supporting the last phase of my research. 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 also gratefully acknowledged.

Many thanks to Steffen Finck, Heiko Hahn, Jens Jägersküpper, and Alexander Melkozerov who read the unfinished thesis.

A lot of thanks to the people at the LS11 at the Technical University of Dortmund and the Institut für Angewandte Systemwissenschaften und Wirtschaftsinformatik at Neubiberg.

Last, but not least, I would like to thank my parents and my whole family for the never faltering support.

List of Symbols and Abbreviations

$(1, \lambda)$-ES	ES with one parent, λ offspring
$(\mu/\mu_I, \lambda)$-ES	ES with μ parents, λ offspring using intermediate recombination
β	Parameter of the two-point distribution
λ	Offspring number
$\langle x \rangle$	Centroid or mean; usually of the parent population
$\mathcal{N}(\mu, \sigma^2)$	Normal (Gaussian) distribution with mean μ and variance σ
μ	Parent number
$\overline{\Delta Q^\star}$	Quality change normalized w.r.t N, i.e., $\overline{\Delta Q^\star} = \overline{\Delta Q} N$
$\overline{\Delta Q}$	Quality change. Expected change of the fitness during one generation. In the case of intermediate ES, the quality change gives the expected change of the fitness of the centroids.
$\Phi(x)$	Cumulative distribution function of standard normal distribution, i.e., $\mathcal{N}(0, 1)$
ρ	Mixing number: Number of recombinants
σ	Abbr. for $\langle \varsigma^{(g)} \rangle$
$\sigma^{(g)}$	Mutation strength
σ^*	Normalized mutation strength w.r.t. R and N, i.e., $\sigma^* = \sigma\ N/R$
$\sigma^\star$	Normalized mutation strength w.r.t. N, i.e., $\sigma^\star = \sigma N$
σ_ϵ	Noise strength: The standard deviation of the noise term in the standard noise model using a normally distributed random variable with zero mean
τ	Learning rate parameter of the log-normal distribution
$\varphi^{(k)}$	kth order progress rate
φ^*	Normalized progress rate w.r.t. R and N, i.e., $\varphi^* = \varphi\ N/R$

$\varphi^\star$	Normalized progress rate w.r.t. N, i.e., $\varphi^\star = \varphi N$
φ_R	Progress rate sphere model: progress towards the optimizer ridge functions: progress towards the axis
φ_x	Progress rate ridge functions: progress parallel to axis
$\varsigma^{(g)}$	Mutation strength
$C^k(U)$	Set of functions $f : U \to \mathbb{R}$ with f $k$times continuously differentiable and U an open subset of $\mathbb{R}^m$
g	Generation number
N	Search space dimensionality
R	Abbreviation for $r^{(g)}$
$R^{(g)}$	Sphere model: distance to the optimizer ridge functions: distance to the ridge in generation g
cdf	cumulative distribution function
CMA	Covariance matrix adaptation
CSA	Cumulative search step adaptation
pdf	probability density function, density function
Truncation ratio	Ratio of the parent and offspring number, i.e., $\mu : \lambda$
w.l.o.g.	Without loss of generality ...
w.r.t	With respect to ...

Contents

List of Figures

1 Introduction

Evolution strategies (ES) are one of the main variants of evolutionary algorithms (EA) invented in 1963 by Bienert, Rechenberg, and Schwefel at the Technical University Berlin. These population-based search heuristics move through the search space by means of variation, i.e., mutation and recombination, and selection. A population consists of several individuals. Each individual represents possible solution which is coded in the object parameters.

The performance of ES strongly depends on the choice of so-called strategy parameters. In ES, the strategy parameter equals usually the mutation strength. This parameter controls the spread of the population due to mutation. Sometimes the mutation strength is also referred to as the step-size in an analogy to classic optimization and numerics. During an optimization run, the mutation strength must be adapted continuously to allow the ES to travel with sufficient speed. To this end, several methods have been developed – e.g., Rechenberg's well-known $1/5$th-rule [81], self-adaptation [81, 88], or the cumulative step-size adaptation (CSA) and covariance matrix adaptation (CMA) of Ostermeier, Gawelczyk, and Hansen, e.g., [78, 53].

Following [23, p. 8], Figure 1.1 illustrates the basic mechanism of a multi-parent $(\mu/\rho, \lambda)$-ES with σ-self-adaptation. The self-adaptation mechanism will be introduced in more detail in the following chapter. In short, in a self-adaptive ES the tuning of the mutation strength(s) is left to the evolution strategy itself. Each individual has its own distinct set of strategy parameters. Similar to the object parameters, the strategy parameters are subject to variation. If an offspring is selected into the parent population, it also has a chance to bequest its strategy parameters to the offspring generation. That is, self-adaptation assumes a statistic/probabilistic connection between strategy parameters and "good" fitness values.

As Fig. 1.1 shows, a $(\mu/\rho, \lambda)$-ES maintains a population $\mathcal{P}_\mu^{(g)}$ of μ candidate solutions in generation g – with the strategy parameters used in their creation. Based on that parent population, λ offspring are created via variation. The variation process usually comprises recombination and mutation.

The offspring are created as follows: For each offspring, ρ of the μ parents are chosen for recombination leading to the set $\mathcal{P}_\rho$. The selection of the parents may be deterministic or probabilistic (see, e.g., [29, 43]).

First, the strategy parameters are changed. The strategy parameters of the chosen ρ parents are recombined and the result is mutated afterwards. The change of the object parameters occurs in the next step. Again, the parameters are first recombined and then mutated. The newly created strategy parameter σ_l is used in the mutation process. Afterwards, the fitness of the offspring is calculated.

After the offspring population of λ individuals is created, the μ-best individuals with respect to their fitness values are chosen as the next parental population $\mathcal{P}_\mu^{(g+1)}$. Two selection schemes are generally distinguished: "comma" and "plus"-selection. In the former case, selection is restricted to the offspring population. In the latter, members of old parent population and the offspring population may be selected into the succeeding parent population.

```
BEGIN
  g:=0;
  INITIALIZATION(P_μ^(0) := {(y_m^(0), σ_m^(0), F(y_m^(0)))});
  REPEAT
    FOR EACH OF THE λ OFFSPRING DO
      P_ρ:=REPRODUCTION(P_μ^(g))
      σ'_l :=RECOMB_σ(P_ρ);
      σ_l :=MUTATE_σ(σ'_l);
      y'_l :=RECOMB_y(P_ρ);
      y_l :=MUTATE_y(y'_l, σ_l);
      F_l := F(y_l);
    END
    P_λ^(g):={(y_l, σ_l, F_l)};
    CASE ","-SELECTION: P_μ^(g+1):=SELECT(P_λ^(g));
    CASE "+"-SELECTION: P_μ^(g+1):=SELECT(P_μ^(g),P_λ^(g));
    g:=g+1
  UNTIL stop;
END
```

Figure 1.1: The $(\mu/\rho, \lambda)$-σSA-ES (cf. [23, p. 8]).

The ES considered in this thesis are intermediate $(\mu/\mu_I, \lambda)$-ES with self-adaptation of a single mutation strength. The term *intermediate* denotes the manner of recombination. Using intermediate recombination for both, the object parameters and the mutation strengths, the offspring are generated according to:

1. Compute the mean $\langle\sigma\rangle = \frac{1}{\mu}\sum_{m=1}^{\mu}\sigma_m$ of the mutation strengths σ_m of the μ individuals of the parent population.
2. Compute the centroid $\langle\mathbf{y}\rangle = \frac{1}{\mu}\sum_{m=1}^{\mu}\mathbf{y}_m$ of the object vectors $\mathbf{y}_m$ of the μ individuals of the parent population.
3. For all offspring $l \in \{1, \ldots, \lambda\}$:
   (a) To derive the new mutation strength: Mutate the mean $\langle\sigma\rangle$ according to $\sigma_l = \langle\sigma\rangle\zeta$ where ζ is a random variable which should fulfill $\mathrm{E}[\zeta] \approx 1$ (see [29] for a discussion of this and further requirements). Typical choices of ζ's distribution include the log-normal distribution, derivatives of normal distributions, or a two-point distribution [16].
   (b) Generate the object vector $\mathbf{y}_l$ according to $y_i = \langle y_i\rangle + \sigma_l\mathcal{N}(0,1)$ where y_i is the vector's ith component and $\mathcal{N}(0,1)$ stands for a standard normally distributed random variable.

Afterwards, the μ best offspring are chosen – according to their fitness. They (along with their mutation strengths) become the parents of the next generation.

The thesis focuses on an analysis of the self-adaptation mechanism in two fitness environments: the sphere model and ridge functions. The first function class comprises functions $f_{sph} : \mathbb{R}^N \to \mathbb{R}$ of the form

$$f_{sph}(\mathbf{y}) \quad := \quad g(\|\mathbf{y} - \hat{\mathbf{y}}\|) \tag{1.1}$$

with $g : \mathbb{R} \to \mathbb{R}$ monotonically in- or decreasing and $\hat{\mathbf{y}} \in \mathbb{R}^N$ the optimizer of f_{sph}. The self-adaptive behavior of ES on the sphere model is addressed in Chapter 4. The second fitness environment are ridge functions $f_{rid} : \mathbb{R}^N \to \mathbb{R}$ given by

$$f_{rid}(\mathbf{y}) \quad := \quad y_1 - d\Big(\sqrt{\sum_{i=2}^{N} y_i^2}\Big)^{\alpha}. \tag{1.2}$$

The parameter α, $\alpha > 0$, denotes the degree of the ridge whereas d, $d > 0$, gives in a sense the "sharpness" of the isofitness lines of the ridge. The larger the value of d , the narrower the isofitness lines nestle to the axis. Ridge functions are described in more detail in Chapter 5.

The thesis is organized as follows: In Chapter 2, an overview over the state of the present research in self-adaptation is given. The focus is entirely on mutative self-adaptation. Therefore, the extensive work for ES using the $1/5$th rule as, e.g., [62, 85] or the cumulative step-size adaptation, e.g., [9, 5, 10] is omitted.

Afterwards in Chapter 3, the analysis approach of this thesis, the evolution equations first introduced by Beyer [21], is described in greater length. The approach considers the stochastic process induced by the ES as a (stochastic) dynamic system. After introducing the approach used, the analysis is started with intermediate ES on the undisturbed sphere model in Chapter 4. One of the aims is to provide an explanation for the experimental findings by Grünz and Beyer [51] that ES using intermediate recombination do not show the same robustness as a $(1, \lambda)$-ES towards the choice of the learning rate. Afterwards, self-adaptive ES on the noisy sphere are considered.

In Chapter 5, ridge functions are considered. The behavior of self-adaptive ES on two representatives of this function class is analyzed: The sharp ridge with $\alpha = 1$ and the parabolic ridge with $\alpha = 2$. Again, the undisturbed functions are treated first before the analysis is continued with noisy ridge functions in the following sections. As said, the thesis uses a dynamic systems approach to analyze self-adaptive ES. Other approaches include runtime analyses of randomized algorithms for example. In continuous search spaces, Jägersküpper was the first to provide a runtime analysis of evolutionary algorithms [61]. He considered several types of EA, $(1 + \lambda)$-ES, $(1, \lambda)$-ES, and $(\mu + 1)$-ES [62]. Instead of self-adaptation, the focus was on the $1/5$th-rule as adaptation mechanism. The work aimed at and succeeded in deriving lower and upper bounds on the expected runtime. Many of his results were obtained for the sphere model or for the more general positive definite quadratic forms. Among the results obtained are the following

- "The $(1 + 1)$-ES performs with overwhelming probability $\mathcal{O}(N)$ steps to halve the approximation error in the search space.
- The $(1+\lambda)$-ES as well as the $(1, \lambda)$-ES get along with $\mathcal{O}(N/\sqrt{\ln(1 + \lambda)})$ steps with overwhelming probability — when the $1/5$-rule bases on the number of successful mutations.
- The $(1 + \lambda)$-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 $\mathcal{O}(N/\sqrt{\ln(1 + \lambda)})$ steps with overwhelming probability, which is asymptotically optimal.
- The $(\mu+1)$-ES using Gaussian mutations adapted by the $1/5$-rule performs $\mathcal{O}(\mu N)$ steps with overwhelming probability[1] to halve the approximation error in the search space, which is also asymptotically optimal." [62]

[1] An event occurs with overwhelming probability w.r.t. N if the probability of nonoccurrence is exponentially small in N (see [62, p.15].

Runtime analyses of randomized algorithms aim at deriving upper and lower bounds for the expected runtime. One of the tasks is to find the relationship between the expected runtime and the search space dimensionality. The aim is on the one hand to provide the lower bounds and on the other to give exact proofs of the results.

The dynamic systems approach follows a different direction and aims at a different type of results. As stated in [30], one of the aims is to provide analytical formulas of the mean-value dynamics. The dynamic systems approach relies on asymptotical simplifications and on approximations. In a sense it considers a model of the actual algorithm. The analytical formulas derived can be used on the one hand to give recommendations for the parameter setting and to provide insights into the working mechanism of ES on the other.

In this thesis, the dynamic systems approach is applied to derive the following findings:

Self-Adaptation on the Sphere Model

Self-Adaptation and Intermediate Recombination, Section 4.1

1. An explanation of the experimental finding by Grütz and Beyer [51] that intermediate recombinative ES are sensitive to the choice of the learning rate can be provided.

2. Main reason of the sensitivity of the progress rate: The sensitivity towards the choice of the learning rate is due to the self-adaptation mechanism, itself. Due to the genetic repair effect, ES with intermediate recombination may operate with higher mutation strength. The self-adaptation mechanism cannot take this into account.

3. Intermediate recombination and progress: While intermediate ES in contrast to $(1, \lambda)$-ES are sensitive to the choice of the learning rate they may perform superiorly to $(1, \lambda)$-ES. Furthermore, they may reach their specific optimal progress.

4. Optimal learning rate: Provided that the search space dimensionality and the number of offspring are large, it is shown that choosing the learning rate proportional to $1/\sqrt{2N}$ is approximately optimal – as long as the parent number is neither close to one nor close to the number of offspring. Especially, this includes the parent-offspring ratio usually recommended.

Self-Adaptation and Noise, Sections 4.2 and 4.3

1. $(1, \lambda)$-ES suffer from a loss of step size control if the noise strength is too high (Section 4.2). Instead of reaching a stationary state, the mutation strength shows a nearly erratic behavior. Using the assumption that selection in the high noise regime is random, it can be shown that the mutation strength performs a random walk. Larger mutation strengths (which would decrease the influence of the noise) are punished, however, because they may lead more often to worse candidates. As result, the ES is biased towards smaller mutation strengths and shows an irregular behavior.

2. Intermediate $(\mu/\mu_I, \lambda)$-ES are biased towards an increase of the mutation strength. This bias safeguards against a loss of step size control (Section 4.2).

3. Concerning the residual location error, $\mu : \lambda$-ratios around $1/2$ are optimal. Evolution strategies with $\mu : \lambda \in [0.2 - 0.7]$ achieve similar location errors. This enables to follow the usual recommendation to choose $\mu : \lambda$ around 0.27. This allows not only nearly optimal progress in the initial optimization phase but nearly minimal residual location errors (Section 4.3).

4. The residual location error is higher than a (hypothetic) minimal error. In case of intermediate ES, this deviation occurs because of the non-zero stationary mutation strength. But again the deviation of the residual location error from the minimal possible error is small if λ is sufficiently high and even improves more if one of the usual $\mu : \lambda$-ratio is chosen. Recombinative ES achieve nearly optimal location errors (Section 4.3).

Self-Adaptation on the Ridge Function Class

The Sharp Ridge, Sections 5.1.1 and 5.2.1

1. It has been shown in experiments [56] that self-adaptive ES on the sharp ridge may converge prematurely. It is shown in Section 5.1.1 that the size of the constant d w.r.t. the population parameters μ and λ is the critical parameter. Large d-values cause premature convergence of the evolution strategies (Section 5.1.1). To a minor extend this can be remedied by increasing λ. Using recombination with the usual parent offspring ratio enhances the problem: Premature convergence occurs for even lower values of d.

2. If d is small, the ES progresses with a positive quality change. It can be shown that the usual recommendation of choosing the truncation ratio $\mu : \lambda \approx 0.27$ does not apply – unless the learning rate is small, of course. Instead, it can be shown that a fixed μ-value around $2 - 5$ is a good choice.

3. Provided, that d is small, an increase of the learning rate increases the performance, i.e., the quality change. The optimizer is unattainable for finite learning rates, though. Therefore, no recommendation of how to choose τ can be given.

4. The sharp ridge is an example for a positive side effect of noisy fitness evaluations. First of all, the size of the d-constant must be sufficiently high so that the axis is approached in the first hand. Additive noise stops the ES from realizing the subgoal of optimizing the embedded sphere: The higher this "residual location error" to the axis, the higher the progress of self-adaptive ES. As result, recombination using the usual truncation ratio is not recommended. Recombination is necessary, though, since a $(1, \lambda)$-ES looses step-size control.

5. The behavior of ES, i.e., the stationary normalized mutation and noise strength, on the noisy sharp ridge is very similar to that on the noisy sphere. The mutation strength reacts towards changes of the distance to the axis but not towards changes in x-direction, i.e., towards changes parallel to the direction of the axis.

The Parabolic Ridge, Sections 5.1.2 and 5.2.2

1. On the parabolic ridge, no premature convergence occurs. The ES reaches a stationary distance to the axis and progresses then with a constant mutation strength (on average). The mutation strength only reflects the distance to the axis but not the position on the axis. Recombination has disadvantages: The progress rate decreases when switching from $\mu = 1$ to $\mu > 1$ (see Section 5.1.2).

2. Noise has only positive effects if the size of the parent population exceeds half of the size of the offspring population. If $\mu < \lambda/2$, noise degrades the performance (see Section 5.2.2).

3. For small noise strengths, recombination with a truncation ratio of 0.27 cannot be recommended. To safeguard against a loss of step-size control, recombination has be used, though. The performance loss due to recombination only holds for small noise strengths. If the noise increases, the progress rates of intermediate ES with $\mu \not\approx 1$ and $\mu \not\approx \lambda$ converge to nearly the same progress rate ($\approx 1/(4d)$).

1.1 Underlying Publications

This thesis is based in part on the following publications

1. S. Meyer-Nieberg, H.-G. Beyer: Mutative Self-Adaptation on the Sharp and Parabolic Ridge, in Stephens, C. et al., editors, Proceedings of the 9th International Workshop on Foundations of Evolutionary Algorithms (FOGA-IX), pages 70-96, 2007 [75]

2. S. Meyer-Nieberg, H.-G. Beyer: Self-Adaptation in Evolutionary Algorithms in F. Lobo, C. Lima, and Z. Michaelewicz: Parameter Settings in Evolutionary Algorithms, pages 47-76, Springer, 2007 [76]

3. H.-G. Beyer, S. Meyer-Nieberg: Self-Adaptation on the Ridge Function Class: First Results for the Sharp Ridge, in T.P. Runarsson et al., editors, Parallel Problem Solving from Nature 9, pages 71-80, Springer, 2006 [28]

4. H.-G. Beyer, S. Meyer-Nieberg: Self-Adaptation of Evolution Strategies under Noisy Fitness Evaluations. Genetic Programming and Evolvable Machines. $7(4)$, 295-328, 2006 [27]

5. S. Meyer-Nieberg, H.-G. Beyer: On the Analysis of Self-Adaptive Evolution Strategies: First Results, in McKay, B. et al., editors, Proc. of the CEC'05, Edinburgh, UK, pages 2341-2348, Piscataway, NJ, 2005, IEEE [74]

The contribution of the author of this thesis is at least 50%. Chapter 2 is based on a revised and extended version of [76]. Results from [74] and [27] are presented in Sections 4.1 and 4.2 of Chapter 4. Chapter 5, i.e., Section 5.1, is based in parts on [28] and [75].

2 Self-Adaptation in Evolutionary Algorithms

Evolutionary algorithms (EA) operate on basis of populations of individuals. Their performance depends on the characteristics of the population's distribution. Self-Adaptation aims at biasing the distribution towards appropriate regions of the search space – keeping up sufficient diversity among individuals in order to enable further evolvability.

Generally, this is achieved by adjusting the setting of control parameters. Control parameters can be of various forms – for instance mutation rates, recombination probabilities, or the population size (see, e.g., [16]).

The goal is not only to find suitable adjustments but to do this efficiently. The task is even further complicated: The EA faces a dynamic problem since a parameter setting that was optimal at the beginning of an EA-run may become unsuitable during the evolutionary process. For this reason, there is generally a need for a steady modification or adaptation of the control parameters during the run of an EA.

This chapter considers the principle of self-adaptation which is explicitly used in evolutionary programming (EP) [47, 48] and evolution strategies (ES) [81, 87] while it is rarely used in genetic algorithms (GA) [58, 59]. The areas of evolutionary algorithms differ in their terminology to some extent: For instance, the term crossover is used more often in the field of genetic algorithms and generally denotes recombination of two parents. Also, the mutation strength is referred to as the mutation rate in GA.

Individuals of a population represent possible solutions. These are coded in a set of object parameters that can be interpreted as the genome of the individual. The basic idea of explicit self-adaptation consists in incorporating control parameters into the genome and evolving them alongside with the object parameters.

In this chapter, an overview over the self-adaptative behavior of evolutionary algorithms is provided. First, a short overview over the historical development of adaptation mechanisms in evolutionary computation is given in Section 2.1. In the following part, i.e., in Section 2.2, classification schemes for grouping the various approaches are presented. Afterwards, self-adaptative mechanisms are considered. The overview is started by some examples – introducing self-adaptation of the strategy parameter and of the crossover operator. Several authors have pointed out that the concept of self-adaptation transcends explicit self-adaptation. Section 2.3.2 is devoted to such ideas. The mechanism of self-adaptation has been examined in various areas in order to find answers to the question under which conditions self-adaptation works and when it could fail. Therefore, the chapter closes with a short overview over some of the research done in this field.

2.1 A Short History of Adaptation in Evolutionary Algorithms

This section sketches shortly the historic development of adaptation mechanisms. The first proposals to adjust the control parameters of a computation automatically date back to the early days of evolutionary computation.

In 1967, Reed, Toombs, and Barricelli [83] experimented with the evolution of probabilistic strategies playing a simplified poker game. Half of a player's genome consisted of strategy parameters de-

termining, e.g., the probabilities for mutation or the probabilities for crossover with other strategies. These strategy parameters were subject to random variation. Interestingly, it was shown for a play with a known optimal strategy that the evolutionary simulation realized nearly optimal plans.

Also in 1967, Rosenberg [84] proposed to adapt crossover probabilities and Bagley [18] suggested incorporating the control parameters into the representation of an individual in GA. Although Bagley's suggestion is one of the earliest proposals of applying classical self-adaptive methods, self-adaptation as usually used in ES appeared relatively late in genetic algorithms. In 1987, Schaffer and Morishima [86] introduced the self-adaptive *punctuated crossover* adapting the number and location of crossover points. Some years later in 1992, a first method to self-adapt the mutation operator was suggested by Bäck [14, 13]. He proposed a self-adaptive mutation rate in genetic algorithms similar to evolution strategies.

The idea of using a meta-GA can be found quite early. Here, an upper-level GA tries to tune the control parameters of a lower-level algorithm which in turn tries to solve the original problem. The first suggestion stems from Weinberg [102] in 1970 and gave rise to the work by Mercer and Sampson [73].

Concerning evolution strategies, the need to adapt the mutation strength (or strengths) appropriately during the evolutionary process was recognized 1973 in Rechenberg's seminal book *Evolutionsstrategie* [81].

He proposed the well-known $1/5$th rule, which was originally developed for $(1+1)$-ES. It relies on counting the successful and unsuccessful mutations for a certain number of generations. If more than $1/5$th of mutations leads to an improvement the mutation strength is increased and decreased otherwise. The aim was to stay in the so-called *evolution window* which guarantees nearly optimal progress.

In addition to the $1/5$th rule, Rechenberg [81] also proposed to couple the evolution of the strategy parameters with that of the object parameters. Both parameter sets were randomly changed. The idea of (explicit) self-adaptation was born. To compare the performance of this *learning population* with that of an ES using the $1/5$th rule, Rechenberg conducted some experiments on the sphere and corridor model. The learning population exhibited a higher convergence speed and even more important it proved to be applicable in cases where it is improper to use the $1/5$th rule. Self-adaptation thus appeared as a more universally usable method.

Since then various methods for adapting control parameters in evolutionary algorithms have been developed – ranging from adapting crossover probabilities in genetic algorithms to a direct adaptation of the distribution [36].

In 1974, Schwefel [87, 89] introduced a self-adaptive method for changing the strategy parameters in evolution strategies which is today commonly associated with the term self-adaptation. In its most general form, the full covariance matrix of a general multidimensional normal distribution is adapted. A similar method of adapting the strategy parameters was offered by Fogel et al. [46] in the area of evolutionary programming – the so-called meta-EP operator for changing the mutation strength.

A more recent technique, the cumulative path-length control, stems from Ostermeier, Hansen, and Gawelczyk [78]. One of the aims is to derandomize the adaptation of the strategy parameters. The methods developed, the cumulative step-size adaptation (CSA) as well as the covariance matrix adaptation (CMA) [53], make use of an *evolution path*, $\mathbf{p}^{(g+1)} = (1-c)\mathbf{p}^{(g)} + \sqrt{c(2-c)}\mathbf{z}_{\text{sel}}^{(g+1)}$, which cumulates the selected mutation steps. The variable $\mathbf{p}^{(g)}$ gives the path at generation g whereas $\mathbf{z}_{\text{sel}}^{(g)}$ denotes the selected mutation steps, i.e., in the case of $(\mu/\mu_I, \lambda)$-ES $\mathbf{z}_{\text{sel}}^{(g)}$ equals the centroid of the mutation vectors of the μ best offspring. The basic working mechanism can be illustrated by a simple example. Consider an evolution path with purely random selection (see [29]): Since the mutations

are normally distributed, the cumulated evolution path is given by $\mathbf{u}^{(g)} = \sum_{k=1}^{g} \sigma\mathcal{N}^{(k)}(\mathbf{0}, \mathbf{1})$, where $\mathcal{N}(\mathbf{0}, \mathbf{1})$ is a random vector with identically independently $\mathcal{N}(0, 1)$ normally distributed components with zero mean and variance one. Therefore, the length of $\mathbf{u}^{(g)}$ is χ-distributed with expectation $\overline{u} = \sigma\overline{\chi}$. Fitness based selection changes the situation: If the mutation steps are too large on average, smaller mutations will be selected. Thus, the path-length is smaller than $\overline{u}$ and the step size should be decreased. Otherwise if the path-length is larger than the expected $\overline{u}$, the step-size should be increased. The cumulative step-size adaptation is also used in the CMA-algorithm. However, additionally CMA adapts the whole covariance matrix [53] and as such it represents the state-of-the-art in real-coded evolutionary optimization algorithms.

2.2 A Taxonomy of Adaptation

As the previous section showed, various methods for changing and adapting control parameters of evolutionary algorithms exist and adaptation can take place on different levels.

Mainly, two taxonomy schemes were proposed – the elder by Angeline [2] in 1995 and the younger by Eiben, Hinterding, and Michaelewicz [42] in 1999. These schemes group adaptive computations into distinct classes – distinguishing evolutionary algorithms by the type of adaptation, i.e., how the parameter is changed, and by the level of adaptation, i.e., where the changes occur.

Let us start with Angeline's classification [2]. Considering the type of adaptation, adaptive evolutionary computations are divided into algorithms with *absolute update rules* and into algorithms with *empirical update rules*.

If an *absolute update rule* is applied, a statistic is computed. This may be done by sampling over several generations or by sampling the population. Based on the result, it is decided by means of a deterministic and fixed rule if and how the operator is to be changed. Rechenberg's $1/5$th-rule [81] is one well-known example of this group.

In contrast to this, evolutionary algorithms with *empirical update rules* control the values of the strategy parameters themselves. The strategy operator may be interpreted as an incorporated part of the individual's genome, thus being subject to "genetic variations" [2]. In case the strategy parameter variation leads to an individual with a sufficiently good fitness, it is selected and "survives". Individuals with appropriate strategy parameters should – on average – have good fitness values and thus a higher chance of survival than those with badly tuned parameters. As a result, the EA should be able to self-control the parameter change.

As Smith [92] points out, the difference between these two types of algorithms lies in the nature of the transition function. The transition function maps the set of operators at generation t on that at $t + 1$. In the case of absolute update rules, it is defined externally. In the case of self-adaptive algorithms, the transition function is a result of the operators and is defined by the algorithm itself.

Both classes of adaptive evolutionary algorithms can be further subdivided based on the level the adaptive parameters operate on. Angeline distinguished between *population-*, *individual-*, and *component-level* adaptive parameters.

Population-level adaptive parameters are changed globally for the whole population. Examples are for instance the mutation strength and the covariance matrix adaptation in CSA and CMA evolution strategies [53]. Adaptation on the *individual level* changes the control parameters of an individual and these changes only affect that individual. The probability for crossover in GA is for instance adapted in [86] on the level of individuals. Finally, *component-level* adaptive methods affect each component of an individual separately. Self-Adaptation in ES with correlated mutations (see Section 2.3.1) belongs to this adaptation type.

Angeline's classification was extended and broadened by Eiben, Hinterding, and Michaelewicz

[42]: Adaptation schemes are again classified firstly by the type of adaptation and secondly – as in [2] – by the level of adaptation. Considering the different levels of adaptation a fourth level, *environment level adaptation*, was introduced to take non static responses of the environment into account.

Concerning the adaptation type, the algorithms are divided into *static*, i.e., no changes of the parameters occur, and *dynamic* algorithms. The term "dynamic adaptation" is used to classify any algorithm that changes the strategy parameters and is doing so without any external control. Based on the *mechanism of adaptation* three subclasses are distinguished: *deterministic*, *adaptive*, and finally *self-adaptive* algorithms. The latter classes comprise the same groups of algorithms as in Angeline's classification [2].

A deterministic adaptation is used if the control parameter is changed according to a deterministic rule *without* taking into account any present information by the evolutionary algorithm itself. Examples of this adaptation class are the time-dependent change of the mutation rates proposed by Holland [59] and the cooling schedule in simulated annealing like selection schemes.

Algorithms with an adaptive dynamic adaptation rule take feedback from the EA itself into account and change the control parameters accordingly. Again, a well known member of this class is Rechenberg's $1/5$th-rule. Further examples include Davis' adaptive operator fitness [35] and Julstrom's adaptive mechanism [63]. The former relates the usage probability of reproduction operators to their success. The latter takes the performance of crossover and mutations as basis to tune their application ratio.

2.3 Self-Adaptation: The Principles

This section sketches the principles of self-adaptation. First, some examples are given to illustrate the use of self-adaptation. Self-Adaptation can be seen in a broader context than given by the original definition. This concept of *generalized self-adaptation* is pointed out in the following subsection. The section ends with general demands for self-adaptive operators.

2.3.1 Self-Adapted Parameters: Some Examples

In this subsection some examples are presented in order to illustrate the basic principle. The subsection starts with self-adaptation of strategy parameters which is probably the best known form before addressing self-adaptation of recombination operators.

Self-Adaptation of Strategy Parameters

The technique most commonly associated with the term self-adaptation was introduced by Rechenberg [82] and Schwefel [87, 88] in the area of evolution strategies and independently by Fogel [45] for evolutionary programming. The control parameters considered here apply to the mutation process and parameterize the mutation distribution. The mutation is usually given by a normally distributed random vector, i.e., $\mathbf{Z} \sim \mathbf{N}(\mathbf{0}, \mathbf{C})$. The entries c_{ij} of the covariance matrix $\mathbf{C}$ are given by $c_{ii} = var(Z_i)$ or by $c_{ij} = cov(Z_i, Z_j)$ if $j \neq i$. The density function reads

$$p_Z(Z_1, \ldots, Z_N) = \frac{\mathrm{e}^{-\frac{1}{2}\mathbf{Z}^{\mathrm{T}}\mathbf{C}^{-1}\mathbf{Z}}}{\sqrt{(2\pi)^N \det(\mathbf{C})}}, \tag{2.1}$$

where N is the dimensionality of the search space. The basic step in the self-adaptation mechanism consists of a mutation of the mutation parameters themselves. In contrast to the additive change of the object variables, the mutation of the mutation strengths (i.e., the standard deviations $\sqrt{c_{ii}}$ in (2.1)) is realized by a multiplication with a random variable. The resulting mutation parameters are then

applied in the variation of the object parameters. It should be mentioned here that concerning evolution strategies, the concept of self-adaptation was originally developed for non-recombinative $(1, \lambda)$-ES. After multi-parent strategies were proposed, self-adaptation was adapted accordingly. The reader is referred to Section 1 for a description of a multi-parent $(\mu/\rho, \lambda)$-ES with σ-self-adaptation. Depending on the form of $\mathbf{C}$, different mutation distributions have to be taken into account. Considering the simplest case $\mathbf{Z} = \sigma\mathcal{N}(\mathbf{0}, \mathbf{I})$, the mutation of σ is given by

$$\sigma' = \sigma \mathrm{e}^{\tau\epsilon} \tag{2.2}$$

and using the new σ', the mutation of the object parameters reads

$$x_i' = x_i + \sigma'\mathcal{N}(0,1). \tag{2.3}$$

The ϵ in Eq. (2.2) is a random number, often chosen as

$$\epsilon \sim \mathcal{N}(0,1), \tag{2.4}$$

thus, producing log-normally distributed σ' variants. This way of choosing ϵ is also referred to as the "log-normal mutation rule". Equation (2.2) contains a new strategy specific parameter – the *learning rate* τ to be fixed. The general recommendation based on experimental findings is to choose $\tau \propto 1/\sqrt{N}$. Later on this recommendation was shown to be optimal with respect to the convergence speed of $(1, \lambda)$-ES on the sphere [23, p. 303].

If different mutation strengths are used for each dimension, i.e., $Z_i = \sigma_i\mathcal{N}(0,1)$, the update rule

$$\sigma_i' = \sigma_i \exp(\tau'\mathcal{N}(0,1) + \tau\mathcal{N}_i(0,1)) \tag{2.5}$$
$$x_i' = x_i + \sigma_i'\mathcal{N}(0,1) \tag{2.6}$$

has been proposed. It is recommended to choose the learning rates $\tau' \propto 1/\sqrt{2N}$ and $\tau \propto 1/\sqrt{2\sqrt{N}}$ [16].

The approach can also be extended to allow for correlated mutations [16]. Here, rotation angles α_i need to be taken into account leading to the update rule

$$\sigma_i' = \sigma_i \exp(\tau'\mathcal{N}(0,1) + \tau\mathcal{N}_i(0,1)) \tag{2.7}$$
$$\alpha_i' = \alpha_i + \beta\mathcal{N}_i(0,1) \tag{2.8}$$
$$\mathbf{x}' = \mathbf{x} + \mathcal{N}(0, \mathbf{C}(\sigma', \alpha)) \tag{2.9}$$

where $\mathbf{C}$ is the covariance matrix [16]. The parameter β is usually chosen as 0.0873 [88].

In EP, a different mutation operator, called *meta-EP* [45], is used

$$\sigma_i' = \sigma_i\Big(1 + \alpha\mathcal{N}(0,1)\Big) \tag{2.10}$$
$$x_i' = x_i + \sigma_i'\mathcal{N}(0,1). \tag{2.11}$$

Both operators lead to similar results – provided that the parameters τ and α are sufficiently small.

The log-normal operator, Eqs. (2.2), (2.3), and the meta-EP operator introduced above are not the only possibilities. Self-Adaptation seems to be relatively robust to the choice of the distribution. Another possible operator is given by $\epsilon = \pm\delta$, where $+|\delta|$ and $-|\delta|$ are generated with the same

probability of $1/2$. That is, the resulting cumulative density function (cdf) of δ belongs to a two-point distribution giving rise to the so-called two-point rule. It is usually implemented using $\delta = 1/\tau \ln(1+\beta)$, thus, leading with (2.2) to

$$\sigma_i' = \begin{cases} \sigma_i(1+\beta) & \text{if } u \leq 0.5 \\ \sigma_i/(1+\beta) & \text{if } u > 0.5 \end{cases}, \tag{2.12}$$

with u uniformly distributed random variable on $]0,1]$.

A further variant was proposed by Yao and Liu [105]: They substituted the normal distribution of the meta-EP operator with a Cauchy-distribution. Their new algorithm, called *fast evolutionary programming*, performed well on a set of seperable test functions and appeared to be preferable in the case of multi-modal functions. The Cauchy-distribution is similar to the normal distribution but has a far heavier tail. Its moments are undefined.

In [68], Lee and Yao introduced yet another alternative. They suggested using a Lévy-distribution. Investigating several seperable test functions, they argued that using Lévy-distributions instead of normal distributions may lead to higher variations and a greater diversity. Compared to the Cauchy-distribution, Lévy-distributions allow for a greater flexibility since the Cauchy-distribution appears as a special case of Lévy-distributions.

Self-Adaptation of Recombination Operators

Crossover is traditionally regarded as the main search mechanism in genetic algorithms and most efforts to self-adapt this operator stem from this area. In evolution strategies the term recombination is usually used instead of crossover.

Schaffer and Morishima [86] proposed the *punctuated crossover* which adapts the positions where crossover occurs. An individual's genome is augmented with a bitstring indicating crossover points. A position in this crossover map is changed in the same manner as its counterpart in the original genome. Schaffer and Morishima reported that punctuated crossover performed better than one-point crossover. Spears [98] points out, however, that the improvement of the performance might not necessarily be due to self-adaptation but to the generic advantage of crossover with more than one crossover point over one-point crossover.

Spears [98] self-adapted the form of the crossover operator using an additional bit to decide whether two-point or uniform crossover should be used for creating the offspring. Again, it should be noted that Spears attributes the improved performance not to the self-adaptation process itself but rather to the increased diversity that is offered to the algorithm.

Smith and Fogarty [95] introduced the so-called LEGO-algorithm, a linkage evolving genetic algorithm. The objects which are adapted are *blocks*, i.e., linked neighboring genes. Each gene has two additional bits which indicate whether it is linked to its neighbor on the right or on the left. These additional bits are also subject to mutation. Two neighboring genes are then called *linked* if the respective bits are set. More than two parents may contribute in the creation of an offspring. The positions of an offspring are filled successively by a competition between parental blocks. The blocks have to be eligible, i.e., they have to start at the position currently considered. The fittest block is copied as a whole and then the process starts anew.

2.3.2 A Generalized Concept of Self-Adaptation

In [16], Bäck identified two key features of self-adaptation: Self-adaptation aims at biasing the population distribution to more appropriate regions of the search space by making use of an indirect link between good strategy parameter or recombination operator values and good object variables.

Furthermore, self-adaptation relies on a population's diversity. While the adaptation of the operator ensures a good convergence speed, the degree of diversity determines the convergence reliability. More generally speaking, self-adaptation controls the relationship between parent and offspring population, i.e., the transmission function (see, e.g., Altenberg [1]). The control can be direct by manipulating control parameters in the genome or more implicit. In the following, we see that self-adaptation can be put into a broader context.

Igel and Toussaint [60] addressed the question of neutral genotype-phenotype mapping. They point out that neutral genome parts give an algorithm the ability to "vary the search space distribution independent of the phenotypic variation" [60]. This may be regarded as one of the main benefits of neutrality. While neutrality induces a redundancy in the relationship between genotype-phenotype, the mapping from the genome to the population distribution has to be taken into account, too. The latter mapping cannot be viewed as redundant in general. This use of neutrality is termed *generalized self-adaptation*. It also comprises the classical form of self-adaptation since the strategy parameters it adapts belong to the neutral part of the genome.

More formally, generalized self-adaptation is defined as "adaptation of the exploration distribution $P_{\mathcal{P}}^{(t)}$ by exploiting neutrality – i.e., independent of changing phenotypes in the population, of external control, and of changing the genotype-phenotype mapping" [60]. Igel and Toussaint showed additionally that neutrality cannot be seen generally as a disadvantage since the enlargement of the search space does not necessarily lead to a significant degradation of the performance.

In [49], Glickman and Sycara referred to an *implicit self-adaptation* caused by a non-injective genotype-phenotype mapping. Again there are variations of the genome that do not alter the fitness value but influence the transmission function which induces a similar effect.

Beyer and Deb [38] pointed out that in well-designed real-coded GA, the parent offspring transmission function is controlled by the characteristics of the parent population. Thus, the GA performs an implicit form of self-adaptation. In contrast to the explicit self-adaptation in ES, an individual's genome does not contain any control parameters. Deb and Beyer [40] examined the dynamic behavior of real-coded genetic algorithm (RCGA) that apply simulated binary crossover (SBX) [37, 41]. In SBX, two parents x^1 and x^2 create two offspring y^1 and y^2 according to

$$\begin{aligned} y_i^1 &= 1/2\Big((1-\beta_i)x_i^1 + (1+\beta_i)x_i^2\Big) \\ y_i^2 &= 1/2\Big((1+\beta_i)x_i^1 + (1-\beta_i)x_i^2\Big). \end{aligned} \tag{2.13}$$

The random variable β has the density

$$p(\beta) = \begin{cases} 1/2(\eta+1)\beta^{\eta} & \text{if } 0 \le \beta \le 1 \\ 1/2(\eta+1)\beta^{-\eta-2} & \text{if } \beta > 1 \end{cases}. \tag{2.14}$$

The authors pointed out that these algorithms show self-adaptive behavior although an individual's genome does not contain any control parameters. Well-designed crossover operators create offspring depending on the difference in parent solutions. The spread of children solutions is in proportion to the spread of the parent solutions. Solutions near the parent solutions are more likely to be created as children solutions than more distant solutions [40]. In this manner, the diversity in the parental population controls that of the offspring population.

Self-adaptation in evolution strategies has similar properties. In both cases, offspring closer to the parents have a higher probability to be created than individuals further away. While the implicit self-adaptability of real-coded crossover operators is well understood today, it is interesting to point out that even the standard one or k-point crossover operators operating on binary strings do have this

property: Due to the mechanics of these operators, bit positions which are common in both parents are transferred to the offspring. However, the other positions are randomly filled. From this point of view, crossover can be seen as a *self-adaptive mutation operator*, which is in contrast to the building block hypothesis [50] usually offered to explain the working of crossover in binary GA.

2.3.3 Demands on the Operators: Real-coded Algorithms

Several postulates and guidelines have been devised that should be fulfilled by self-adaptative evolutionary algorithms. Many of them address the mutation operators. In [26, 23, 29] several rules for the design of mutation operators were introduced that stem from analyses of implementations and theoretical considerations in evolution strategies:

1. *reachability*: every finite state must be reachable,
2. *scalability*: the mutation operator must be tunable in order to adapt to the fitness landscape, and
3. *unbiasedness*: it must not introduce a bias on the population.

A detailed discussion can be found in [29], for example. The necessity of the first two requirements can be immediately discerned. The demand of unbiasedness is explained in the following. It should be noted that unbiasedness is also required in the case of the recombination operator [26, 65]. The demand of unbiasedness becomes clear when considering that the evolutionary search behavior of an EA can be divided into two phases: Exploitation of the search space by selecting good solutions (reproduction) and exploration of the search space by means of variation. Only the former generally makes use of fitness information, whereas the latter should ideally rely on search space information of the population alone. Thus, under a variation operator, the expected population mean should remain unchanged, i.e., the variation operators should not bias the population. This requirement, first made explicit in [26], may be regarded as a basic design principle for variation operators in EA. The basic work [26] additionally proposed design principles with respect to the changing behavior of the population variance. Generally, selection changes the population variance. In order to avoid premature convergence, the variation operator must counteract that effect of the reproduction phase to some extent. General rules how to do this are, of course, nearly impossible to give but some *minimal* requirements can be proposed concerning the behavior on certain fitness landscapes [26].

For instance, Deb and Beyer [26] postulated that the population variance should increase exponentially with the generation number on flat or linear fitness functions. As pointed out by Hansen [52] this demand might not be sufficient. He proposed a linear increase of the expectation of the logarithm of the variance instead. Based on the desired behavior in flat fitness landscapes, Beyer and Deb [26] advocated applying variation operators that also increase the population variance in the general case of unimodal fitness functions. While the variance should be decreased if the population brackets the optimum, this should not be done by the variation operator. Instead, this task should be left to the selection operator.

In the case of crossover operators in real-coded genetic algorithms (RCGA), similar guidelines have been proposed by Kita and Yamamura [65]. They supposed that the distribution of the parent population indicates an appropriate region for further search. As before, the first guideline states that the statistics of the population should be preserved: Both, the mean as well as the variance-covariance matrix, should be retained. Additionally, the crossover operator should lead to as much diversity in the offspring population as possible. The first guideline may be violated, though, since the selection operator typically reduces the variance. Therefore, it may be necessary to increase the present search region.

2.4 Self-Adaptation in EAs: Theoretical and Empirical Results

In this section, empirical and theoretical research is reviewed that aims at understanding the working of self-adaptive EA and at evaluating their performance. First, genetic algorithms are addressed before research approaches of self-adaptation in evolution strategies and evolutionary programming are described.

2.4.1 Genetic Algorithms

In genetic algorithms, self-adaptation is applied to the crossover operator and to the mutation rate. First, a review of self-adaptation of the crossover operator is given before the question of adaptation of the mutation rate is addressed.

Self-Adaptation of the Crossover Operator: Real-Coded Genetic Algorithms in Flat Fitness Landscapes

Beyer and Deb [39] analyzed three crossover operators commonly used in real-coded genetic algorithms, i.e., the simulated binary crossover (SBX) by Deb and Agrawala [37], the blend crossover operator (BLX) of Eshelman and Schaffer [44], and the *fuzzy recombination* of Voigt et al. [101]. All crossover operators use the following recombination operator

$$\begin{aligned} y_{1,k} &:= \frac{1}{2}\Big((1-\beta_k)x_{1,k} + (1+\beta_k)x_{2,k}\Big) \\ y_{2,k} &:= \frac{1}{2}\Big((1+\beta_k)x_{1,k} + (1-\beta_k)x_{2,k}\Big) \end{aligned} \tag{2.15}$$

with $x_{1,k}$ and $x_{2,k}$ drawn independently from the parent population and β_k a random variable (see [39]). The crossover operators differ in the distribution of the random variable β_k.

The analysis was aimed at ascertaining if and under which conditions the postulates proposed in Section 2.3.3 are fulfilled [39]. To this end, expressions for the mean and the variance of the offspring population in relation to the parent population were derived. The fitness environments considered were flat fitness landscapes and the sphere. As mentioned before in Section 2.3.3, self-adaptation should not change the population mean in the search space, i.e., it should not introduce a bias, but it should – since a flat fitness function is considered – increase the population variance and this exponentially fast.

It was shown in [39] that the crossover operator leaves the population mean unchanged regardless of the chosen distribution of the random variable β_k. Concerning the population variance, an exponential change can be asserted. Whether the variance expands or contracts depends on the population size and on the second moment of the random variable. Thus, a relationship between the population size and the distribution parameters of the random variables can be derived which ensures an expanding population.

A further investigation of self-adaptation of the crossover operator was offered by Kita [64]. He analyzed real-coded genetic algorithms using UNDX-crossover (unimodal normal distribution) and performed a comparison with evolution strategies. Based on empirical results, he pointed out that both appear to work reasonably well although naturally some differences in their behavior was observed. The ES for example widens the search space faster if the system is far away from an optimum. But the RCGA appears to have a computational advantage in high-dimensional search spaces compared to an ES which adapts the rotation angles of the covariance matrix according to Eqs. (2.7)-(2.9). Kita used a $(15, 100)$-ES with the usual recommendations for setting the learning rates.

Self-Adaptation of the Mutation Rate in Genetic Algorithms

Traditionally, the crossover (recombination) operator is seen as the main variation operator in genetic algorithms, whereas the mutation operator was originally proposed as a kind of background operator endowing the algorithm with the potential ability to explore the whole search space. Actually, there are good reasons to consider this as a reasonable recommendation in genetic algorithms with genotype-phenotype mapping from $\mathbb{B}^\ell \rightarrow \mathbb{R}^\ell$. As has been shown in [26], standard crossover of the genotypes does not introduce a bias on the population mean in the phenotype space. Interestingly, this does *not* hold for bit-flip mutations. That is, mutations in the genotype space result in a biased phenotypic population mean – thus violating the postulates formulated in [26]. On the other hand, over the course of the years it was observed that for genetic algorithms on (pseudo) boolean functions (i.e., the problem specific search space is the $\mathbb{B}^\ell$) the mutation operator might also be an important variation operator to explore the search space (see, e.g., [99]). Additionally, it was found that the optimal mutation rate or mutation probability does not only depend on the function to be optimized but also on the search space dimensionality and the current state of the search (see, e.g., [15]).

A mechanism to self-adapt the mutation rate was proposed by Bäck [13, 14] for GA using the standard ES approach. The mutation rate is encoded as a bit-string and becomes part of the individual's genome. As it is common practice, the mutation rate is mutated first which requires its decoding to $[0, 1]$. The decoded mutation rate is used to mutate the positions in the bit-string of the mutation rate itself. The mutated version of the mutation probability is then decoded again in order to be used in the mutation of the object variables.

Several investigations have been devoted to the mechanism of self-adaptation in genetic algorithms. Most of the work is concentrated on empirical studies. These are often directed to possible designs of mutation operators trying to identify potential benefits and drawbacks.

Bäck [14] investigated the asymptotic behavior of the encoded mutation rate – neglecting the effects of recombination and selection. The evolution of the mutation rate results in a Markov chain[1]. The absorbing state of this chain is zero which shows the convergence of the simplified algorithm.

The author showed empirically that an GA with an extinctive selection scheme[2] with self-adaptation performs better than a reference GA without adaptation [14]. For the comparison, three high-dimensional test functions (two unimodal, one multimodal) were used.

In [13], a self-adaptive GA optimizing the bit-counting function was examined. Comparing its performance with a GA that applies an optimal deterministic schedule to tune the mutation strength, it was shown that the self-adaptive algorithm realizes nearly optimal mutation rates.

The representation of the mutation rate as a bit-string may hamper its fine-tuning by self-adaptation. To overcome this problem, the genome is extended with a real-coded mutation rate $p \in]0, 1[$ in [17]. Using a real-coded mutation rate in GA, however, necessitates several requirements: The expected change of p should be zero and small changes should occur with a higher probability than large ones. Also, it is required that a change by a factor c has the same probability as by $1/c$. The authors used a logistic change function with parameter γ. The algorithm was compared with a GA without any adaptation and with a GA using a deterministic time-dependent schedule. The GA with the deterministic time-dependent schedule performed best on the test-problems chosen. The self-adaptive GA was ranked in second place. Unfortunately, the learning rate γ was found to have a high impact.

Considering the originally proposed algorithm [14], Smith [94] demonstrated that it may get stuck in suboptima with prematurely reduced mutation strength. He showed that the algorithm becomes

[1] A Markov chain is a stochastic process which possesses the Markov property, i.e., the future behavior depends on the present state but not on the past.

[2] A selection scheme is extinctive iff at least one individual is not selected (see [14]).

more robust by using a fixed learning rate for the bitwise mutation of the mutation strength.

In 1996, Smith and Fogarty [96] examined empirically a self-adaptive steady state $(\mu + 1)$-GA finding that self-adaptation may improve the performance of a GA. The mutation rate was encoded again as a bit-string and several encoding methods were applied. Additionally, the impact of crossover in combination with a self-adaptive mutation rate was investigated. The self-adaptive GA appeared to be relatively robust with respect to changes of the encoding or crossover.

In [97], the authors examined the effect of self-adaptation when the crossover operator and the mutation rate are both simultaneously adapted. It appeared that at least on the fitness functions considered synergistic effects between the two variation operators come into play.

To investigate the behavior of self-adaptive genetic algorithms more closely, Smith [93] developed a model to predict the mean fitness of the population. In the model, several simplifications are made. Most importantly, the mutation rate is only allowed to assume q different values. Because of this, Smith also introduced a new scheme for mutating the mutation rate. The probability of changing the mutation rate is given by $p_z = z(q-1)/q$, where z is the so-called *innovation rate*.

In [100], Stone and Smith compared a self-adaptive GA using the log-normal operator with a GA with discrete self-adaptation, i.e., a GA implementing the model proposed in [93]. To this end, they evaluated the performance of a self-adaptive GA with continuous self-adaptation and the performance of their model on a set of five test functions. Stone and Smith found that the GA with discrete self-adaptation behaves more reliably whereas the GA with continuous self-adaptation may get stuck in local optima. They attributed this behavior to the fact that the mutation rate gives the probability of bitwise mutation. As a result, smaller differences between mutation strengths are lost and more or less the same amount of genes are changed. The variety the log-normal operator provides in continuous search spaces cannot be carried over to the genome effectively and the likelihood of large changes is small. In addition, they argued that concerning the discrete self-adaptation a innovation rate of one is connected with an explorative behavior of the algorithm. This appears more suitable for multimodal problems whereas smaller innovation rates are preferable for unimodal functions.

2.4.2 Evolution Strategies and Evolutionary Programming

Research on self-adaptation in evolution strategies has a long tradition. The first theoretic in-depth analysis has been presented by Beyer [21]. It focused on the conditions under which a convergence of the self-adaptive algorithm can be ensured. Furthermore, it also provided an estimate of the convergence order.

The evolutionary algorithm leads to a stochastic process or more exactly to a Markov chain [77]. The random variables chosen to characterize the system's behavior are the object vector (or its distance to the optimizer, respectively) and the mutation strength.

There are several approaches to analyze the Markov chain. The first [31, 12] considers the chain directly whereas the second [90, 91, 55] analyzes induced supermartingales. The third [23, 38] uses a model of the Markov chain in order to determine the dynamic behavior.

Convergence Results using Markov Chains

Bienvenüe and François [31] examined the global convergence of adaptive and self-adaptive $(1, \lambda)$-evolution strategies on spherical functions. To this end, they investigated the induced stochastic process $z_t = \|x_t\|/\sigma_t$. The parameter σ_t denotes the mutation strength, whereas x_t stands for the object parameter vector.

They showed that (z_t) is a homogeneous Markov chain, i.e., z_t only depends on z_{t-1}. This also confirms an early result obtained in [21] that the evolution of the mutation strength can be decoupled

from the evolution of $\|x_t\|$. Furthermore, they showed that (x_t) converges or diverges log-linearly – provided that the chain (z_t) is Harris-recurrent[3].

Auger [12] followed their line of research focusing on $(1, \lambda)$-ES optimizing the sphere model. She analyzed a general model of a $(1, \lambda)$-ES with

$$\begin{aligned} x_{t+1} &= \arg\min \left\{ f(x_t + \sigma_t \eta_t^1 \xi_t^1), \ldots, f(x_t + \sigma_t \eta_t^\lambda \xi_t^\lambda) \right\} \\ \sigma_{t+1} &= \sigma_t \eta^*(x_t),\ \eta^* \text{ given by } x_{t+1} = x_t + \sigma_t \eta^*(x_t) \xi^*(x_t), \end{aligned} \tag{2.16}$$

i.e., σ_{t+1} is the mutation strength which accompanies the best offspring. The function f is the sphere and η and ξ are random variables. Auger proved that the Markov chain given by $z_t = x_t/\sigma_t$ is Harris-recurrent and positive if some additional assumptions on the distributions are met and the offspring number λ is chosen appropriately. As a result, a law of large numbers can be applied and $1/t \ln(\|x_t\|)$ and $1/t \ln(\sigma_t)$ converge almost surely[4] to the same quantity – the convergence rate. This ensures either log-linearly convergence or divergence of the ES – depending on the sign of the limit. Auger further showed that the Markov chain (z_t) is also geometrically ergodic (see, e.g., [77]) so that the Central Limit Theorem can be applied. As a result, it is possible to derive a confidence interval for the convergence rate. This is a necessary ingredient, because the analysis still relies on Monte-Carlo simulations in order to obtain the convergence rate (along with its confidence interval) numerically for the real $(1, \lambda)$-ES.

In order to perform the analysis, it is required that the random variable ξ is symmetric and that both random variables ξ and η must be absolutely continuous with respect to the Lebesgue-measure. Furthermore, the density p_ξ is assumed to be continuous almost everywhere, $p_\xi \in L^\infty(\mathbb{R})$, and zero has to be in the interior of the support of the density[5], i.e., $0 \in \mathring{\operatorname{supp}}\, p_\xi$. Additionally, it is assumed that $1 \in \mathring{\operatorname{supp}}\, p_\eta$ and that $\mathrm{E}[|\ln(\eta)|] < \infty$ holds. The requirements above are met by the distribution functions normally used in practice, i.e., the log-normal distribution (mutation strength) and normal distribution (object variable). In order to show the Harris-recurrence, the positivity, and the geometric ergodicity, so-called Forster-Lyapunov drift conditions need to be obtained [77, 12]. To this end, new random variables are to be introduced

$$\hat{\eta}(\lambda)\hat{\xi}(\lambda) = \min \left\{ \eta^1 \xi^1, \ldots, \eta^\lambda \xi^\lambda \right\}. \tag{2.17}$$

They denote the minimal change of the object variable. For the drift conditions a number α is required. Firstly, α has to ensure that the expectations $\mathrm{E}[|\xi|^\alpha]$ and $\mathrm{E}[(1/\eta)^\alpha]$ are finite. Provided that also $\mathrm{E}[|1/\hat{\eta}(\lambda)|^\alpha] < 1$, α can be used to give a drift condition V. More generally stated, α has to decrease the reduction velocity of the mutation strength associated with the best offspring of λ trials sufficiently. Thus, additional conditions concerning α and the offspring number λ are introduced leading to the definition of the sets

$$\Gamma_0 = \{\gamma > 0 : \mathrm{E}\left[|1/\eta|^\gamma\right] < \infty \text{ and } \mathrm{E}\left[|\xi|^\gamma\right] < \infty\} \tag{2.18}$$

and

$$\Lambda = \bigcup_{\alpha \in \Gamma_0} \Lambda_\alpha = \bigcup_{\alpha \in \Gamma_0} \{\lambda \in N : \mathrm{E}\left[1/\hat{\eta}(\lambda)^\alpha\right] < 1\}. \tag{2.19}$$

[3] Let N_A be the number of passages in the set A. The set A is called Harris-recurrent if $P_z(N_A = \infty) = 1$ for $z \in A$. Or in other words: If the process starting from z visits A infinitely often with probability one. A process (z_t) is Harris-recurrent if a measure ψ exists such that (z_t) is ψ-irreducible and for all A with $\psi(A) > 0$, A is Harris-recurrent (see, e.g., [77]).

[4] A sequence of random variables x_t defined on the probability space $(\Omega, \mathcal{A}, P)$ converges almost surely to a random variable x if $P(\{\omega \in \Omega | \lim_{t\to\infty} x_t(\omega) = x(\omega)\}) = 1$. Therefore, events for which the sequence does not converge have probability zero.

[5] The support of a density f is the closure of the set of all non-zero points of f.

Finally, the almost sure convergence of $1/t \ln(\|x_t\|)$ and $1/t \ln(\sigma_t)$ can be shown for all $\lambda \in \Lambda$. It is not straightforward to give expressions for Λ or Λ_α in the general case although Λ_α can be numerically obtained for a given α. Only if the densities of η and ξ have bounded support, it can be shown that Λ_α is of the form $\Lambda_\alpha = \{\lambda : \lambda \geq \lambda_0\}$.

Convergence Theory with Supermartingales

Several authors [90, 91, 55] use the concept of martingales or supermartingales[6] to show the convergence of an ES or to give an estimate of the convergence velocity. As before, the random variables most authors are interested in are the object variable and the mutation strength.

Semenov [90] and Semenov and Terkel [91] examined the convergence and the convergence velocity of evolution strategies. To this end, they considered the stochastic Lyapunov function V_t of a stochastic process X_t. By showing the convergence of the Lyapunov function, the convergence of the original stochastic process follows under certain conditions.

From the viewpoint of probability theory, the function V_t may be regarded as a supermartingale. Therefore, a more general framework in terms of convergence of supermartingales can be developed. The analysis performed in [91] consists of two independent parts. The first concerns the conditions that imply almost surely convergence of supermartingales to a limit set. The second part (see also [90]) proposes demands on supermartingales which allow for an estimate of the convergence velocity. Indirectly, this also gives an independent convergence proof.

The adaptation of the general framework developed for supermartingales to the situation of evolution strategies requires the construction of an appropriate stochastic Lyapunov function. Because of the complicated nature of the underlying stochastic process, the authors did not succeed in the rigorous mathematical treatment of the stochastic process. Similar to the Harris-recurrent Markov chain approach, the authors had to resort to Monte-Carlo simulations in order to show that the necessary conditions are fulfilled.

In [90] and [91], $(1, \lambda)$-ES are considered where the offspring are generated according to

$$\begin{aligned} \sigma_{t,l} &= \sigma_t \mathrm{e}^{\vartheta_{t,l}} \\ x_{t,l} &= x_t + \sigma_{t,l}\zeta_{t,l} \end{aligned} \tag{2.20}$$

and the task is to optimize $f(x) = -|x|$. The random variables $\vartheta_{t,l}$ and $\zeta_{t,l}$ are uniformly distributed with $\vartheta_{t,l}$ assuming values in $[-2, 2]$ whereas $\zeta_{t,l}$ is defined on $[-1, 1]$. For this problem, it can be shown that the object variable and the mutation strength converge almost surely to zero – provided that there are at least three offspring. Additionally, the convergence velocity of the mutation strength and the distance to the optimizer is bounded from above by a function of the form $\exp(-at)$ which holds asymptotically almost surely.

Hart, DeLaurentis, and Ferguson [55] also used supermartingales in their approach. They considered a simplified $(1, \lambda)$-ES where the mutations are modeled by discrete random variables. This applies to the mutations of the object variables as well as to those of the mutation strengths. Offspring are generated according to

$$\begin{aligned} \sigma_{t,l} &= \sigma_t D \\ x_{t,l} &= x_t + \sigma_{t,l} B. \end{aligned} \tag{2.21}$$

The random variable D may assume three values $\{\gamma, 1, \eta\}$ with $\gamma < 1 < \eta$. The random variable B takes a value of either $+1$ or -1 with probability $1/2$ each. Under certain assumptions, the strategy

[6] A random process X_t is called a supermartingale if $\mathrm{E}[|X_t|] < \infty$ and $\mathrm{E}[X_{t+1}|\mathcal{F}_t] \leq X_t$ where $\mathcal{F}_t$ is, e.g., the σ-field that is induced by X_t.

converges almost surely to the minimum x^* of a function $f : \mathbb{R} \to \mathbb{R}$ which is assumed to be strictly monotonically increasing for $x > x^*$ and strictly monotonically decreasing for $x < x^*$.

As a second result, the authors proved that their algorithm fails to locate the global optimum of a specific multimodal function with probability one. We will return to this aspect of their analysis in Section 2.5.

Instead of using a Lyapunov function as Semenov and Terkel, they introduced a random variable that is derived from the (random) object variable and the mutation strength. It can be shown that this random variable is a nonnegative supermartingale if certain requirements are met. In that case, the ES converges almost surely to the optimal solution if the offspring number is sufficiently high.

The techniques introduced in [55] can be applied to the multi-dimensional case [54] provided that the fitness function is separable, i.e., $g(x) = \sum_{k=0}^{N} g_k(x_k)$, and the g_k fulfill the conditions for f. The authors considered an ES-variant where only one coordinate is changed in each iteration. The coordinate k is chosen uniformly at random. Let $X^t_{\lambda,k}$ and $\Sigma^t_{\lambda,k}$ denote the stochastic processes that result from the algorithm. It can be shown that $X^t_{\lambda,1}, \ldots, X^t_{\lambda,N}$ are independent of each other. This also holds for $\Sigma^t_{\lambda,1}, \ldots, \Sigma^t_{\lambda,N}$. Therefore, the results of the one-dimensional analysis can be directly transferred.

Although the analysis in [55, 54] provides an interesting alternative, it is restricted to very special cases: Due to the kind of mutations used, the convergence results in [55, 54] are, however, not practically relevant if the number of offspring exceeds six.

Dynamic Systems Approach: The Evolution Equations

In 1996, Beyer [21] was the first to provide a theoretical framework for the analysis of self-adaptive EAs. He used approximate equations to describe the dynamics of self-adaptive evolution strategies. Let the random variable $r^{(g)} = \|X^{(g)} - \hat{X}\|$ denote the distance of the present search point to the optimizer and $\varsigma^{(g)}$ the mutation strength. The dynamics of an ES can be interpreted as a Markov process as we have already seen. But generally, the transition kernels for

$$\begin{pmatrix} r^{(g)} \\ \varsigma^{(g)} \end{pmatrix} \quad \to \quad \begin{pmatrix} r^{(g+1)} \\ \varsigma^{(g+1)} \end{pmatrix} \tag{2.22}$$

cannot be analytically determined. One way to analyze the system is therefore to apply a step by step approach extracting the important features of the dynamic process and thus deriving approximate equations.

The change of the random variables can be divided into two parts. While the first denotes the expected change, the second covers the stochastic fluctuations

$$\begin{aligned} r^{(g+1)} &= r^{(g)} - \varphi(r^{(g)}, \varsigma^{(g)}) + \epsilon_R(r^{(g)}, \varsigma^{(g)}) \\ \varsigma^{(g+1)} &= \varsigma^{(g)}\Big(1 + \psi(r^{(g)}\varsigma^{(g)})\Big) + \epsilon_\sigma(r^{(g)}, \varsigma^{(g)}). \end{aligned} \tag{2.23}$$

The expected changes φ and ψ of the variables are termed *progress rate* if the distance is considered and *self-adaptation response* in the case of the mutation strength.

The distributions of the fluctuation terms are approximated using Gram-Charlier series' (or Edgeworth series'), usually cut off after the first term: The stochastic term is approximated using a normal distribution. The variance remains to be determined which can be done using the evolution equations, themselves. In short, this requires the calculations of the second moments and leads to the corresponding second-order progress rate and to the second-order self-adaptation response.

To analyze the self-adaptation behavior of the system, expressions for the respective progress rate and self-adaptation response have to be found. Generally, no closed analytical solution can be derived. Up to now, only results for $(1, 2)$-ES using two-point mutations could be obtained [23, p. 283f][21]. Therefore, several simplifications have to be introduced. For instance, if the log-normal operator is examined, the most important simplification is to consider $\tau \ll 1$. The so derived expressions are then verified by experiments.

Self-Adaptation on the Sphere Model It is shown in [23, p. 306] that an $(1, \lambda)$-ES with self-adaptation convergences to the optimum log-linearly. Also the usually recommendation of choosing the learning rate proportionally to $1/\sqrt{N}$, where N is the search space dimensionality, is indeed approximately optimal. In the case of $(1, \lambda)$-ES, the dependency of the progress on the learning rate is weak provided that $\tau \geq c/\sqrt{N}$ with a constant c holds. As a result, it is not strictly necessary to have N-dependent learning parameters.

As has been shown in [23, p. 305], the time to adapt an ill-fitted mutation strength to the fitness landscape is proportionally to $1/\tau^2$. Adhering to the scaling rule $\tau \propto 1/\sqrt{N}$ results in an adaptation time that linearly increases with the search space dimensionality. Therefore, it is recommended to work with a *generation-dependent* or constant learning rate τ, respectively, if N is large.

The maximal progress rate that can be obtained in experiments is always smaller than the theoretical maximum predicted by the progress rate theory (without considering the stochastic process dynamics). The reason for this is that the fluctuations of the mutation strength degrade the performance. The average progress rate is deteriorated by a loss part stemming from the variance of the strategy parameter. The theory developed in [23] is able to predict this effect qualitatively.

If recombination is introduced in the algorithm the behavior of the ES changes qualitatively. Beyer and Grünz [51] showed that multi-recombinative ES that use intermediate or dominant recombination do not exhibit the same robustness with respect to the choice of the learning rate as $(1, \lambda)$-ES. Instead their progress in the stationary state has a clearly defined optimum and nearly optimal progress is only attainable for a relatively narrow range of the learning rate τ. If the learning rate is chosen suboptimally, the performance of the ES degrades but the ES still converges log-linearly to the optimum. The reason for this behavior [74] is due to the different effects recombination has on the distance to the optimizer (i.e., on the progress rate) and on the mutation strength. An intermediate recombination of the object variables reduces the harmful parts of the mutation vector also referred to as "genetic repair effect". Thus, it reduces the loss part of the progress rate. This enables the algorithm to work with higher mutation strengths. However, since the strategy parameters are necessarily selected before recombination takes place, the self-adaptation response cannot reflect the after selection genetic repair effect and remains relatively inert to the effect of recombination.

Flat and Linear Fitness Landscapes In [26], the behavior of multi-recombinative ES on flat and linear fitness landscapes was analyzed. Accepting the variance postulates proposed in [26] (see Section 2.3.3) the question arises whether the standard ES variation operators comply with these postulates, i.e., whether the strategies are able to increase the population variance in flat and linear fitness landscapes. Several common recombination operators and mutation operators were examined such as intermediate/dominant recombination of the object variables and intermediate/geometric recombination of the strategy parameters. The mutation rules applied for changing the mutation strength are the log-normal and the two-point distribution.

The analysis started with considering flat fitness landscapes which are selection neutral. Thus, the evolution of the mutation strength and the evolution of the object variables can be fully decoupled and the population variance can be easily computed. Beyer and Deb showed that if intermediate

recombination is used for the object variables, the ES is generally able to increase the population variance exponentially. The same holds for dominant recombination. However, there is a memory of the old population variances that gradually vanishes. Whether this is a beneficial effect has not been investigated up to now.

In the case of linear fitness functions, only the behavior of $(1, \lambda)$-ES has been examined so far. It has been shown that the results obtained in [23] for the sphere model can be transferred to the linear case if $\sigma^* := \sigma(N/R) \to 0$ is considered because the sphere degrades to a hyperplane. As a result, it can be shown that the expectation of the mutation strength increases exponentially if log-normal or two-point operators are used.

Beyond the Sphere Model: Ridge Functions

The self-adaptive behavior of evolution strategies on the ridge function class $f(\mathbf{y}) = y_1 - (\sum_{i=2}^{N} y_i^2)^{(\alpha/2)}$ was only addressed recently. Many analyses, e.g.,[5, 9] focus on the cumulative path length adaption rather than self-adaptation.

Lunacek and Whitley [72] presented an investigation of self-adaptive ES using the two-point rule for creating new mutation strength. They focused on $(1, \lambda)$-ES on two ridge function classes and provided experimental evidence for the conjecture

> "The global step-size of a self-adaptive $(1, \lambda)$-ES will stabilize when the selection of σ is unbiased toward larger or smaller values. If the ridge bias cannot be removed, self-adaptation will continue to decrease σ by selecting smaller step-sizes" [72].

To support this conjecture, they ran 100 trials of a $(1, 60)$-ES. In the experiments, different d-values, $d > 1$, were examined.

Very recently, Arnold and MacLeod [11] presented a comparison of several adaptation methods for ES analyzing the influence of noisy fitness evaluations. The self-adaptive ES investigated used the two-point rule to update the mutation strength. Furthermore, the mutation strengths were recombined using

$$\sigma' \leftarrow \sigma\Big(\prod_{m=1}^{\mu} \varsigma^{(m;\lambda)}\Big)^{\frac{1}{\mu\kappa}} \tag{2.24}$$

instead of the arithmetic recombination introduced in Section 1. The parameter κ is used to dampen the change of the mutation strength. Under some assumptions similar to the ones introduced by Lunacek and Whitley, they succeeded in deriving equations giving the stationary distance, mutation strength, and progress parallel to the axis direction. Among the results obtained are the following: Self-Adaptive ES fail in the creation of useful mutation strengths if $\mu \geq \lambda/2$ [11]. In addition, non–recombinative $(1, \lambda)$-ES are superior to recombinative ES. Compared to other adaptation means, e.g., CSA-ES, self-adaptation was found to perform worst of all.

2.5 Problems and Limitations of Self-Adaptation

Most of the research done so far seems to be centered on the effects of self-adapting the mutation strengths. Some of the problems that were reported refer to divergence and premature convergence of the algorithm (see, e.g., Kursawe [67]). Premature convergence may occur if the mutation strength and the population variance are decreased too fast. This generally results in a convergence towards a suboptimal solution. While the problem is well-known, it appears that only a few theoretical investigations have been done. However, premature convergence is not a specific problem of self-adaptation.

Rudolph [85] analyzed an $(1+1)$-ES applying Rechenberg's $1/5$th-rule. He showed for a test problem that the ES's transition to the global optimum cannot be ensured when the ES starts at a local optimum and if the step-sizes are decreased too fast.

Stone and Smith [100] investigated the behavior of GA on multimodal functions applying Smith's discrete self-adaptation algorithm. Premature convergence was observed for low innovation rates and high selection pressure since this combination causes a low diversity of the population. Diversity can be increased by using high innovation rates. Stone and Smith additionally opted for a scheme that passes through the present value of the strategy parameter while still introducing different choices thus providing a suitable relation between exploration and exploitation.

Liang et al. [69, 70] considered the problem of a prematurely reduced mutation strength. They started with an empirical investigation on the loss of step size control for EP on five benchmark functions [69]. The EP used a population size of $\mu = 100$ and a tournament size of $q = 10$. Stagnation of the search occurred even for the sphere model. As they argued, this might be due to the selection of an individual with a mutation strength far too small but with a high fitness value. This individual bequests its ill-adapted mutation strength to all descendants and, therefore, the search stagnates.

In [70], Liang et al. examined the probability of loosing the step size control. To simplify the calculations, a $(1 + 1)$-EP was considered. Therefore, the mutation strength changes whenever a successful mutation happens. A loss of step size control occurs if the mutation strength is smaller than an arbitrarily small positive number ϵ after κ successful mutations. The probability of such an event can be computed. It depends on the initialization of the mutation strength, the learning parameter, on the number of successful mutations, and on ϵ. As the authors showed, the probability of loosing control of the step size increases with the number of successful mutations.

A reduction of the mutation strength should occur if the EP is already close to the optimum. However, if the reduction of the distance to the optimizer cannot keep pace with that of the mutation strength, the search stagnates. This raises the question whether the operators used in this EP implementation comply with the design principles postulated in [39] (compare Section 2.3.3). An analysis of the EP behavior in flat or linear fitness landscapes might reveal the very reason for this failure. It should be noted also that similar premature convergence behaviors of self-adaptive ES are rarely observed. A way to circumvent such behavior is to introduce a lower bound for the step size. Fixed lower bounds are considered in [69]. While this surely prevents premature convergence of the EP, it does not take into account the fact that the ideal lower bound of the mutation strength depends on the actual state of the search.

In [70], two schemes are considered proposing a dynamic lower bound (DLB) of the mutation strength. The first is based on the success rate reminiscent of Rechenberg's $1/5$th-rule. The lower bound is adapted on the population level. A high success rate leads to an increase of the lower bound, a small success decreases it. The second DLB-scheme is called "mutation step size based" since it uses the median of the mutation strengths of all successful offspring as the next lower bound. These two schemes appear to work well on most fitness functions of the benchmark suite. On functions with many local optima, however, both methods experience difficulties.

As mentioned before, Hart, Delaurentis, and Ferguson analyzed an evolutionary algorithm with discrete random variables on a multi-modal function [55]. They showed the existence of a bimodal function for which the algorithm fails to converge to the global optimizer with probability one if it starts close to the local optimal solution.

Won and Lee [104] addressed a similar problem although in contrast to Hart, DeLaurentis, and Ferguson they proved sufficient conditions for premature convergence avoidance of a $(1 + 1)$-ES on a one-dimensional bimodal function. The mutations were modeled using Cauchy-distributed random variables and the two-point operator was used to change the mutation strengths themselves.

Glickman and Sycara [49] identified possible causes for premature reduction of the mutation strength. They investigated the evolutionary search behavior of a $(10, 100)$-EA without any crossover on a complex problem arising from the training of neural networks with recurrent connections.

What they have called *bowl effect* may occur if the EA is close to a local minimum. Provided that the mutation strength is below a threshold, the EA is confined in a local attractor and cannot find any better solution. As a result, small mutation strengths will be preferred.

A second cause is attributed to the selection strength. Glickman and Sycara suspect that if the selection strength is high, high mutation rates have a better chance of survival compared to using low selection strength: A high mutation rate increases the variance. This is usually connected with a higher chance of degradation as compared to smaller mutation rates. But if an improvement occurs it is likely to be considerably larger than those achievable with small mutation rates. If only a small percentage of the offspring is accepted, there is a chance that higher mutation strengths "survive". Thus, using a high selection strength might be useful in safeguarding against premature stagnation. In their experiments, though, Glickman and Sycara could not observe a significant effect. They attributed this in part to the fact that the search is only effective for a narrow region of the selection strength.

Recently, Hansen [52] resumed the investigation of the self-adaptive behavior of multiparent evolution strategies on linear fitness functions started in [39]. Hansen's analysis is aimed at revealing the causes why self-adaptation usually works adequately on linear fitness functions. He offered conditions under which the control mechanism of self-adaptation fails, i.e., that the EA does not increase the step size as postulated in [39]. The expectation of the mutation strength is not measured directly. Instead, a function h is introduced the expectation of which is unbiased under the variation operators. The question that now remains to be answered is whether the selection will increase the expectation of $h(\sigma)$. In other words, is the effect of an increase of the expectation a consequence of selection (and therefore due to the link between good object vectors and good strategy values) or is it due to a bias introduced by the recombination/mutation-operators chosen?

Hansen proposes two properties an EA should fulfill: First, the descendants' object vectors should be point-symmetrically distributed after mutation and recombination. Additionally, the distribution of the strategy parameters given the object vectors after recombination and mutation has to be identical for all symmetry pairs around the point-symmetric center. Evolution strategies with intermediate multirecombination fulfill this symmetry assumption. Their descendents' distribution is point-symmetric around the recombination centroid.

Secondly, Hansen offers a so-called *σ-stationarity assumption*. It postulates the existence of a monotonically increasing function h whose expectation is left unbiased by recombination and mutation. Therefore, $\mathrm{E}[h(\mathcal{S}_k^{\sigma_{i;\lambda}|i=1,\ldots,\mu})] = (1/\mu)\sum_{i=1}^{\mu} h(\sigma_{i;\lambda})$ must hold for all offspring. The term $\mathcal{S}_k^{\sigma_{i;\lambda}|i=1,\ldots,\mu}$ denotes the mutation strength of an offspring k created by recombination and mutation.

Hansen showed that if an EA fulfills the assumptions made above, self-adaptation does not change the expectation of $h(\sigma)$ provided that the offspring number is twice the number of parents.

The theoretical analysis was supplemented by an empirical investigation of the self-adaptation behavior of some evolution strategies examining the effect of several recombination schemes on the object variables and on the strategy parameter. It was shown that an ES which applies intermediate recombination to the object variables and to the mutation strength increases the expectation of $\log(\sigma)$ for all choices of the parent population size. On the other hand, evolution strategies that fulfill the symmetry and the stationarity assumption, increase the expectation of $\log(\sigma)$ if $\lambda < \mu/2$, keep it constant for $\lambda = \mu/2$ and decrease it for $\lambda > \mu/2$.

Intermediate recombination of the mutation strengths results in an increase of the mutation strength. This is beneficial in the case of linear problems and usually works as desired in practice. However, as

Hansen states the presence of a bias may entail “the danger of divergence or premature convergence” [52].

2.6 Conclusions

Self-adaptation usually refers to an adaptation of control parameters which are incorporated into an individual’s genome. These are subject to variation and selection – evolving together with the object parameters. Stating it more generally: A self-adaptive algorithm controls the transmission function between parent and offspring population by itself without any external influence. For this reason the concept can be broadened to include algorithms where the representation of an individual is augmented with genetic information that does not code information regarding the fitness but influences the transmission function instead. Interestingly, real-coded genetic algorithms where the diversity of the parent population controls that of the offspring may be regarded as self-adaptive. Surprisingly, even binary genetic algorithms with crossover operators like 1-point or k-point crossover share this property to a certain extent.

Self-Adaptation is common in the area of evolutionary programming and evolution strategies. Here, generally the mutation strength or the full covariance matrix is adapted. Analyses conducted so far focus mainly on the convergence to the optimal solution. Nearly all analyses use either a simplified model of the algorithm or have to resort to numerical calculations in their study. The results obtained are similar: On simple fitness functions, conditions can be derived that ensure the convergence of the EA to local optimal solutions. The convergence is usually log-linear.

The explicit use of self-adaptation techniques is rarely found in genetic algorithm and if at all mainly used to adopt the mutation rate. Most of the studies found are directed at finding suitable ways to introduce self-adaptive behavior in GA. As we have pointed out, however, crossover in binary standard GA does provide a rudimentary form of self-adaptive behavior. Therefore, the mutation rate can be often kept at a low level provided that the population size is reasonably large. However, unlike the clear goals in real-coded search spaces, it is by no means obvious to formulate desired behaviors the self-adaptation should realize in binary search spaces. This does not apply to some real-coded genetic algorithms where it can be shown mathematically that they can exhibit self-adaptive behavior in simple fitness landscapes.

It should be noted that self-adaptation techniques are not the means to solve all adaptation problems in evolutionary algorithms. Concerning evolution strategies, multi-recombinative self-adaptation strategies are sensitive to the choice of the external learning rate τ. As a result, an optimal or a nearly optimal mutation strength is not always realized.

More problematic appears a divergence or a premature convergence to a suboptimal solution. The latter is attributed to a too fast reduction of the mutation strength. Several reasons for that behavior have been proposed although not rigorously investigated up to now. However, from our own research we have found that the main reason for a possible failure is due to the opportunistic way how self-adaptation uses the selection information obtained from just one generation. Self-adaptation rewards short-term gains. In its current form, it cannot look ahead. As a result, it may exhibit the convergence problems mentioned above.

Regardless of the problems mentioned, self-adaptation is a state-of-the-art adaptation technique with a high degree of robustness, especially in real-coded search spaces and in environments with uncertain or noisy fitness information. It also bears a large potential for further developments both in practical applications and in theoretical as well as empirical evolutionary computation research.

3 Analyzing Self-Adaptive Evolution Strategies

In this chapter, the evolution equations – the approach used in the analysis of self-adaptive ES in this thesis – are described in greater detail. The approach was first introduced in [21]. Before the dynamics of evolution strategies can be analyzed, the variables that characterize the system must be determined. In other words, the state variables need to be given. Considering ES, one might be interested in monitoring the fitness values, the distance to the optimizer (depending on the fitness model), and, since self-adaptation is considered, the mutation strength. The approach then aims at modeling and analyzing the evolution of these state variables over time. In the following, the sphere model is used for further explanations. Since the sphere model consists of functions of the form $f(\mathbf{y})=g(\|\mathbf{y}-\hat{\mathbf{y}}\|) = g(R)$, the state variables are chosen as the distance to the optimizer $R^{(g)} = \|\mathbf{y}^{(g)}-\hat{\mathbf{y}}\|$ and the mutation strength $\varsigma^{(g)}$ at generation g. The dynamics of $(\mu/\mu_I,\lambda)$-ES generate a stochastic process

$$\begin{pmatrix} R^{(g)} \\ \varsigma^{(g)} \end{pmatrix} \rightarrow \begin{pmatrix} R^{(g+1)} \\ \varsigma^{(g+1)} \end{pmatrix}. \tag{3.1}$$

As mentioned in Chapter 2, up to now no closed solution for the transition kernels could be derived in general. The only exception is a $(1,2)$-ES using the two-point rule for the mutation of the mutation strength (see [21] of [23, p. 287].

In this thesis, therefore, the step-by-step approach introduced in [21] is followed. The approach relies on the evolution equations. These are stochastic difference equations or iterative maps, respectively, used to describe the change of the state variables during one generation. The change of the random variables can be divided into two parts: The first denotes the expected change. The second part covers the random fluctuations and is denoted by ϵ_R or ϵ_σ. In their most general form, the evolution equations read

$$R^{(g+1)} = R^{(g)} - \mathrm{E}[R^{(g)} - R^{(g+1)}|R^{(g)},\sigma^{(g)}] + \epsilon_R(R^{(g)},\varsigma^{(g)}) \tag{3.2}$$

$$\varsigma^{(g+1)} = \varsigma^{(g)}\left(1 + \mathrm{E}\left[\frac{\varsigma^{(g+1)}-\varsigma^{(g)}}{\varsigma^{(g)}}|R^{(g)},\sigma^{(g)}\right]\right) + \epsilon_\sigma(R^{(g)},\varsigma^{(g)}). \tag{3.3}$$

In (3.2), a well known progress measure appears: the progress rate φ_R. The progress rate measures the expected change of the distance in one generation

$$\varphi_R(\varsigma^{(g)},R^{(g)}) := \mathrm{E}[R^{(g)} - R^{(g+1)}|\varsigma^{(g)},R^{(g)}]. \tag{3.4}$$

The progress rate is an example for a so-called local performance measure – local because it depends on the present state of the system.

In the case of the evolution of the mutation strength, a different progress measure is used. Note, since the mutation of the mutation strength is generally realized by a multiplication with a random

variable, the equation in (3.3) gives the relative change. The progress measure is called the (first-order) self-adaptation response (SAR) ψ. The SAR gives the expected relative change of the mutation strength in one generation

$$\psi(\varsigma^{(g)}, R^{(g)}) := \mathrm{E}\left[\frac{\varsigma^{(g+1)} - \varsigma^{(g)}}{\varsigma^{(g)}} | \varsigma^{(g)}, R^{(g)}\right]. \tag{3.5}$$

Let us now address the fluctuation terms. Their distribution is not known and must be approximated using a reference density. Note, given pdfs p_1 and p_2, it is possible to relate p_i to p_j in general (see, e.g., [66, 32]). Common approaches comprise an expansion into a Gram-Charlier or Edgeworth series. The reference distribution is usually (but not necessarily) chosen to be the normal distribution. In order to expand an unknown distribution at all, it must be possible to determine some of its moments or cumulants.

First of all, the fluctuation terms are standardized using the expected value and standard deviation. Clearly, the conditional expectation of ϵ_σ and ϵ_R is zero. Therefore, only the standard deviation remains to be determined.

The main points of the derivation are explained considering the case of ϵ_R. The case of the mutation strength may be treated analogously. Let D_φ denote the standard deviation. Therefore the standardized random part ϵ'_R is related to ϵ_R by $\epsilon_R = D_\varphi \epsilon'_R$. The standard deviation can be derived via (3.2) since its square equals the second conditional moment of ϵ_R

$$\begin{aligned}
D^2_\varphi(\varsigma^{(g)}, R^{(g)}) &= \mathrm{E}[\epsilon^2_R | \varsigma^{(g)}, R^{(g)}] \\
&= \mathrm{E}\Big[\Big(R^{(g+1)} - R^{(g)} + \varphi_R(\varsigma^{(g)}, R^{(g)})\Big)^2 | \varsigma^{(g)}, R^{(g)}\Big] \\
&= \mathrm{E}\Big[\Big(R^{(g+1)} - R^{(g)}\Big)^2 - 2\Big(R^{(g)} - R^{(g+1)}\Big)\varphi_R(\varsigma^{(g)}, R^{(g)}) \\
&\quad + \varphi^2_R(\varsigma^{(g)}, R^{(g)}) | \varsigma^{(g)}, R^{(g)}\Big] \\
&= \mathrm{E}\Big[\Big(R^{(g+1)} - R^{(g)}\Big)^2 | \varsigma^{(g)}, R^{(g)}\Big] - \varphi^2_R(\varsigma^{(g)}, R^{(g)}).
\end{aligned} \tag{3.6}$$

The distribution of ϵ'_R is expanded into an Edgeworth series. For the analysis, the expansion is cut off after the first term (cf. [23, p.265]). That is to say, it is supposed that the deviations from the normal distribution are negligible in the analysis scenario the equations will be applied to. The random variable ϵ_R reads

$$\epsilon_R = D_\varphi(\varsigma^{(g)}, R^{(g)})\mathcal{N}(0,1) + \ldots \tag{3.7}$$

The expectation $\mathrm{E}[(R^{(g+1)} - R^{(g)})^2 | \varsigma^{(g)}, R^{(g)}]$ appearing in (3.6) is called the second-order progress rate

$$\varphi^{(2)}(\varsigma^{(g)}, R^{(g)}) = \mathrm{E}\Big[\Big(R^{(g+1)} - R^{(g)}\Big)^2 | \varsigma^{(g)}, R^{(g)}\Big]. \tag{3.8}$$

The random variable ϵ_σ is obtained similarly. As in the case of the distance, a first-order approach (i.e., the first term of the series expansion) is used

$$\epsilon_\sigma = D_\psi(\varsigma^{(g)}, R^{(g)})\mathcal{N}(0,1) + \ldots. \tag{3.9}$$

The derivation of the standard deviation is exactly the same as previously. We have

$$\begin{aligned} D_\psi^2(\varsigma^{(g)}, R^{(g)}) &= \mathrm{E}[\epsilon_\sigma^2|\varsigma^{(g)}, R^{(g)}] \\ &= \mathrm{E}\Big[\Big(\varsigma^{(g+1)} - \varsigma^{(g)} - \varsigma^{(g)}\psi(\varsigma^{(g)}, R^{(g)})\Big)^2|\varsigma^{(g)}, R^{(g)}\Big] \\ &= \mathrm{E}\Big[\Big(\varsigma^{(g+1)} - \varsigma^{(g)}\Big)^2 - 2\varsigma^{(g)}\Big(\varsigma^{(g+1)} - \varsigma^{(g)}\Big)\psi(\varsigma^{(g)}, R^{(g)}) \\ &\quad +\psi(\varsigma^{(g)}, R^{(g)})^2|\varsigma^{(g)}, R^{(g)}\Big] \\ &= \mathrm{E}\Big[\Big(\varsigma^{(g+1)} - \varsigma^{(g)}\Big)^2 - (\varsigma^{(g)})^2\psi^2(\varsigma^{(g)}, R^{(g)})|\varsigma^{(g)}, R^{(g)}\Big] \end{aligned} \tag{3.10}$$

(cf. 3.3). Again, this introduces a new measure, the second-order SAR

$$\psi^{(2)}(\varsigma^{(g)}, R^{(g)}) := \mathrm{E}\Big[\Big(\frac{\varsigma^{(g+1)} - \varsigma^{(g)}}{\varsigma^{(g)}}\Big)^2|\varsigma^{(g)}, R^{(g)}\Big]. \tag{3.11}$$

Using the results obtained so far, the evolution equations can be rewritten to

$$R^{(g+1)} = R^{(g)} - \varphi_R(\varsigma^{(g)}, R^{(g)}) + D_\varphi(\varsigma^{(g)}, R^{(g)})\mathcal{N}(0,1) + \ldots \tag{3.12}$$

$$\begin{aligned} \varsigma^{(g+1)} &= \varsigma^{(g)}\Big(1 + \psi(\varsigma^{(g)}, R^{(g)})\Big) + D_\psi(\varsigma^{(g)}, R^{(g)})\mathcal{N}(0,1) + \ldots \\ &= \varsigma^{(g)}\Big(1 + \psi(\varsigma^{(g)}, R^{(g)}) + D'_\psi(\varsigma^{(g)}, R^{(g)})\mathcal{N}(0,1) + \ldots\Big) \end{aligned} \tag{3.13}$$

with $D'_\psi = D_\psi/\varsigma^{(g)}$.

The Deterministic Evolution Equations

In this thesis, the fluctuation parts are neglected in most cases with the exception of Section 4.4. The evolution equations without perturbation parts are generally termed deterministic evolution equations [23]. This approach –though rather crude– serves well to extract the general characteristics of self-adaptive evolution strategies. The deterministic evolution equations read

$$R^{(g+1)} = R^{(g)} - \mathrm{E}[R^{(g)} - R^{(g+1)}] \tag{3.14}$$

$$\varsigma^{(g+1)} = \varsigma^{(g)}\Big(1 + \mathrm{E}\Big[\frac{\varsigma^{(g+1)} - \varsigma^{(g)}}{\varsigma^{(g)}}\Big]\Big). \tag{3.15}$$

An equilibrium (steady state, or stationary state) is characterized by $R^{(g+1)} = R^{(g)}$ and $\varsigma^{(g+1)} = \varsigma^{(g)}$. Note, demanding stationarity of the $R^{(g)}$-evolution equals a complete standstill of the ES in most cases. Often more interesting is the evolution equation of the normalized mutation strength $\varsigma^{*(g)} = \varsigma^{(g)}(N/R^{(g)})$

$$\varsigma^{*(g+1)} = \varsigma^{*(g)}\left(\frac{1 + \psi(\varsigma^{*(g)}, R^{(g)})}{1 - \frac{\varphi_R^*(\varsigma^{*(g)}, R^{(g)})}{N}}\right) \tag{3.16}$$

with $\varphi_R^* = \varphi_R(N/R^{(g)})$ and $\varsigma^{*(g+1)} = \varsigma^{(g+1)}(N/R^{(g+1)})$ since it allows for a stationary state without requiring a stationary state of the $R^{(g)}$-evolution. The assumption of the existence of a stationary state is motivated by findings that it is optimal in many cases for the mutation strength to scale with the distance to the optimizer. Optimal in this case refers to a local progress measure, i.e., to a maximal expected gain during one generation.

Including the Fluctuations

If the perturbation parts are included in the analysis, the situation becomes more complicated. Equations (3.12) and (3.13) describe a Markov process, the transition densities p_{tr} of which have to be determined. The variables $R^{(g+1)}$, $R^{(g)}$, $\varsigma^{(g+1)}$, and $\varsigma^{(g)}$ are now all random variables. Assuming that the distribution of each has a density, the density of the distance r at generation g is denoted with $p(R^{(g)})$ and the density of the mutation strength with $p(\varsigma^{(g)})$. As pointed out in [23, p. 313], it generally suffices not to determine the complete distribution but to concentrate on some of the moments, generally the expectation of course. The expectations read

$$\begin{aligned}
\overline{R^{(g+1)}} &= \int_0^\infty R^{(g+1)} p(R^{(g+1)})\,\mathrm{d}R^{(g+1)} \\
&= \int_0^\infty \int_0^\infty \int_0^\infty \Big(R^{(g)} - \varphi_R(R^{(g)}, \varsigma^{(g)})\Big) \\
&\quad \times p_R(R^{(g+1)}|\varsigma^{(g)}, R^{(g)}) p(\varsigma^{(g)}) p(R^{(g)})\,\mathrm{d}\varsigma^{(g+1)}\,\mathrm{d}R^{(g)}\,\mathrm{d}R^{(g+1)} \\
&= \int_0^\infty \int_0^\infty \Big(R^{(g)} - \varphi_R(R^{(g)}, \varsigma^{(g)})\Big) p(\varsigma^{(g)}) p(R^{(g)})\,\mathrm{d}R^{(g)}\,\mathrm{d}\varsigma^{(g)} \qquad (3.17) \\
\overline{\varsigma^{(g+1)}} &= \int_0^\infty \varsigma^{(g+1)} p(\varsigma^{(g+1)})\,\mathrm{d}\varsigma^{(g+1)} \\
&= \int_0^\infty \int_0^\infty \int_0^\infty \varsigma^{(g)} \Big(1 + \psi(\varsigma^{(g)}, R^{(g)})\Big) \\
&\quad \times p_\sigma(\varsigma^{(g+1)}) p(\varsigma^{(g)}) p(R^{(g)})\,\mathrm{d}\varsigma^{(g)}\,\mathrm{d}R^{(g)}\,\mathrm{d}\varsigma^{(g+1)} \\
&= \int_0^\infty \int_0^\infty \varsigma^{(g)} \Big(1 + \psi(\varsigma^{(g)}, R^{(g)})\Big) p(\varsigma^{(g)}) p(R^{(g)})\,\mathrm{d}\varsigma^{(g)}\,\mathrm{d}R^{(g)}. \qquad (3.18)
\end{aligned}$$

As can be inferred from (3.17) and (3.18), the transition densities are not needed if only the expectations are to be determined.

An equilibrium of a stochastic process is then characterized by a convergence of the densities to an equilibrium distribution, i.e., $\lim_{g\to\infty} p(\varsigma^{(g+1)}) = \lim_{g\to\infty} p(\varsigma^{(g)}) = p_\infty(\varsigma)$. Note, again it is normally the normalized mutation strength which converges towards an equilibrium as long as the ES progresses still. If a stationary state is reached, the invariant density solves the eigenvalue equation

$$c p_\infty(\varsigma) = \quad = \int_0^\infty p_{tr}(\varsigma|\sigma) p_\infty(\sigma)\,\mathrm{d}\sigma \qquad (3.19)$$

with $c = 1$ and p_{tr} the transition density. In general, the equilibrium distribution p_∞ is unknown. As pointed out in [23, p. 318], it is possible to determine p_∞ numerically or even analytically. The results, however, tend to be quite complicated and do not allow for further analytical treatment. Instead of trying to obtain the distribution itself, the expected value is obtained analyzing thus the mean value dynamics of the system. Unfortunately, the form of the evolution equations hinders a direct determination of the expectation since in general lower order moments depend on higher order moments leading to a non-ending recursion.

Therefore, a so-called ansatz is used [23, p. 319]: Instead of determining the solution of (3.19), the equilibrium distribution is set to a known (similar) distribution. This approach is reminiscent of the Edgeworth or Gram-Charlier expansion. The ansatz distribution takes the place of the baseline density and the expansion is cut off after the very first term.

Generally, the equations obtained are non-linear and can be solved only numerically. Special cases may exist, though, which allow for an analytical treatment.

In the following chapters, the evolution equations are applied to self-adaptive ES in two fitness environments: the sphere model and ridge functions. In both cases, the analysis is divided into two parts. First, the undisturbed fitness function is analyzed before noisy fitness evaluations are taken into account.

4 Self-Adaptation on the Sphere Model

The investigation of self-adaptive ES is started with the sphere model.

Definition 1. A function $f : \mathbb{R}^N \to \mathbb{R}$ is called a sphere (model) if

$$f(\mathbf{y}) \quad = \quad g(\|\mathbf{y} - \hat{\mathbf{y}}\|) \tag{4.1}$$

with $g : \mathbb{R} \to \mathbb{R}$ a monotonously in- or decreasing function, $\hat{\mathbf{y}}$ the optimizer of f, and $\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{N} x_i^2}$ the Euclidean norm on $\mathbb{R}^N$. □

The sphere only depends on the distance to the optimizer. It generally serves to model more general fitness functions in the vicinity of the optimum.

The analysis presented here can be seen as an extension of the analysis first carried out in [21] broadening the subject of the analysis from non-recombinative evolution strategies to evolution strategies using intermediate recombination on the one hand and to noisy fitness evaluations on the other.

4.1 Self-Adaptation and Intermediate Recombination

Self-Adaptation was originally proposed for non-recombinative $(1, \lambda)$-evolution strategies as a means to adapt the mutation strength. Recall, the mutation strength is treated in a similar manner as the object parameters. Therefore, it is subject to variation and selection. The random change is realized by a multiplication with a random variable. Common choices of distribution functions for this random variable include, e.g., the log-normal distribution. Here given the parental σ, the new mutation strength σ' of an offspring is generated according to

$$\sigma' \quad = \quad \sigma \mathrm{e}^{\tau \mathcal{N}(0,1)} \tag{4.2}$$

as mentioned in Chapter 2. The parameter τ is referred to as the learning rate. Another common choice is the symmetric two-point distribution with

$$\sigma' \quad = \quad \begin{cases} \sigma(1+\beta) & \text{if } u \leq 0.5 \\ \sigma/(1+\beta) & \text{if } u > 0.5 \end{cases} . \tag{4.3}$$

The random variable u follows a uniform distribution on $(0, 1]$. Both distributions – the log-normal and the two-point distribution – depend on one free parameter. The choice of this parameter influences the performance of ES. Therefore, one of the first questions to be asked is how τ (or β) is to be chosen so that the ES progresses with optimal speed. For $(1, \lambda)$-ES on the sphere model, this question is already answered: It is optimal to choose $\tau \propto 1/\sqrt{N}$ [21]. Apart from this condition, self-adaptation in $(1, \lambda)$-ES is remarkably robust with regard to the learning rate. Interestingly, this does not hold anymore once recombination comes into play [51]. The reasons for this behavior are investigated in this section.

4.1.1 Modeling the Self-Adaptive ES

To analyze the ES variables are needed to characterize the behavior. Since the sphere is considered, the fitness functions are of the form $f(\mathbf{y}) = g(\|\mathbf{y} - \hat{\mathbf{y}}\|)$, with optimizer $\hat{\mathbf{y}}$. Therefore, two variables suffice for the analysis: the distance to the optimizer, i.e., $R^{(g)} := \|\langle \mathbf{y}^{(g)} \rangle - \hat{\mathbf{y}}\|$ and the mutation strength $\langle \sigma^{(g)} \rangle$ at generation g. The evolution equations introduced in Chapter 3 are used to describe the change of these state variables from one generation to the next. Remember, the change is divided into an expected change and into a random perturbation part. Using the state variables $R^{(g)}$ for the distance of the centroid of the parental population to the optimizer and $\langle \varsigma^{(g)} \rangle$ for the mean of the mutation strengths at generation g, the evolution of the ES can be described by

$$\begin{pmatrix} R^{(g+1)} \\ \langle \varsigma^{(g+1)} \rangle \end{pmatrix} = \begin{pmatrix} R^{(g)} - \varphi_R(\langle \varsigma^{(g)} \rangle, R^{(g)}) + \epsilon_R^{(g)} \\ \langle \varsigma^{(g)} \rangle \Big(1 + \psi(\langle \varsigma^{(g)} \rangle, R^{(g)})\Big) + \epsilon_\sigma^{(g)} \end{pmatrix}. \tag{4.4}$$

The deterministic changes of the variables are given by the progress rate

$$\varphi_R(\langle \varsigma^{(g)} \rangle, R^{(g)}) := \mathrm{E}\left[R^{(g)} - R^{(g+1)} | \langle \varsigma^{(g)} \rangle, R^{(g)}\right] \tag{4.5}$$

in the case of the distance and in the case of the mutation strength by the self-adaptation response function (SAR)

$$\psi(\langle \varsigma^{(g)} \rangle, R^{(g)}) := \mathrm{E}\left[\frac{\langle \varsigma^{(g+1)} \rangle - \langle \varsigma^{(g)} \rangle}{\langle \varsigma^{(g)} \rangle} | \langle \varsigma^{(g)} \rangle, R^{(g)}\right] \tag{4.6}$$

whereas $\epsilon_R^{(g)}$ and $\epsilon_\sigma^{(g)}$ denote the random fluctuations.

To start the analysis, the perturbation parts of (4.4) are neglected. Furthermore, the notations are simplified. Unless the dependence on the generation number is explicitely needed, let $R := R^{(g)}$, $r := R^{(g+1)}$, and $\sigma := \langle \varsigma^{(g)} \rangle$. Finally, the usual normalizations are introduced to eliminate the R-dependency of the equations with $\sigma^* := \sigma\,(N/R)$, $\langle \varsigma^{*(g+1)} \rangle := \langle \varsigma^{(g+1)} \rangle\,(N/r)$, and $\varphi_R^* := \varphi_R\,(N/R)$.

From this point, the normalized system

$$\begin{pmatrix} r \\ \langle \varsigma^{*(g+1)} \rangle \end{pmatrix} = \begin{pmatrix} R\Big(1 - \varphi_R^*(\sigma^*)/N\Big) \\ \sigma^* \Big(\frac{1+\psi(\sigma^*)}{1-\varphi_R^*(\sigma^*)/N}\Big) \end{pmatrix} \tag{4.7}$$

of the deterministic evolution equations serves as the starting point of our analysis. Before continuing, the progress rate (4.5) and the self-adaptation response function (4.6) need to be determined. The progress rate $\varphi_R^* = (N/R)\mathrm{E}[R - r]$ is given for $\tau = 0$ and $N \to \infty$ by

$$\varphi_R^*(\sigma^*) = c_{\mu/\mu,\lambda}\sigma^* - \frac{\sigma^{*2}}{2\mu}. \tag{4.8}$$

The derivation of (4.8) can be found in Appendix B.1.2 with $\sigma_\epsilon^* = 0$ or in [23]. The self-adaptation response (SAR) is obtained in Appendix C.1.1. For $N \to \infty$ and $\tau \ll 1$ it is given by

$$\psi(\sigma^*) = \tau^2\Big(1/2 + e_{\mu,\lambda}^{1,1} - c_{\mu/\mu,\lambda}\sigma^*\Big). \tag{4.9}$$

The coefficients $c_{\mu/\mu,\lambda}$ and $e_{\mu,\lambda}^{1,1}$ are special cases of the so-called generalized progress coefficients [23, p. 172] $e_{\mu,\lambda}^{\alpha,\beta}$ (A.24).

The approximation errors made by using (4.8) and (4.9) diminish for increasing N and decreasing τ. Therefore, the analysis is restricted to ES operating in high-dimensional search spaces and to small learning rates τ.

Before continuing, it is important to note a result first obtained in [21]:

"The evolution of the mutation strength can be decoupled from that of the distance."

Why this is the case can immediately be inferred from the form of (4.7), (4.8), and (4.9): There is no direct influence of R on the evolution of the normalized mutation strength. The evolution of the mutation strength can be considered and analyzed isolated. This does not hold for the evolution of R which is directly influenced by σ^*.

4.1.2 Analyzing the Stationary Points

Considering (4.7), the behavior of the ES is described by deterministic difference equations or by an iterated map. Using the theory of dynamic systems [103], one of the first questions to be raised is whether the system admits stationary points. The analysis of stationary points has an additional justification: The ES should strive to operate with the best mutation strength it can achieve. The size of the mutation strength obviously depends on the position in the search space, i.e., on the distance to the optimizer in the case of the sphere.

Definition 2. Let $f : \mathbb{R}^N \to \mathbb{R}^M$. Stationary points or fixed points (fix-points, equilibrium solutions, stationary solutions) $\mathbf{y}_S$ of the difference equation (or iterated map) $\mathbf{y}^{(t+1)} = f(\mathbf{y}^{(t)})$ are given by $\mathbf{y}_S = f(\mathbf{y}_S)$. □

Stationary points are time-invariant solutions of the dynamic system. If the system reaches a fixed point, it comes to a halt and no movement occurs – unless the system is perturbed. As can be seen easily and will be shown below, system (4.7) as a whole does not admit stationary points unless very specific situations occur. Seen isolated, the evolution equations for the mutation strength and the distance admit stationary points, though.

Let us start with the mutation strength and consider system (4.7) and Eqs. (4.8) (progress rate) and (4.9) (SAR). Stationary points of the σ^*-evolution of (4.7) that is points for which

$$\begin{aligned} \langle \varsigma^{*(g+1)} \rangle = \sigma^* \quad &\Leftrightarrow \quad \sigma^* = 0 \bigvee \frac{1+\psi}{1-\frac{\varphi_R^*}{N}} = 1 \\ &\Leftrightarrow \quad \sigma^* = 0 \bigvee c_{\mu/\mu,\lambda}\sigma^* - \frac{\sigma^{*2}}{2\mu} = -N\tau^2\Big(1/2 + e_{\mu,\lambda}^{1,1} - c_{\mu/\mu,\lambda}\sigma^*\Big) \end{aligned} \tag{4.10}$$

holds (see (4.8) and (4.9)) are given by $\varsigma^*_{stat_1} = 0$ or by

$$\varsigma^*_{stat_2} \quad = \quad \mu c_{\mu/\mu,\lambda}\left((1-N\tau^2) + \sqrt{(1-N\tau^2)^2 + 2N\tau^2\frac{1/2 + e_{\mu,\lambda}^{1,1}}{\mu c_{\mu/\mu,\lambda}^2}}\right). \tag{4.11}$$

The detailed derivation of the stationary points can be found in Appendix D.1.1. Stationary points are characterized by either a loss of step-size control or by a mutation strength which is a function of the learning rate τ (if the other parameters are considered to be fixed). Therefore, the learning rate can be used to calibrate the value of the non-zero stationary mutation strength.

The stationary points of the R-evolution remain to be addressed. To this end, system (4.7) and Eq. (4.8) have to be considered. Fixed points of the R-evolution, i.e., points for which

$$r = R \quad \Leftrightarrow \quad R = 0 \bigvee \varphi_R^* = c_{\mu/\mu,\lambda}\sigma^* - \frac{\sigma^{*2}}{2\mu} = 0 \tag{4.12}$$

holds, are then given by $(R, \sigma^*)^{\mathrm{T}} = (0, c)^{\mathrm{T}}$, $(R, \sigma^*)^{\mathrm{T}} = (c, 0)^{\mathrm{T}}$ with $c \in \mathbb{R}, c \geq 0$ or by $(R, \varsigma^*_{\varphi_{R_0}})^{\mathrm{T}}$ with $R > 0$ and

$$\varsigma^*_{\varphi_{R_0}} \quad = \quad 2\mu c_{\mu/\mu,\lambda}. \tag{4.13}$$

Stationary solutions of system (4.7) are thus characterized as follows:

1. A loss of step-size control occurs in an arbitrary distance to the optimizer,
2. the optimum is reached, or
3. the second stationary solution of the σ^*-evolution (4.11) and the second stationary point of the R-evolution (4.13) match.

The question remain whether these possibilities actually occur and if (4.7) admits them whether they are stable solutions.

It is easy to show that the first possibility: a loss of step-size control leads to an instable fixed point. In other words, if the system is in the fixed point $\varsigma^*_{stat_1} = 0$ and small perturbations occur, it will move away from it. Let us first recall the definition of asymptotic stability.

Definition 3. Let $f : \mathbb{R}^N \to \mathbb{R}^N$ and $\mathbf{y}_S \in \mathbb{R}^N$ a fixed point of $\mathbf{y}^{(t+1)} = f(\mathbf{y}^{(t)})$. The fixed point is called (locally) asymptotically stable if an $\epsilon > 0$ exists so that for all $\Delta^{(t)}$ with $\|\Delta^{(0)}\| < \epsilon$, $\Delta^{(t)} = \mathbf{y}^{(t)} - \mathbf{y}_S$

$$\lim_{t\to\infty} \Delta^{(t)} \quad = \quad \lim_{t\to\infty} \mathbf{y}^{(t)} - \mathbf{y}_S = 0 \tag{4.14}$$

holds. In other words: After a perturbation, the system returns to the equilibrium provided that the perturbation is sufficiently small. □

A well established means to show the locally asymptotic stability is via the linear approximation using the Taylor series (see, e.g., [103, 71]).

Lemma 1. *Let $f : \mathbb{R}^N \to \mathbb{R}^M$ be a twice continuously differentiable function. Then it follows*

$$\Delta^{(t+1)} = Df(\mathbf{y})|_{\mathbf{y}=\mathbf{y_S}}(\Delta^{(t)}) + \mathcal{O}({\Delta^{(t)}}^{\mathrm{T}}\Delta^{(t)}). \tag{4.15}$$

□

Provided that the fixed point is hyperbolic (i.e., no eigenvalue has a real part of ± 1) the stability of the fixed point can be established considering the linear system. To this end, the Jacobian matrix $Df(\mathbf{y}_S)$ at $\mathbf{y}_S$ must be obtained and analyzed.

Lemma 2. *Consider an iterated map. A hyperbolic fixed point $\mathbf{y_S}$ is stable if the absolute value of the real part of all eigenvalues is smaller than one. It is instable if the absolute value of the real part of one eigenvalue is greater than one (see, e.g., [103]).*□

It is easy to see that the stationary solution $\varsigma^*_{stat_1} = 0$ is an unstable fixed point of the evolution equation of the mutation strength in (4.7): To this end, the first derivative of $f(\sigma^*) = \sigma^*[(1 + \psi(\sigma^*))/(1 - \varphi^*_R/N)]$ must be determined. First of all, note that f is $C^2(U)$ for a ball $U(0)$ The derivative is easily obtained as

$$f'(\sigma^*) = \frac{1+\psi(\sigma^*)}{1-\varphi^*_R(\sigma^*)/N} + \sigma^*\Big(\frac{\psi'(\sigma^*)}{1-\varphi^*_R(\sigma^*)/N} + \frac{(1+\psi(\sigma^*))\varphi^{*\prime}_R(\sigma^*)/N}{(1-\varphi^*_R(\sigma^*)/N)^2}\Big). \quad (4.16)$$

Inserting the fixed point, $f'(\sigma^*)|_{\sigma^*=0} = 1 + \tau^2(1/2 + e^{1,1}_{\mu,\lambda})$ is obtained, which is greater than one as long as $\tau > 0$. The fixed point $\varsigma^*_{stat_1} = 0$ is therefore unstable.

The last possibility – an intersection of the second stationary solution of the σ^*-evolution (4.11) with the second stationary solution of the R-evolution (4.13) does not occur for finite τ. As already noted in [23, p. 300] for $(1, \lambda)$-ES and also revealed by (4.10), the τ-parameter steers the stationary point (4.12) between the zero of the SAR

$$\varsigma^*_{\psi_0} = \frac{1/2 + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}} \quad (4.17)$$

and the second zero (4.13) of the progress rate: For $N\tau^2 \to \infty$, (4.12) goes to (4.17), whereas for $N\tau^2 \to 0$, (4.12) approaches the zero of the progress rate (4.13). It can be shown by case inspection, that

$$\varsigma^*_{\psi_0} = \frac{1/2 + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}} < \varsigma^*_{\varphi_{R_0}} = 2\mu c_{\mu/\mu,\lambda} \quad (4.18)$$

expect for $\mu \approx \lambda$. That is, the zero of the SAR is smaller than the zero of the progress rate.

The stationary mutation strength (4.11), p. 35,

$$\varsigma^*_{stat_2} = \mu c_{\mu/\mu,\lambda}\left((1-N\tau^2) + \sqrt{(1-N\tau^2)^2 + 2N\tau^2\Big(\frac{1/2 + e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}}\Big)}\right)$$

remains as the only (possibly) stable fixed point of the ς^*-evolution in (4.7). It can be shown that it is indeed stable provided that either τ is sufficiently small or N is sufficiently large.

Since the calculations are rather lengthy, they can be found in Appendix D.1.3, p. 195.

The stationary mutation strength $\varsigma^*_{stat_2}$, (4.11), is the only (locally) stable invariant solution of (4.7). It is associated with a positive progress: $\varsigma^*_{stat_2}(\tau) < 2\mu c_{\mu/\mu,\lambda}$ for every $\tau < \infty$ with $\lim_{\tau\to 0}\varsigma^*_{stat_2}(\tau) = 2\mu c_{\mu/\mu,\lambda}$. That is, self-adaptation works in the sense that it is always associated with a positive expected progress. The system thus moves towards the optimum (on average) – regardless of the choice of the learning rate. The stationary progress rate itself can be determined by inserting the mutation strength (4.11) into the progress rate (4.8)

$$\varphi^*_{st} = \frac{\mu c^2_{\mu/\mu,\lambda}}{2}\left(1 - \left(N\tau^2 - \sqrt{(1-N\tau^2)^2 + 2N\tau^2\frac{1/2 + e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}}}\right)^2\right). \quad (4.19)$$

As the stationary mutation strength (4.11), (4.19) is a function of the learning rate. Before discussing the dependency on this parameter, the results obtained so far are compared with the results of experiments.

4.1.3 Comparison with Experiments

Figure 4.1 compares the stationary mutation strength (4.11) with the result of experiments for two search space dimensionalities, $N = 100$ and $N = 10,000$. While there are large deviations for the lower dimensional search space, the prediction quality improves for $N = 10,000$.

The predictions of (4.19) are compared with the results of experiments in Figs. 4.2 and 4.3 for two multi-recombinative ES, a $(10/10_I, 60)$- and a $(20/20_I, 60)$-ES. Also depicted are the results of a numerical calculation of the stationary progress rate using the N-dependent progress rate formula [23, p. 216f]

$$\varphi^*(\sigma^*) = \frac{c_{\mu/\mu,\lambda}\sigma^*(1+\frac{\sigma^{*2}}{2\mu N})}{\sqrt{1+\frac{\sigma^{*2}}{2N}}\sqrt{1+\frac{\sigma^{*2}}{\mu N}}} - N\Big(\sqrt{1+\frac{\sigma^{*2}}{\mu N}}-1\Big) + \mathcal{O}(\frac{1}{\sqrt{N}}) \tag{4.20}$$

in the derivation. As representative of sphere functions, $f(\mathbf{y}) = \|\mathbf{y}\|^2$ was used. The sampling process was started once a stationary normalized mutation strength was reached and kept up as long as $r^{(g)} > 10^{-75}$. In the case of nearly optimal learning rates and $N = 30$, the stationary phase consists of only $2,000$ - $3,000$ generations. Therefore, the experiments were repeated until each data point represents the average of at least $95,000$ experiments.

Since (4.19) has been derived using the N-independent progress rate formula (4.8), the agreement with the experiments for low-dimensional search spaces is rather poor. However, its general tendency as a function of τ is similar. Furthermore, the agreements improves for larger values of τ. The quality of the prediction of (4.19) increases steadily with the search space dimensionality (see Fig. 4.3).

If the N-dependent progress rate (4.20) is used, the agreement with the experiments improves. Although there are still relatively large deviations as long as τ is small, the curves of the predicted and the observed τ-values are closer together.

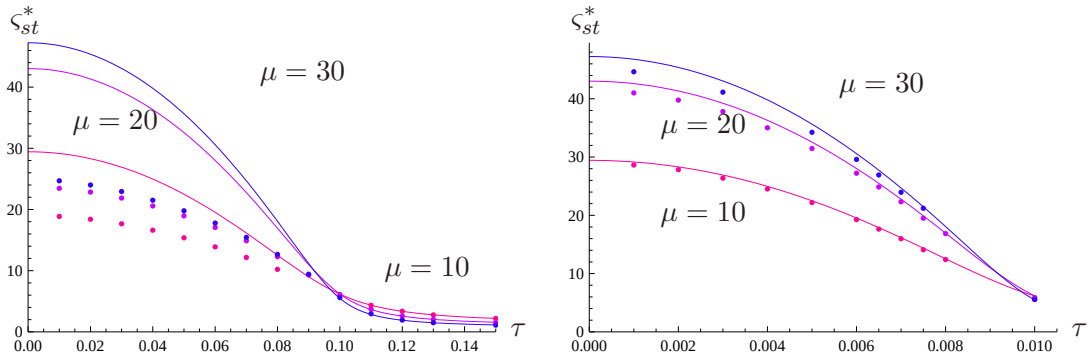
The experiments, the N-dependent progress rate, and (4.19) show a strong dependency on the choice of τ. In all cases, the progress increases with τ until a maximum is reached and the progress deteriorates. In the experiments and if the N-dependent progress rate is used, this behavior is more pronounced in high-dimensional than in low-dimensional search spaces: The maximal progress depends on the search space dimensionality. The position of the maximum, i.e., the optimal learning rate depends in all three cases on the search space dimension – decreasing with increasing N. Generally, using (4.19) leads to an underestimate of the measured optimal τ but improves if N growths.

The results of the experiments are in accordance with the results reported in [51] where the performance of $(\mu/\mu_I, \lambda)$-ES was investigated experimentally. The most astonishing observation reported in that work was that the performance of the ES sensitively depends on the choice of learning parameter. Therefore, the adjustment of the mutation strength is only nearly optimal in a narrow τ-range leading to a deterioration of the performance of the ES otherwise.

4.1.4 Self-Adaptation and Optimal Progress

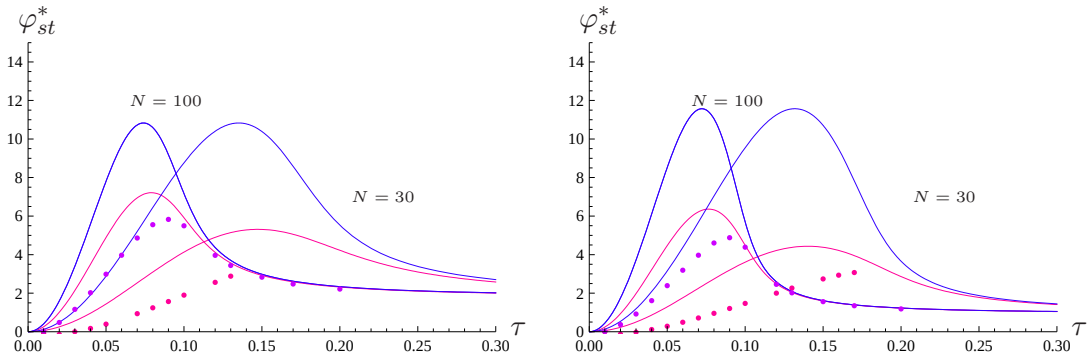
As the revealed by the experiments and as predicted by (4.19), intermediate self-adaptive ES exhibit a positive progress rate for a wide choice of τ-values. But the ES are sensitive to the choice of the learning rate. Nearly optimal progress in high-dimensional search spaces can only be achieved in a relatively narrow range of learning rates in the vicinity of an optimum. This optimal learning rate is easily obtained. To this end, maximizer of the progress rate (4.8) is needed. As stated in [23], the optimal progress rate and mutation strength are given by

$$\varphi^*_{R_{opt}} = \max_{\sigma^*}(c_{\mu/\mu,\lambda}\sigma^* - \frac{\sigma^{*2}}{2\mu}) = \frac{\mu c^2_{\mu/\mu,\lambda}}{2} \text{ and} \tag{4.21}$$



a) $N = 100$ b) $N = 10,000$

Figure 4.1: The stationary mutation strength as a function of the learning rate. Shown are the results for $(10/10_I, 60)$, $20/20_I, 60)$, and $(30/30_I, 60)$-ES. The data points denote the results of experiments, whereas the solid lines depict (4.11).

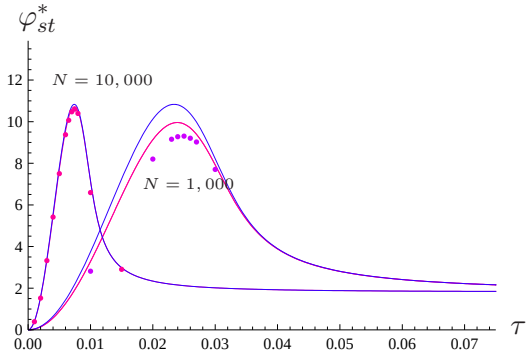


a) $(10/10_I, 60)$-ES, $N = 100$ and 30 b) $(20/20_I, 60)$-ES, $N = 100$ and 30

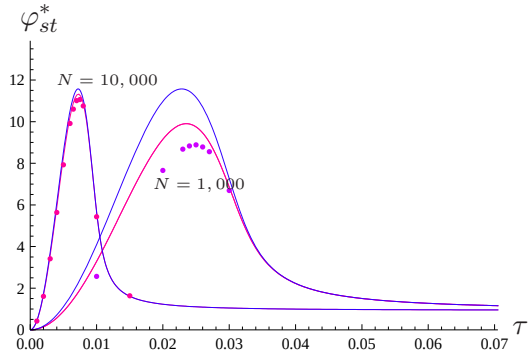
Figure 4.2: The stationary progress rate as a function of the learning parameter τ. Shown are from left to right the results for $N = 100$ and $N = 30$. The results of (4.19) are presented by the blue curves, whereas the red depict the results of using the N-dependent progress rate (4.20). The points indicate the results of experiments.

$$\varsigma^*_{\varphi_{R_{opt}}} = \arg\max_{\sigma^*} \varphi^*_R(\sigma^*) = \mu c_{\mu/\mu,\lambda,\lambda}. \tag{4.22}$$

Let us now consider the stationary mutation strength (4.11). Recall, by variating τ, (4.11) can be varied between the zero of the SAR (4.17), $\varsigma^*_{\psi_0}$, and the second zero of the progress rate (4.13), $\varsigma^*_{\varphi_{R_0}}$. The optimal mutation strength (4.22) is reachable since it lies inside the admissible interval, $[\varsigma^*_{\psi_0}, \varsigma^*_{\varphi_{R_0}}[$. Equation (4.22) can be used to determine an optimal learning rate by requiring that $\varsigma^*_{stat_2}(\tau) = \varsigma^*_{\varphi^*_{R_{opt}}} = 2\mu c_{\mu/\mu,\lambda}$ and solving the equation for τ_{opt} (see Appendix D.1.2). After a short calculation, the optimal learning rate τ_{opt} is given by



a) $(10/10_I, 60)$-ES, $N = 10,000$ and $1,000$



b) $(20/20_I, 60)$-ES, $N = 10,000$ and $1,000$

Figure 4.3: The stationary progress rate as a function of the learning parameter τ. Shown are from left to right the results for $N = 10,000$ and $N = 1,000$. The results of (4.19) are presented as the blue curves, whereas the red curves depict the results of using the N-dependent progress rate (4.20). The points indicate the results of experiments.

$$\tau_{opt} = \frac{1}{\sqrt{2N}} \sqrt{\frac{\mu c^2_{\mu/\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda} - 1/2 - e^{1,1}_{\mu,\lambda}}}. \tag{4.23}$$

As (4.23) shows, the optimal learning rate scales with $1/\sqrt{2N}$. Equation (4.23) can be rewritten to

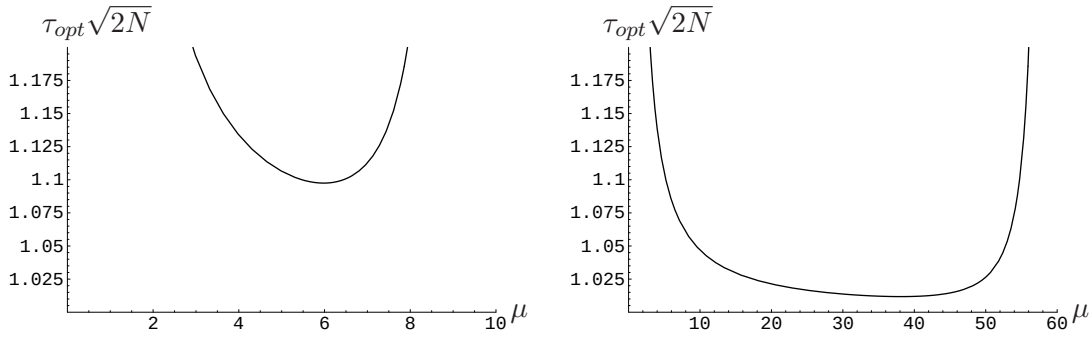
$$\tau_{opt} = \frac{1}{\sqrt{2N}} \sqrt{\frac{1}{1 - \frac{1/2 - e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}}}}. \tag{4.24}$$

Provided that $\mu c_{\mu/\mu} \gg (1/2 + e^{1,1}_{\mu,\lambda})/c_{\mu/\mu,\lambda}$ holds, the optimal learning rate is close to $1/\sqrt{2N}$. This requires sufficiently large offspring populations and choosing neither $\mu \not\approx 1$ nor $\mu \not\approx \lambda$. Figure 4.4 shows exemplary the dependency of the optimal learning rate on the parent number μ for $\lambda = 10$ and $\lambda = 60$. Provided that λ is not small, it can be seen that the optimal learning rate is close to $1/\sqrt{2N}$ for a relatively wide range of μ. That is, choosing $\tau \approx 1/\sqrt{2N}$ may be a good approximate for the optimal learning rate for typical truncation ratios in the interval $[0.125, 0.8]$.

Having derived an optimal learning rate, the question remains why ES with intermediate recombination suffer more severe performance losses than $(1, \lambda)$-ES from a non-optimal choice of the learning rate.

4.1.5 Investigating the τ-Sensitivity of Intermediate ES

The performance sensitivity of $(\mu/\mu_I, \lambda)$-ES on the choice of the learning rate is in pronounced contrast to the $(1, \lambda)$-ES. A $(1, \lambda)$-ES has a nearly optimal performance on the sphere test function for a wide range of τ-values. But what are the reasons for these different responses? In this section, the underlying causes are investigated more closely.

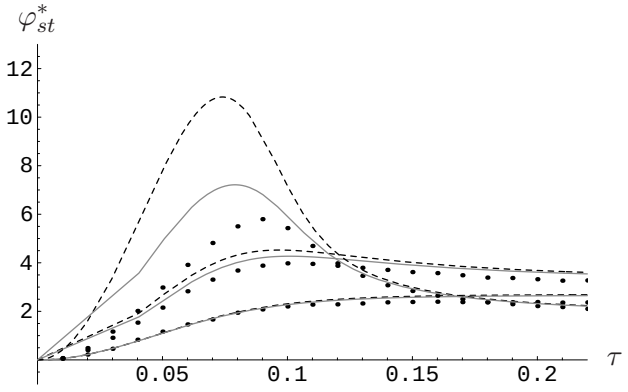


a) $(\mu/\mu_I, 10)$-ES b) $(\mu/\mu_I, 60)$-ES

Figure 4.4: The optimal learning rate (4.23) as a function of the parent number μ for $(\mu/\mu_I, 10)$-ES and $(\mu/\mu_I, 60)$-ES.

Deviations from the optimal learning rate

Let us start with some exemplary results for $N = 100$. Figure 4.5 depicts the stationary progress rate for some $(\mu/\mu_I, \lambda)$-ES. The transition from $\mu = 1$ to $\mu > 1$ leads to a qualitative different behavior: If there is only one parent, the stationary progress rate stabilizes on a nearly optimal level for a relatively wide range of $\tau \geq \tau_{opt}$. If μ increases, this is not the case anymore. The stationary progress rates show sharper peaks and the region with nearly optimal values becomes narrower.



$(\mu/\mu_I, 60)$-ES, $N = 100$

Figure 4.5: The stationary progress rate as a function of the learning parameter τ for $\mu = 1$, $\mu = 2$, and $\mu = 10$ bottom to top. The dashed curves represent the results of (4.19), whereas the solid lines depict the results obtained using the N-dependent progress rate. The points indicate the results of experiments.

Deviating from τ_{opt}: The Stationary Mutation Strength

But why does the stationary progress rate behave in this manner? To answer that question, consider first the stationary mutation strength (4.11). As stated, it depends on τ. The stationary mutation strength is furthermore determined and influenced by the progress rate (4.8) and the self-adaptation response (4.9). Recall, the progress rate (4.8), $\varphi^*_R(\sigma^*) = c_{\mu/\mu,\lambda}\sigma^* - \sigma^{*2}/(2\mu)$, reaches its optimum $\varphi^*_{opt} = \mu c^2_{\mu/\mu,\lambda}/2$, (4.21), at $\varsigma^*_{\varphi_{opt}} = \mu c_{\mu/\mu,\lambda}$, (4.22) and is positive for $0 < \sigma^* < 2\mu c_{\mu/\mu,\lambda}$. The SAR (4.9), $\psi(\sigma^*) = \tau^2(1/2 + e^{1,1}_{\mu,\lambda} - c_{\mu/\mu,\lambda}\sigma^*)$, is a monotonously decreasing function with zero (4.17), $\varsigma^*_{\psi_0} = (1/2 + e^{1,1}_{\mu,\lambda})/c_{\mu/\mu,\lambda}$.

It will be shown in the following that the relation between the zero of the SAR (4.17) to the optimal mutation strength (4.22) that is the size of $a := (1/2 + e^{1,1}_{\mu,\lambda})/(\mu c^2_{\mu/\mu,\lambda})$ is a decisive parameter.

First of all note that any deviation with Δ, $\Delta \geq 0$ from the optimal learning rate cannot have a significant effect if the limit of the stationary mutation strength is close to the optimal mutation strength, i.e., if

$$\lim_{\tau\to\infty} \varsigma^*_{stat_2}(\tau) = \frac{1/2 + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}} = \varsigma^*_{\psi_0} \approx \varsigma^*_{\varphi_{opt}} = \mu c_{\mu/\mu,\lambda}. \tag{4.25}$$

As can be verified by case inspection, this is only the case for $(1,\lambda)$-ES but not for $(\mu/\mu_I,\lambda)$-ES. Non-recombinative ES with only one parent can be expected to be robust against choices of larger learning rates. This also translates to the progress rate (see Fig. 4.5).

As can be seen, intermediate ES do have a potential problem in the sense that their limit for $N\tau^2 \to \infty$ is smaller than the optimizer. For too large learning rates, problems occur and the stationary mutation strength deviates far from the optimizer. This is amplified if the parent-offspring ratio is chosen around 0.27 which is recommended as optimal in the case of the sphere: The zero of the SAR is significantly smaller than the optimal mutation strength. In the case of $(\mu/\mu_I, 60)$-ES for example, the ratio drops to < 0.2 for $\mu \in (5, 55)$ with a minimal value of ≈ 0.023.

This is not the only problem, though. If the decline in the performance were gradual, the difference between limit and optimal value would not be so decisive. The question remains: What are the effects of smaller deviations from the optimal τ?

In the following part, this question is answered by taking a closer look at the influence of a deviation on the stationary progress rate (4.19) and stationary mutation strength (4.11). But first, the equations are simplified. A straightforward comparison of $\varsigma^*_{\varphi_{opt}}$ and the zero of the SAR $\varsigma^*_{\psi_0} = (1/2 + e^{1,1}_{\mu,\lambda})/c_{\mu/\mu,\lambda}$ (4.17) with the stationary mutation strength (4.11) shows that (4.11) can be re-expressed by a very simple equation

$$\begin{aligned} \varsigma^*_{st} &= \varsigma^*_{\varphi_{opt}}\left((1-x) + \sqrt{(1-x)^2 + 2ax}\right) \\ &:= \varsigma^*_{\varphi_{opt}} f(x) \end{aligned} \tag{4.26}$$

with $a = (1/2 + e^{1,1}_{\mu,\lambda})/(\mu c^2_{\mu/\mu,\lambda})$ and $x = N\tau^2$. Considering the optimal progress in the stationary state, one would like to have $f(x) = 1$ so that the optimal mutation strength is assumed. This (cf. (4.23)) equals the condition

$$\begin{aligned} x_{opt} &= \sqrt{(1-x_{opt})^2 + 2ax_{opt}} \\ \Rightarrow x_{opt} &= \frac{1}{2(1-a)}. \end{aligned} \tag{4.27}$$

Equation (4.27) is well-defined for all $a \in (0,1)$. The case $a = 0$ cannot occur. If $a > 1$, i.e., $\varsigma^*_{\psi_0} > \varsigma^*_{\varphi_{opt}}$, the ES is unable to work with the optimal progress rate at any rate.

In the following, only the function f is addressed that is the results obtained are relative to the optimal mutation strength and do not depend on its height. Let the deviation be given by Δ, $\Delta \geq 0$. Assuming smallness of the deviations, f can be expanded into its Taylor series around x_{opt}. The Taylor series of $f(x) = 1 - x + \sqrt{(1-x)^2 + 2ax}$ around $x_{opt} = 1/(2(1-a))$is given by

$$T_f(x_{opt} + \Delta) = f(x_{opt}) + f'(x_{opt})\Delta + \frac{f''(x_{opt})}{2}\Delta^2 + \mathcal{O}(\Delta^3). \tag{4.28}$$

The first derivative of f is given by

$$f'(x) = \frac{x - (1-a)}{\sqrt{(1-x)^2 + 2ax}} - 1 \tag{4.29}$$

whereas the second reads

$$f''(x) = \frac{1}{\sqrt{(1-x)^2 + 2ax}} - \frac{(x-(1-a))^2}{\sqrt{(1-x)^2 + 2ax}^3}. \tag{4.30}$$

First, note the following:

1. The function f approaches a for $x \to \infty$.
2. For all $a \in [0,1)$, $f'(x) \leq 0$ for all $x \geq 0$.
3. For all $a \in [0,1)$, $f''(x) \geq 0$ for all $x \geq 0$.

In other words, the first derivative is negative but monotonously increasing. Using the mean value theorem, the absolute value of the deviation of f is

$$|f(x_{opt} + \Delta) - f(x_{opt})| = |f'(\theta)|\Delta$$

for a θ with $0 \leq \theta \leq \Delta$. Therefore,

$$\begin{aligned} |f(x_{opt} + \Delta) - f(x_{opt})| &\leq |f'(x_{opt})|\Delta \\ &= 2(1-a)^2\Delta \end{aligned} \tag{4.31}$$

follows. For $\Delta \to 0$, the inequality becomes "=". Assuming that Δ is small enough so that the "="-sign roughly holds, the effect of the deviation depends on the parameter a, i.e., the quotient $\varsigma^*_{\psi_0}/\varsigma^*_{\varphi_{opt}}$. The effect of a deviation is enhanced for all choices of a with $a \leq 1/\sqrt{2}$. The question remains though, whether this translates to the stationary optimal progress.

Deviating from τ_{opt}: The Stationary Progress Rate

The stationary progress rate can similarly be written as

$$\begin{aligned} \varphi^*_{st} &= \varphi^*_{opt} 2\, f(x)(1 - \frac{f(x)}{2}) \\ &:= \varphi^*_{opt}\, g(x) \end{aligned} \tag{4.32}$$

which can be seen by plugging (4.26) into (4.8). The question remains how (4.32) responds to deviations from x_{opt}. This is analyzed in this section. Note that since x_{opt} leads to a global maximum

$g'(x_{opt}) = 0$ holds. The quantity of interest is the rate by which the optimum is left. Therefore, let us consider the first derivative. The Taylor-Series of g' around x_{opt} is given by

$$T'_g(x_{opt} + \Delta) \quad = \quad g'(x_{opt}) + g''(x_{opt})\Delta + g'''(x_{opt})/2\Delta^2 + \mathcal{O}(\Delta^3). \tag{4.33}$$

The first derivative is given by $g'(x) = 2f'(x)(1 - f(x))$ and the second by $g''(x) = 2f''(x)(1 - f(x)) - 2(f'(x))^2$. The rate by which the optimum is left can be given by

$$|g'(x_{opt} + \Delta) - g'(x_{opt})| \quad = \quad 4(1-a)^4\Delta + \mathcal{O}(\Delta^2). \tag{4.34}$$

Thus, the behavior of the stationary mutation strength and therefore of the stationary progress rate can be traced back over $a = (1/2 + e^{1,1}_{\mu,\lambda})/(\mu c^2_{\mu/\mu,\lambda})$ to the SAR and the progress rate. Only if $a = (1/2 + e^{1,1}_{\mu,\lambda})/(\mu c^2_{\mu/\mu,\lambda}) \approx 1$ the stationary progress rate can be expected to be robust against all choices of $\tau \geq \tau_{opt}$. Otherwise if $a = (1/2 + e^{1,1}_{\mu,\lambda})/(\mu c^2_{\mu/\mu,\lambda}) < 1$ which equals $\varsigma^*_{\psi_0} < \varsigma^*_{\varphi_{opt}}$, the system eventually deviates from the actual optimum since $\varsigma^*_{stat_2}$ approaches $\varsigma^*_{\psi_0}$ for $N\tau^2 \to \infty$. The rate by which the optimal progress rate (relative to the optimum, of course) is left also depends on this ratio. The smaller $\varsigma^*_{\psi_0}$ is in comparison to $\varsigma^*_{\varphi_{opt}}$, the sooner the optimal progress rate is left and the stronger the limit progress rate deviates from φ^*_{opt}.

It remains to investigate the effects of recombination on the SAR and the progress rate. Keeping λ constant, the mutation strength $\varsigma^*_{\varphi_{opt}} = \mu c_{\mu/\mu,\lambda}$ is a function of μ. Its plot (see Fig. 4.6) is symmetrical around the maximum $\mu = \lambda/2$. The free μ factor stems actually from the loss term of the progress rate which is dampened by recombination. Considering the derivation of the progress rate [23, p. 210f] or B.1.2 this loss term results from the perpendicular $\langle \mathbf{z}_B \rangle$-component of (B.20). Recombination actually leads to a *genetic repair* effect because these harmful components are statistically averaged out.

The zero of the SAR defines the mutation strength for which no change with respect to $R^{(g)}$ is expected. That is, the non-normalized $\langle \varsigma^{*(g+1)} \rangle$ equals $\langle \varsigma^{*(g)} \rangle$ and any change from $\langle \varsigma^{*(g)} \rangle$ to $\langle \varsigma^{*(g+1)} \rangle$ is a result of $\varphi^*/N \neq 0$. As the parental number μ increases, the zero $\varsigma^*_{\psi_0}$ decreases first. Once μ is closer to λ, it assumes larger values until it gets greater than $\varsigma^*_{\varphi_{opt}}$ and finally even greater than $\varsigma^*_{\varphi_0}$.

In contrast to the progress rate, the SAR is not directly influenced by the recombination of the object parameters: Here, the average is taken over the mutation strengths and the selection only considers the fitness values, i.e., the resulting distances to the optimum. The recombination of the object parameters from which the progress rate benefits occurs afterwards and thus cannot play a role in the case of the self-adaptation response. The SAR is a linear function of the mutation strength with no free μ-term and is influenced by the parental number over the progress coefficients.

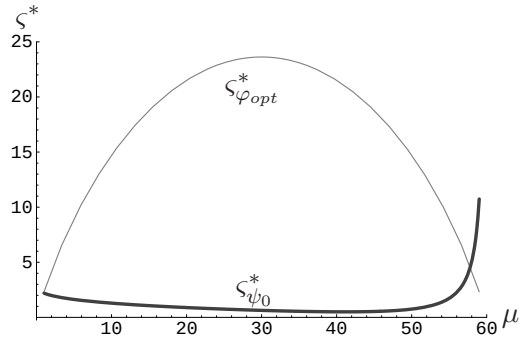
As a result, the effect of changing μ is somewhat more damped – compared to that of $\varsigma^*_{\varphi_{opt}}$ which can be seen in Fig. 4.6.

4.2 Self-Adaptation and Noisy Fitness Evaluations: $(1, \lambda)$-ES

In this section, the analysis is extended to self-adaptation under noisy fitness evaluations. The noise term is represented by the standard noise model, that is, by an additive normally distributed noise term with zero mean and standard deviation or noise strength σ_ϵ.

Definition 4. The noisy sphere model with the standard noise model is given by

$$f(\mathbf{y}) \quad = \quad g(\|\mathbf{y} - \hat{\mathbf{y}}\|) + \epsilon \tag{4.35}$$



$(\mu/\mu_I, 60)$-ES, comparison of $\varsigma^*_{\psi_0}$ and $\varsigma^*_{\varphi_{opt}}$

Figure 4.6: Comparison between the optimal point of the progress rate (symmetric curve), $\varsigma^*_{\varphi_{opt}} = \mu c_{\mu/\mu,\lambda}$, and the zero of the SAR, $\varsigma^*_{\psi_0} = (1/2 + e^{1,1}_{\mu,\lambda})/c_{\mu/\mu,\lambda}$, i.e., the limit of stationary mutation strength for $N\tau^2 \to \infty$.

with $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$, $g : \mathbb{R} \to \mathbb{R}$ a monotonously in- or decreasing function, and $\hat{\mathbf{y}}$ the optimizer of g. □

In the following, only the case of quadratic sphere functions is explicitely considered, i.e., $g(R) = \pm R^2$. The equations can be easily adapted to include the general case. The noise strength σ_ϵ can be used to model several situations. This section focuses on the most common scenario: The strength of the noise is independent of the position of the ES in the search space. The noise strength σ_ϵ is assumed to be a constant value, $\sigma_\epsilon = c$. This causes the influence of the noise to change through the search space. Dependent on the distance $\|\mathbf{y} - \hat{\mathbf{y}}\|$, it may have high influence if the value of $|g|$ is small or it may be negligible for large $|g|$-values. Note, this noise model actually prevents the ES (recombinative or non-recombinative) to converge to the optimal $\hat{\mathbf{y}}$ as was shown in various papers by Beyer and Arnold (see, e.g., [24, 25, 4]). In the following the evolution of the ES under this type of noise is referred to as *evolution under permanent noise* σ_ϵ.

4.2.1 Modeling the Evolution Strategy

To model the evolution strategies, again the evolution equations

$$R^{(g+1)} = R^{(g)} - \varphi_R(\varsigma^{(g)}, R^{(g)}, \sigma_\epsilon) + \epsilon_R(\varsigma^{(g)}, R^{(g)}, \sigma_\epsilon) \tag{4.36}$$

$$\varsigma^{(g+1)} = \varsigma^{(g)}\Big(1 + \psi(\varsigma^{(g)}, R^{(g)}, \sigma_\epsilon)\Big) + \epsilon_\sigma(\varsigma^{(g)}, R^{(g)}, \sigma_\epsilon) \tag{4.37}$$

are used. The terms ϵ_R and ϵ_σ cover the perturbations whereas the progress rate φ_R and self-adaptation response ψ stand for the expected changes. In the following, the usual normalizations are introduced – setting $\sigma^* := (N/R)\varsigma^{(g)}$, $\varphi^*_R := (N/R)\varphi_R$, and $\sigma^*_\epsilon := [N/(2R^2)]\sigma^{(g)}_\epsilon$. As before, $R := R^{(g)}$ is used in order to shorten the notation. The last normalization

$$\sigma^*_\epsilon := \frac{N}{2R^2}\sigma^{(g)}_\epsilon. \tag{4.38}$$

gives raise to a third evolution equation

$$R^{(g+1)} = R\Big(1 - \frac{1}{N}\varphi^*_R(\sigma^*, R, \sigma^*_\epsilon)\Big) + \epsilon^*_R(\sigma^*, R, \sigma^*_\epsilon))\Big) \tag{4.39}$$

$$\varsigma^{*(g+1)} = \sigma^* \Big(\frac{1 + \psi(\sigma^*, R, \sigma_\epsilon^*) + \epsilon_\sigma^*(\sigma^*, R, \sigma_\epsilon^*))}{1 - \frac{1}{N}\varphi_R^*(\sigma^*, R, \sigma_\epsilon^*) + \epsilon_R^*(\sigma^*, R, \sigma_\epsilon^*))} \Big) \tag{4.40}$$

$$\sigma_\epsilon^{*(g+1)} = \frac{\sigma_\epsilon^*}{\Big(1 - \frac{1}{N}\varphi_R^*(\sigma^*, R, \sigma_\epsilon^*) + \epsilon_R^*(\varsigma^*, R, \sigma_\epsilon^*))\Big)^2}. \tag{4.41}$$

The progress rate is derived in Appendix B.1. For $N \to \infty$ and $\tau = 0$

$$\varphi_R^*(\sigma^*, R, \sigma_\epsilon^*) = c_{1,\lambda} \frac{\sigma^{*2}}{\sqrt{\sigma^{*2} + \sigma_\epsilon^{*2}}} - \frac{\sigma^{*2}}{2} \tag{4.42}$$

is obtained. The derivation of the SAR (cf. C.1.1, Eq. (C.36)) gives

$$\psi(\sigma^*) = \tau^2 \Big((d_{1,\lambda}^{(2)} - 1) \frac{\sigma^{*2}}{\sigma^{*2} + \sigma_\epsilon^{*2}} - c_{1,\lambda} \frac{\sigma^{*2}}{\sqrt{\sigma^{*2} + \sigma_\epsilon^{*2}}} \Big) \tag{4.43}$$

for $N \to \infty$ and $\tau \ll 1$. The progress coefficient $d_{1,\lambda}^{(2)}$ in (4.43) is a special case of the progress coefficients and is defined by

$$d_{1,\lambda}^{(k)} := \frac{\lambda}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^k \mathrm{e}^{-\frac{t^2}{2}} \Phi(t)^{\lambda - 1} \, \mathrm{d}t \tag{4.44}$$

([23, p. 119]). Note, $d_{1,\lambda}^{(2)} - 1 = 1/2 + e_{1,\lambda}^{1,1}$ holds. The evolution of the σSA-ES is fully described by the system of stochastic evolution equations (4.39), (4.40), and (4.41). Due to the stochasticity, the general solution would be given by a time-dependent pdf $p(r, \varsigma^*, \sigma_\epsilon^*)^{(g)}$ to be obtained by solving the corresponding Chapman-Kolmogorov-Equations. In this section, it is abstained from trying to solve these equations by means of analytical approximations in general. Instead, only the stationary state (also referred to as steady state) is considered which is observed for a sufficiently large generation time g, i.e., in the limit $g \to \infty$. Furthermore, we will not search for the steady state pdf, but rather for its first moment assuming that the fluctuating parts in the evolution equations (4.39), (4.40), and (4.41) can be neglected. This is a rather crude approximation, therefore it will be compared with simulations.

4.2.2 The Stationary State

As already mentioned, the stochastic perturbation parts of the evolution equations (4.39), (4.40), and (4.41) are neglected. Applying thus a deterministic approach, the equations simplify to

$$R^{(g+1)} = R \left(1 - \frac{1}{N} \varphi^*(\varsigma^{*(g)}, \sigma_\epsilon^{*(g)}) \right) \tag{4.45}$$

$$\varsigma^{*(g+1)} = \sigma^* \frac{1 + \psi(\varsigma^{*(g)}, \sigma_\epsilon^{*(g)})}{\big(1 - \frac{1}{N}\varphi^*(\varsigma^{*(g)}, \sigma_\epsilon^{*(g)})\big)} \tag{4.46}$$

$$\sigma_\epsilon^{*(g+1)} = \frac{\sigma_\epsilon^{*(g)}}{\big(1 - \frac{1}{N}\varphi^*(\varsigma^{*(g)}, \sigma_\epsilon^{*(g)})\big)^2}. \tag{4.47}$$

As (4.45) to (4.47), (4.42), and (4.43) show, the R-evolution, Eq. (4.45), is governed by the evolution of the mutation and the noise strength, Equations (4.46) and (4.47). However, (4.46) and (4.47) do *not* depend on (4.45). That is why only the system (4.46) and (4.47) has to be considered whereas the R-dynamics is fully controlled by the solution of (4.46) and (4.47).

Evolution under Permanent Noise σ_ϵ

Let us now consider the case of a constant noise strength σ_ϵ. The normalized noise strength defined in (4.38), $\sigma_\epsilon^{*(g)} = \sigma_\epsilon[N/(2(R^{(g)})^2)]$, gradually increases during the course of the evolution until no progress is possible anymore and the evolution of the $R^{(g)}$ comes to a halt (on average).

Three phases can be distinguished: As long as the system is far away from the optimum, the influence of the normalized noise strength can be neglected. The situation resembles the undisturbed sphere. As a consequence, the steady state formula

$$\varsigma_{st}^* = c_{1,\lambda}(1 - N\tau^2) + \sqrt{c_{1,\lambda}^2(1 - N\tau^2)^2 + N\tau^2(2d_{1,\lambda}^{(2)} - 1)}, \tag{4.48}$$

obtained in [23], holds. Considering the maximizer $\varsigma^* = c_{1,\lambda}$ of the noise-free progress rate, the optimal learning rate reads $\tau = c_{1,\lambda}/\sqrt{N(2c_{1,\lambda}^2 + 1 - 2d_{1,\lambda}^{(2)})}$.

As the ES progresses and the normalized noise strength increases, $\varsigma^* = c_{1,\lambda}$ does not fulfill the steady state condition anymore. The former steady state is lost. The increasing noise strength $\sigma_\epsilon^{*(g)}$ influences the equations more and more and leads to a continuously changing mutation strength.

Finally, the R- and ς^*-dynamics converge to a stationary state which is characterized by $\varphi_R^*(\sigma^*, \sigma_\epsilon^*) = 0$ and $\psi(\sigma^*, \sigma_\epsilon^*) = 0$.

The focus of this section lies on the stationary state behavior. Before continuing, the zero points of the progress rate and SAR have to be determined. Let us start with the progress rate. There are two qualitatively different zeros of φ_R^* (4.42), $\varsigma_{\varphi R1}^* = 0$ (associated ideally with $\sigma_\epsilon^* = 2c_{1,\lambda}$) and

$$\varsigma_{\varphi R2}^* = \sqrt{4c_{1,\lambda}^2 - \sigma_\epsilon^{*2}}. \tag{4.49}$$

Demanding stationarity of the σ^*-evolution, i.e., $\psi = 0$, the latter condition (4.49) can be used to determine a stationary mutation strength ς_{st}^* and thus the corresponding noise strength $\sigma_{\epsilon\,st}^*$. Setting $\psi(\varsigma_{st}^*) = 0$ gives

$$\begin{aligned} 0 &= \frac{1}{2} + \frac{(\varsigma_{st}^*)^2}{(\varsigma_{st}^*)^2 + (\sigma_{\epsilon\,st}^*)^2}(d_{1,\lambda}^{(2)} - 1) - \frac{c_{1,\lambda}(\varsigma_{st}^*)^2}{\sqrt{(\varsigma_{st}^*)^2 + (\sigma_{\epsilon\,st}^*)^2}} \\ \Rightarrow 0 &= \frac{1}{2} + (\varsigma_{st}^*)^2 \frac{d_{1,\lambda}^{(2)} - 1 - 2c_{1,\lambda}^2}{4c_{1,\lambda}^2} \end{aligned} \tag{4.50}$$

The so obtained stationary mutation strength

$$\varsigma_{st}^* = 2c_{1,\lambda} \frac{1}{\sqrt{2(2c_{1,\lambda}^2 + 1 - d_{1,\lambda}^{(2)})}}. \tag{4.51}$$

can be used together with (4.49) to determine the stationary noise strength

$$\sigma_{\epsilon\,st}^* = 2c_{1,\lambda}\sqrt{1 - \frac{1}{2(2c_{1,\lambda}^2 + 1 - d_{1,\lambda}^{(2)})}} \tag{4.52}$$

and using $\sigma_\epsilon^{*(g)} = \sigma_\epsilon[N/(2(R^{(g)})^2)]$ to obtain a residual location error

$$R_{st} = \sqrt[2]{\frac{\sigma_\epsilon N}{4c_{1,\lambda}}\sqrt{\frac{2(2c_{1,\lambda}^2+1-d_{1,\lambda}^{(2)})}{2(2c_{1,\lambda}^2+1-d_{1,\lambda}^{(2)})-1}}} \tag{4.53}$$

defined for $2c_{1,\lambda}^2 + 1 - d_{1,\lambda}^{(2)} > 1/2$.

Discussion of the Stationary State

As explained above, the R-evolution is governed by the evolutions of the mutation strength and the noise strength. Therefore, it suffices to consider the evolution equations for the latter. Taking (4.46) and (4.47) into account, there are two different pairs of equilibrium points of the evolution equations The first with $\mathbf{e}_1 = (0, w)^{\mathrm{T}}$ with $w \in \mathbb{R}$ and ideally $w = 2c_{1,\lambda}$ and the second at $\mathbf{e}_2 = (s_2, w_2)^{\mathrm{T}}$ with s_2 given by (4.51) and w_2 by (4.52). The question arises which of these pairs is locally stable, i.e., stable w.r.t. small disturbances.

To this end, a linear approximation in the vicinity of the fixed point or equilibrium solution, respectively, is used again. The first equlibrium solution, $\mathbf{e}_1 = (0, w)^{\mathrm{T}}$, is not stable since it admits an unstable local manifold (see D.2.1). The stability of the second equilibrium point (4.51) and (4.52) is determined numerically since the expression obtained is rather clumsy. In Appendix D.2.1, it is shown that the second stationary solution is stable via the linear approximation if $\tau > 0$ – at least for the sphere. Figure 4.7 illustrates the behavior of the equilibrium points if small disturbances occur.

Interestingly, the distance $R_{st}^B = \sqrt[2]{\sigma_\epsilon N/(4c_{1,\lambda})}$ obtained as an ideal case for a vanishing mutation strength and for a noise strength $\sigma_{\epsilon\,st}^* = 2c_{1,\lambda}$ does not differ much from (4.53) (see Fig. 4.8). If the size of the offspring population is sufficiently large, the difference is negligible. This means in turn that any mutation strength between zero and (4.51) leads to similar residual location errors.

Simulations: Comparison with Experiments

In this section, the predicted stationary mutation strength (4.51) and the residual location error (4.53) are compared with the results of experiments. The quadratic sphere was chosen as test function in all experiments.

Figure 4.8 compares the predicted expected R-value at the steady state with simulations of real ES runs depending on the number of offspring individuals. As one can see, the predictive quality of (4.53) is rather good, however, one observes some randomly appearing small deviations of some data points from the curve. There is a deeper reason for this behavior which can be traced back to the σ^*-evolution.

Figure 4.9 a) presents the long-term σ^*-dynamics of a typical run of an $(1, 100)$-ES on a sphere with constant noise strength. After approaching the vicinity of the steady state (within a few hundred generations if the learning rate is chosen appropriately) the initial steady state is lost again. Unlike the prediction of the *deterministic* approximation, the ES is generally not able to regain the predicted steady state ς^* (4.51). Sometimes short nearly stationary phases exist, but they appear only sporadically. The only observable tendency seems to be a general preference of small mutation strengths. That is, the predicted stationary mutation strength (4.51) cannot be observed after reaching the vicinity of R_{st}. However, the resulting effect on the finally observed steady state R is rather small: Since any mutation strength between zero and (4.51) leads to nearly the same residual location error, both

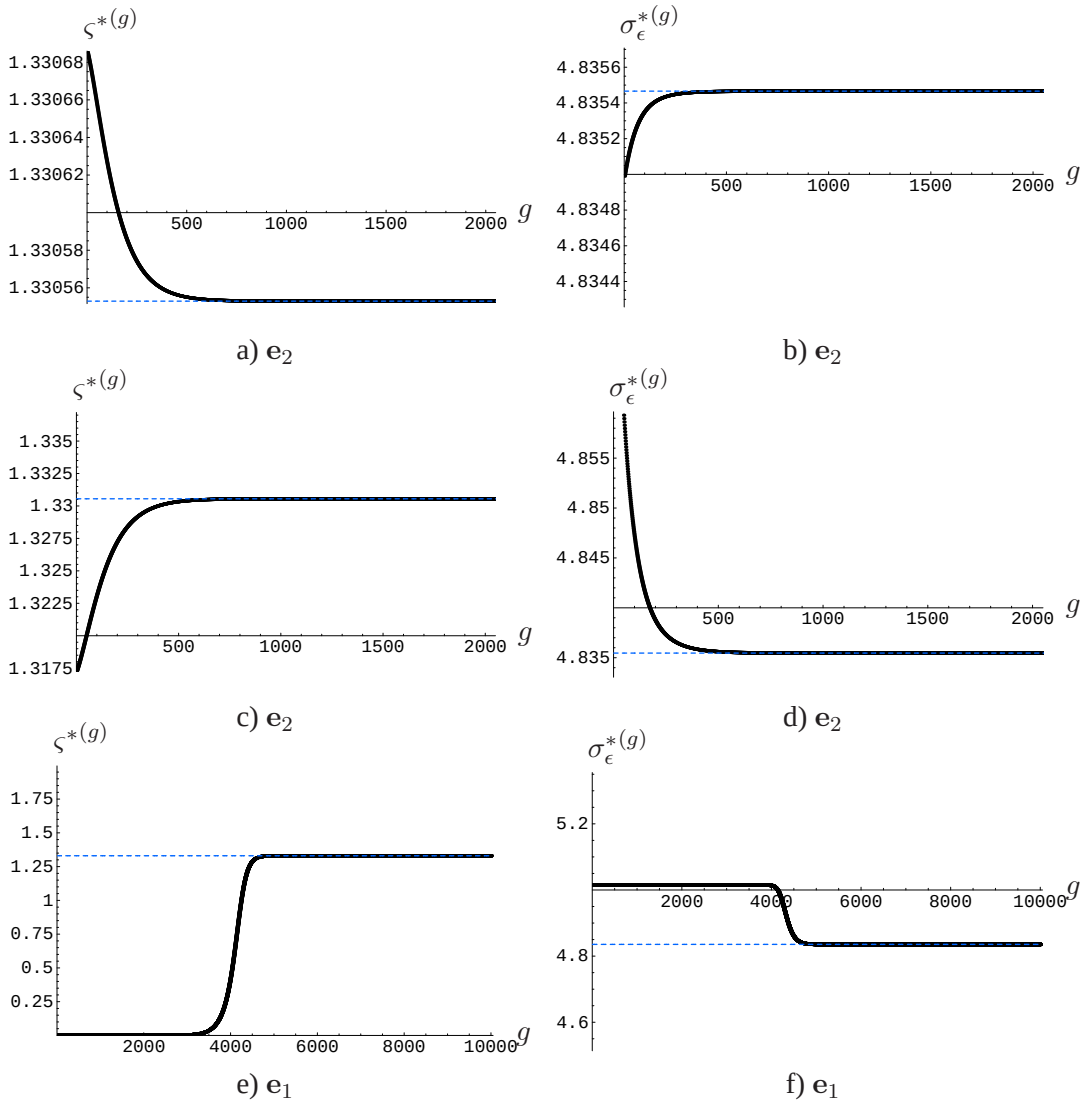


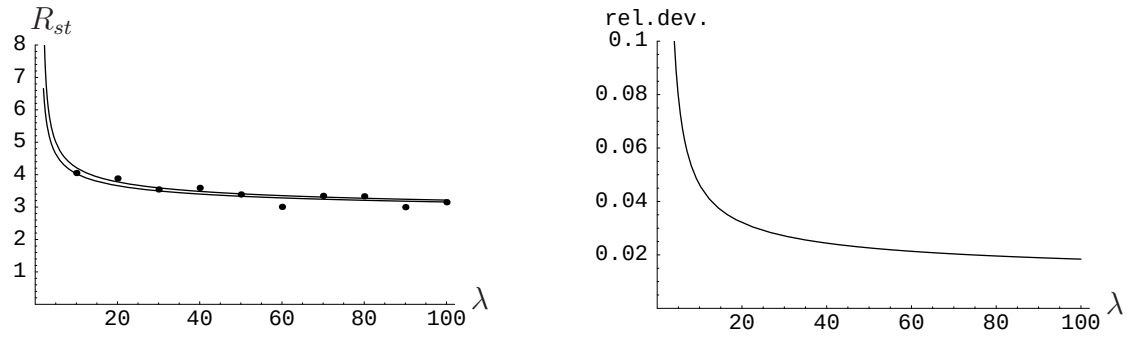
Figure 4.7: Behavior of the evolution equations (4.46) and (4.47) close to the fixed points. As parameters $\lambda = 100$, $N = 100$, and $\tau = 0.1$ were chosen. The dashed lines represent the steady state mutation strength (4.51) and the noise strength (4.52), respectively.

estimates (4.53) and $R_{st}^B = \sqrt[2]{(\sigma_\epsilon N)/(4c_{1,\lambda})}$ serve relatively well as predictors of the final R_{st} which can be seen in Fig. 4.8.

Interestingly, it can be seen in Fig. 4.9 that the non-existence of a final stationary state of the mutation strength seems to occur only in the case of $(1, \lambda)$-ES. If intermediate recombinative $(\mu/\mu_I, \lambda)$-ES are used, the behavior changes qualitatively: The mutation strength fluctuates very stably around a stationary value. This interesting phenomenon is discussed in the next section.

On the Erratic Behavior of the $(1, \lambda)$-ES and a Possible Remedy

In order to discuss the steady state behavior of the ES, it should be recalled that the ES is operating in the large-noise regime. After having reached the vicinity of R_{st}, the noise with strength $\sigma_\epsilon = \text{const.}$



a) R_{st} (4.53) and R_{st}^B b) Relative deviation

Figure 4.8: Final residual location errors as obtained by (4.53) (upper curve) and $R_{st}^B = \sqrt[2]{\sigma_\epsilon N/(4c_{1,\lambda})}$. The parameters were set to $N = 100$, $\tau = 0.1$, and $\sigma_\epsilon = 1$. The points denote the results of $(1, \lambda)$-ES runs. Each data point was obtained by averaging over $500,000$ generations. Figure b) shows the relative deviation of (4.53) from R_{st}^B.

is so large that it totally overshadows the actual fitness information. Thus, the selection process becomes nearly random, i.e., the σ^*-evolution is basically driven by random samples from a log-normal distribution with parameter τ. Under this condition, the probability of an in- or decrease of the mutation strength equals $1/2$

$$\begin{aligned} \Pr\left(\varsigma^{*(g+1)} \leq \varsigma^{*(g)}\right) &= \int_0^{\varsigma^{*(g)}} \frac{\mathrm{e}^{-\frac{(\ln(\varsigma^*/\varsigma^{*(g)}))^2}{2\tau^2}}}{\tau\varsigma^*\sqrt{2\pi}}\,\mathrm{d}\varsigma^* \\ &= \int_{-\infty}^{0} \frac{\mathrm{e}^{-\frac{t^2}{2\tau^2}}}{\tau\sqrt{2\pi}}\,\mathrm{d}t = \Phi_{0,\tau^2}(0) = \frac{1}{2}. \end{aligned} \tag{4.54}$$

Put it another way, the σ^*-evolution of the $(1, \lambda)$-σSA-ES performs a biased random walk: It probabilistically accepts any ς^*-decrease, however, it punishes large ς^* values due to their selective disadvantage. As a result, the $(1, \lambda)$-σSA-ES has a slight tendency towards smaller mutation strengths. This is a clear disadvantage of the standard version of $(1, \lambda)$-σSA-ES. A possible remedy would be to increase the probability of σ^*-increases slightly. This idea will be taken up again.

But before let us consider recombinative strategies. The question arises why recombinative strategies exhibit a qualitatively different behavior. For sake of simplicity, the case of an infinite number of parents is considered. Without loss of generality, let $\varsigma^{*(g)} = 1$. Since the mutation strengths Y_i of the μ parents are independently identically distributed random variables with mean $m = \exp(\tau^2/2)$ and variance $s^2 = \exp(\tau^2)[\exp(\tau^2) - 1]$, the sum $1/\mu \sum_{i=1}^{\mu} Y_i$ converges to a normally distributed random variable $S \sim \mathcal{N}(m, s^2/\mu)$. If μ is sufficiently large, the probability that the mutation strength decreases can be estimated using the cdf of the normal distribution. The probability of $(1/\mu) \sum_{i=1}^{\mu} Y_i \leq 1$ becomes

$$Pr\left(\frac{1}{\mu}\sum_{i=1}^{\mu} Y_i \leq 1\right) \;\rightarrow\; \Phi\left(\sqrt{\mu}\frac{1 - \mathrm{e}^{\frac{\tau^2}{2}}}{\sqrt{\mathrm{e}^{\tau^2}(\mathrm{e}^{\tau^2} - 1)}}\right) \tag{4.55}$$

which is smaller than $1/2$ if $\tau > 0$. Actually, this preference for σ^*-increases can also be shown for the smallest parental population size $\mu = 2$. Therefore, an intermediate recombinative strategy possesses a natural tendency to provide more increases than decreases.

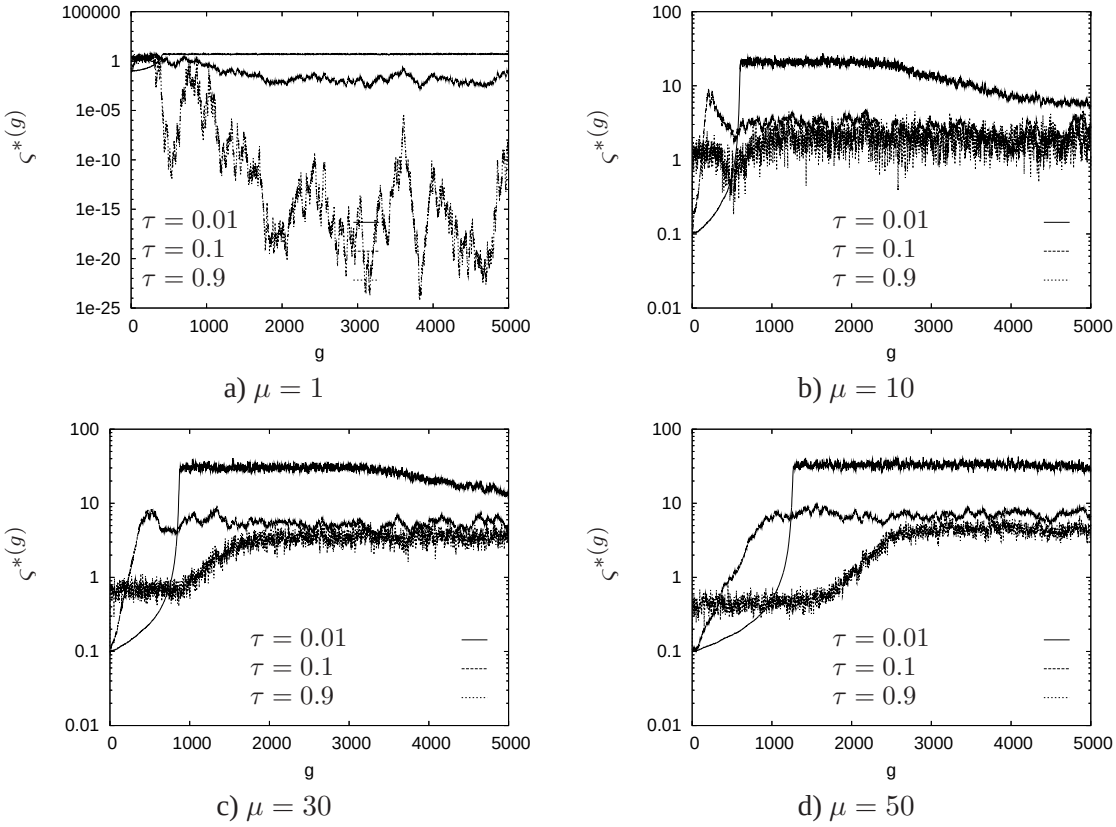


Figure 4.9: The σ^*-evolution of some typical $(\mu/\mu_I, 100)$-ES runs ($N = 100$) on the quadratic sphere. Shown are the results for $\tau = 0.01$ (topmost curve), $\tau = 0.1$, and $\tau = 0.9$ (lowest curve). The duration of the initial steady state for $\varsigma^{*(g)}$ depends on τ and thus on the convergence velocity of the R-variable towards the final steady state.

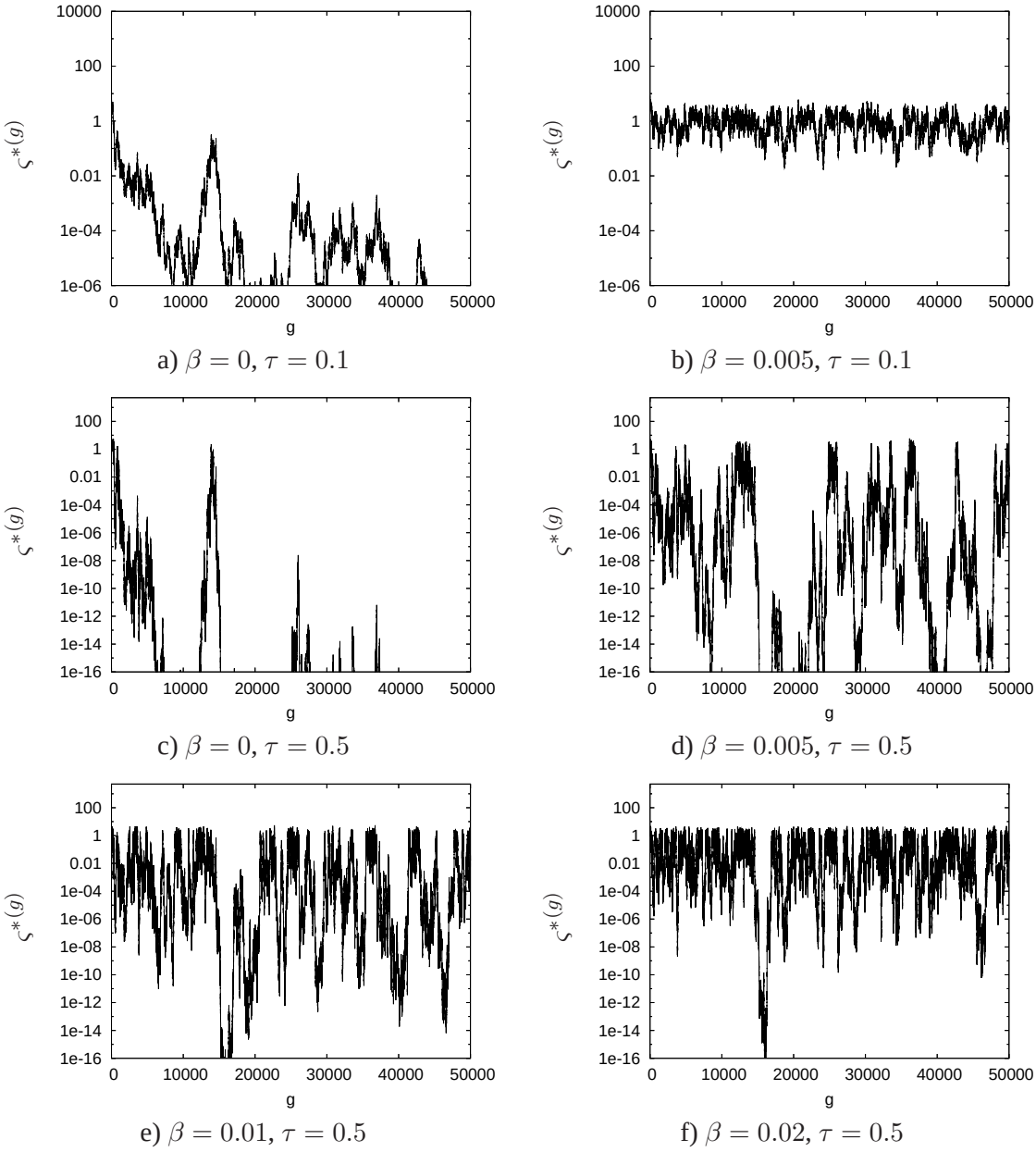
As to the $(1, \lambda)$-ES, this suggests the introduction of a slight preference for σ^*-increases in the mutation operator by using a log-normal distribution

$$p^*_\sigma(\varsigma^*|\sigma^*) = \frac{1}{\varsigma^* \tau \sqrt{2\pi}} \exp\left(-\frac{(\ln(\varsigma^*/\sigma^*) - \beta)^2}{2\tau^2}\right) \tag{4.56}$$

with a bias $\beta > 0$. The question remains how to choose β. On the one hand, it has to be sufficiently large to induce a trend towards larger mutation strengths. One the other hand considering the change $\sigma_l = \sigma^{(g)}\zeta$, the $\mathrm{E}[\zeta] \approx 1$ condition still has to be fulfilled.

Figure 4.10 shows the results of some ES-runs with different choices of β. The effect of the bias β also depends on the learning rate: If τ is relatively large, the ES tends towards smaller values and shows irregular patterns. An increase of β changes the behavior. Larger learning rates seem to require larger biases in turn. Otherwise, a learning rate that is too small may lead to divergent behavior.

In order to investigate this behavior theoretically, one can apply the techniques developed in this section. In what follows, only a short sketch of the derivations is provided. Introducing $\beta > 0$ changes the raw moments of the log-normal distribution to $\overline{\varsigma^{*k}} = (\varsigma^{*(g)})^k \exp(k\beta)\exp(k^2\tau^2/2)$. Thus, if β is chosen sufficiently small, approximations with Taylor series as used in Appendix C.1 are still valid.



a) $\beta = 0$, $\tau = 0.1$

b) $\beta = 0.005$, $\tau = 0.1$

c) $\beta = 0$, $\tau = 0.5$

d) $\beta = 0.005$, $\tau = 0.5$

e) $\beta = 0.01$, $\tau = 0.5$

f) $\beta = 0.02$, $\tau = 0.5$

Figure 4.10: Dynamics of the normalized mutation strength of $(1, \lambda)$-ES. Shown are the results of typical ES runs on the quadratic sphere. The dimension of the search space is $N = 100$ and the noise strength is set to $\sigma_\epsilon = 1$.

Therefore, the derivation of the SAR remains the same. The only change occurs in the last step of the calculations leading from (C.21), p. 143, over (C.22) to (C.23), because the expectations of $[(\varsigma^* - \sigma^*)/\sigma^*]^k$ in (C.21) w.r.t. the log-normal density with bias $\beta = 0$ must be replaced. Finally SAR (C.23) becomes

$$\psi = \tau^2 \mathrm{e}^{\beta}\Big[\frac{1}{2} + \mathrm{e}^{\beta}(d^{(2)}_{1,\lambda} - 1)\frac{(\sigma^*)^2}{(\sigma^*_\epsilon)^2 + (\varsigma^*)^2} - \frac{\mathrm{e}^{\beta} c_{1,\lambda}(\sigma^*)^2}{\sqrt{(\sigma^*_\epsilon)^2 + (\sigma^*)^2}}\Big]. \tag{4.57}$$

Now the stationary points, i.e., the solutions of $\varphi^* = 0$ and $\psi = 0$ using (4.42) and (4.57) are determined. The condition $\varphi^* = 0$ gives $(\varsigma^{*(g)})^2 + (\sigma^{*(g)}_\epsilon)^2 = 4c^2_{1,\lambda}$. Inserting this into (4.57) leads to the stationary mutation strength

$$\begin{aligned} 0 &= \frac{1}{2} + \mathrm{e}^{\beta}(d^{(2)}_{1,\lambda} - 1)\frac{{\varsigma^*_{st}}^2}{{\sigma^*_{\epsilon\, st}}^2 + {\varsigma^*_{st}}^2} - \frac{\mathrm{e}^{\beta} c_{1,\lambda}{\varsigma^*_{st}}^2}{\sqrt{{\sigma^*_{\epsilon\, st}}^2 + {\varsigma^*_{st}}^2}} \\ \Rightarrow 0 &= \frac{1}{2} + \mathrm{e}^{\beta}{\varsigma^*_{st}}^2\left(\frac{(d^{(2)}_{1,\lambda} - 1)}{4c^2_{1,\lambda}} - \frac{1}{2}\right) \\ \Rightarrow \varsigma^*_{st} &= \frac{2c_{1,\lambda}\mathrm{e}^{-\frac{\beta}{2}}}{\sqrt{2(2c^2_{1,\lambda} + 1 - d^{(2)}_{1,\lambda})}}. \end{aligned} \tag{4.58}$$

Finally, the associated noise strength $\sigma^*_{\epsilon\, st} = 2c_{1,\lambda}\sqrt{1 - \frac{\mathrm{e}^{-\beta}}{2\left(2c^2_{1,\lambda}+1-d^{(2)}_{1,\lambda}\right)}}$ gives an estimate of the residual location error

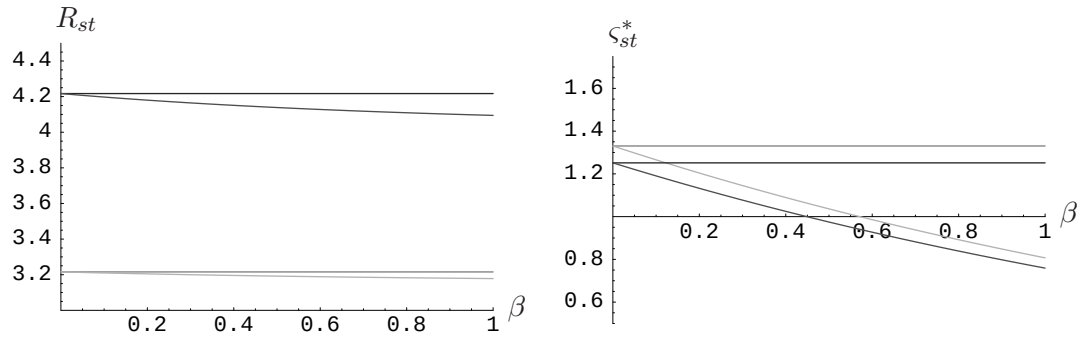
$$R^{\beta}_{st} = \sqrt[2]{\frac{\sigma_\epsilon N}{4c_{1,\lambda}}\sqrt{\frac{1}{1 - \frac{\mathrm{e}^{-\beta}}{2\left(2c^2_{1,\lambda}+1-d^{(2)}_{1,\lambda}\right)}}}}. \tag{4.59}$$

As can be shown numerically (see Fig. 4.11), as long as β is sufficiently small, the estimates (4.58) and (4.59) do not differ significantly from (4.51) and (4.53) obtained for $\beta = 0$.

Several caveats must be added here. It seems to be difficult to find a value of β that on the one hand raises the mutation strength sufficiently and on the other hand does not lead to a deterioration of the residual location error. In addition, the estimates only hold for sufficiently small β-values and they do not account for the interplay with the learning parameter τ. Considering the results of the experiments (see Fig. 4.12), one finds that in the case of larger β-values, i.e., here already for $\beta \geq 0.01$, the predicted mutation strength (4.58) is lower than the experimentally observed one. Also, the ES shows a significant greater sensitivity to the choice of β than predicted by (4.58). These deviations clearly indicate the limits of the deterministic analysis presented.

4.3 Intermediate ES on the Noisy Sphere

In this section, the analysis of evolution strategies on the noisy sphere is extended to ES with intermediate recombination. The approach mirrors that of Section 4.2 closely. Therefore, this section is kept short – only pointing out the differences between intermediate $(\mu/\mu_I, \lambda)$-ES and non-recombinative $(1, \lambda)$-ES.



a) residual location error b) mutation strength

Figure 4.11: Comparison of the predictions of the stationary mutation strength and the residual location error. Figure a) shows the prediction obtained by R_{st} (4.53) and R^{β}_{st} (4.59). Figure b) compares the mutation strengths (4.58) and (4.51). The dimension is $N = 100$ and the noise strength $\sigma_\epsilon = 1$. The gray lines indicate the results for $\lambda = 100$ whereas the black stand for $\lambda = 10$.

The Evolution Equations for Intermediate Evolution Strategies

As in the case of $(1, \lambda)$-ES, two variables are initially used to describe the system: The distance of the centroid to the optimizer $R^{(g)} = \|\langle \mathbf{y}^{(g)} \rangle - \hat{\mathbf{y}}\|$ and the mean of the mutation strength $\langle \varsigma^{(g)} \rangle$. To simplify the notations, the usual normalizations are introduced with $R := R^{(g)}$, $\sigma^* := (N/R)\langle \varsigma^{(g)} \rangle$, $\sigma^*_\epsilon := [N/(2R^2)]\sigma_\epsilon$, and $\varphi^*_R := (N/R)\varphi_R$. After normalizing, the normalized noise strength appears as an additional time-dependent variable. Using the same arguments as in the previous section, the analysis can be restricted to the study of the evolution of the noise and the mutation strength. Starting point of the analysis are therefore the deterministic evolution equations

$$\begin{aligned} \langle \varsigma^{*(g+1)} \rangle &= \sigma^* \left(\frac{1 + \psi(\sigma^*, \sigma^*_\epsilon)}{1 - \frac{\varphi^*_R(\sigma^*, \sigma^*_\epsilon)}{N}} \right) \\ \sigma^{*(g+1)}_\epsilon &= \frac{\sigma^*_\epsilon}{(1 - \frac{\varphi^*_R(\sigma^*, \sigma^*_\epsilon)}{N})^2}. \end{aligned} \tag{4.60}$$

The progress rate φ^*_R and SAR ψ are obtained as

$$\varphi^*_R(\sigma^*, \sigma^*_\epsilon) = \frac{\sigma^{*2}}{\sqrt{\sigma^{*2} + \sigma^{*2}_\epsilon}} c_{\mu/\mu,\lambda} - \frac{\sigma^{*2}}{2\mu} \tag{4.61}$$

for $N \to \infty$ and $\tau = 0$ (see Appendix B.1) and

$$\psi(\sigma^*, \sigma^*_\epsilon) = \tau^2 \Big(\frac{1}{2} + \frac{\sigma^{*2}}{\sigma^{*2} + \sigma^{*2}_\epsilon} e^{1,1,}_{\mu,\lambda} - \frac{\sigma^{*2}}{\sqrt{\sigma^{*2} + \sigma^{*2}_\epsilon}} c_{\mu/\mu,\lambda} \Big) \tag{4.62}$$

for $N \to \infty$ and $\tau \ll 1$ (see Appendix C.1.1).

4.3.1 The Evolution of Intermediate Evolution Strategies under Noise

Let us assume that the ES starts far away from the optimizer. Again, three phases can be distinguished:

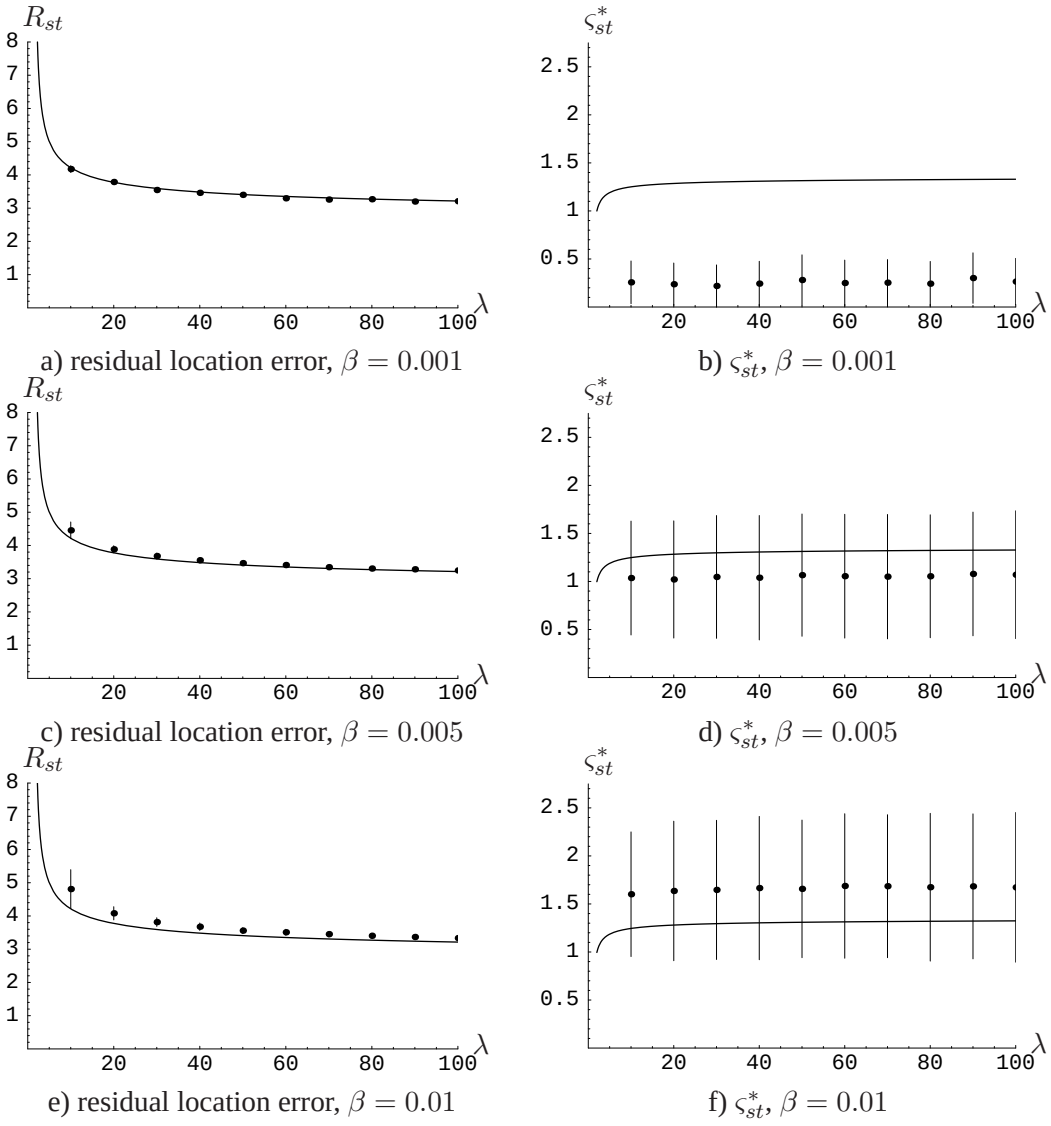


Figure 4.12: Comparison of the predictions of the stationary mutation strength (4.58) and the residual location error (4.59) with the results of experiments on the sphere function for some choices of β. The search space dimension is $N = 100$, the noise was set to $\sigma_\epsilon = 1$, and $\tau = 0.1$ was chosen as the learning parameter. Each data point was averaged over $500,000$ generations. The vertical bars indicate the measured standard deviations.

1. An initial stationary phase: As long as the ES is far away from the optimum, the influence of the noise is negligible. The ES behaves in a similar manner as in the undisturbed case and reaches a temporary stationary point of the normalized $\langle \varsigma^{*(g)} \rangle$-evolution.

2. A transitional phase: Since the ES progresses towards the optimum, the noise term gains more and more influence. This results in a loss of the stationary state and a nearly chaotic movement until the progress towards the optimum stops entirely (on average).

3. A final stationary phase: This is due to the fact that uniform additive noise hinders the ES

from approaching the optimum. Instead a new stationary state is reached with the distance R fluctuating around a positive value. The same holds for the mutation strength.

In the following, the different stationary states are characterized and the influence of recombination on the behavior is discussed.

The Initial Stationary State and the Influence of Recombination

Recall from Section 4.1, that the initial stationary state (after a transient time) is a stationary state of the $\langle \varsigma^{*(g)} \rangle$-evolution only. If $R \gg 1$, the influence of σ_ϵ^* is negligible and the results obtained in Sec. 4.1 apply:

1. The stationary state (4.11) reads
$$\varsigma_{st}^* = \mu c_{\mu/\mu,\lambda} \left((1 - N\tau^2) + \sqrt{(1 - N\tau^2)^2 + N\tau^2 \frac{1/2 + e_{\mu,\lambda}^{1,1}}{\mu c_{\mu/\mu,\lambda}^2}} \right).$$
2. The stationary mutation strength and progress rate depend strongly on the correct choice of the learning rate (4.23)
$$\tau_{opt} = \frac{1}{\sqrt{2N}} \sqrt{\frac{\mu c_{\mu/\mu,\lambda}^2}{\mu c_{\mu/\mu,\lambda}^2 - 1/2 - e_{\mu,\lambda}^{1,1}}}.$$
Otherwise, the progress may degrade significantly.
3. Nevertheless, recombination is beneficial since the maximal possible progress depends on the $\mu : \lambda$-ratio and is highest for $\mu \approx 0.27\lambda$.

As mentioned, this steady state is lost eventually. But choosing the $\mu : \lambda$ ratio and the learning rate accordingly ensures that the progress of the ES is nearly optimal as long as the stationary state persists.

The Final Stationary State

The influence of recombination on the final stationary state needs to be discussed. It was claimed in the previous section that recombination of the mutation strengths is beneficial since it introduces a bias. In contrast to non-recombinative ES, no loss of mutation strength control occurs. For an analysis, the respective stationary mutation strength and distance for recombinative ES need to be obtained. The approach followed mirrors the one taken in the previous section. The stationary mutation strength reads

$$\varsigma_{st}^* = \frac{2\mu c_{\mu/\mu,\lambda}}{\sqrt{4\mu c_{\mu/\mu,\lambda}^2 - 2e_{\mu,\lambda}^{1,1}}} \tag{4.63}$$

and is connected with the stationary noise strength

$$\sigma_{\epsilon\, st}^* = 2\mu c_{\mu/\mu,\lambda} \sqrt{\frac{4\mu c_{\mu/\mu,\lambda}^2 - 2e_{\mu,\lambda}^{1,1} - 1}{4\mu c_{\mu/\mu,\lambda}^2 - 2e_{\mu,\lambda}^{1,1}}}, \tag{4.64}$$

and the residual location error

$$R_{st} = \sqrt{\frac{\sigma_\epsilon N}{4\mu c_{\mu/\mu,\lambda}}} \sqrt[4]{\frac{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}}{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1}} \tag{4.65}$$

A derivation can be found in Appendix D.2.2. Note, in the case of the usual $\mu : \lambda$-ratios $2\mu c^2_{\mu/\mu,\lambda} \gg e^{1,1}_{\mu,\lambda}$ holds and the stationary mutation strength (4.63) scales with $\sqrt{\mu}$ – provided that λ is large. Therefore, recombination increases the normalized mutation strength.

The normalized noise strength (4.64) and residual location error (4.65) are given as a product of two factors: The first stems from demanding stationarity of the R-evolution and therefore $\varphi^*_R = 0$ which leads to the condition $\sigma^{*2} + \sigma^{*2}_\epsilon = 4\mu^2 c^2_{\mu/\mu,\lambda}$. Setting $\sigma^* = 0$ leads to the first factor in (4.64) and (4.65). The second factor gives the deviation due to the non-zero stationary mutation strength (4.63). But the normalized noise strength (4.64) does not deviate far from the maximally possible noise strength $2\mu c_{\mu/\mu,\lambda}$ if the offspring population is large.

A similar result holds for the location error. First of all, the minimal location error given by $\sqrt{\sigma_\epsilon N/(4\mu c_{\mu/\mu,\lambda})}$ is symmetric around its minimum for $\mu : \lambda = 0.5$. The region around the minimum is relatively flat and nearly optimal distances are obtainable for $\mu : \lambda \in [0.2 - 0.7]$. The ES with (4.65) deviates from this optimal value, though, which is due to the non-zero mutation strength. However, this deviation is small. Recombination may lower (4.65), so that it gets even closer to the minimal location error: For relatively large λ-values and if μ is neither close to one or to λ, the following approximate steady state values hold

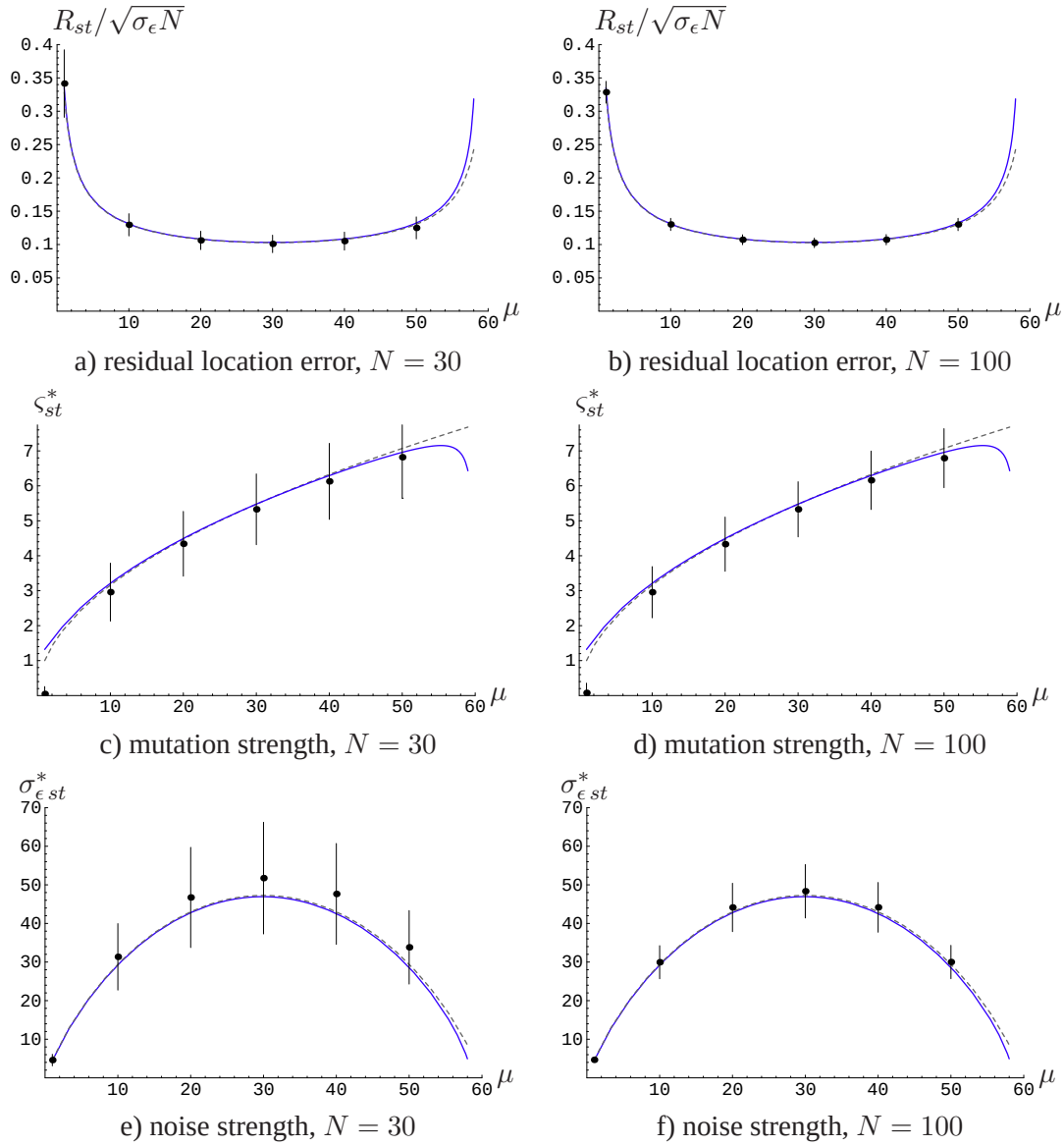
$$\begin{pmatrix} \sigma^*_{\epsilon\, app} \\ \varsigma^*_{app} \\ R_{app} \end{pmatrix} = \begin{pmatrix} 2\mu c_{\mu/\mu,\lambda} \\ \sqrt{\mu} \\ \sqrt{\frac{\sigma_\epsilon N}{4\mu c_{\mu/\mu,\lambda}}} \end{pmatrix}. \tag{4.66}$$

To summarize, recombination on the noisy sphere is beneficial: Recombination of the object variables enables a closer approach to the actual optimum. Recombination of the mutation strengths enforces a positive stationary mutation strength and does not result in a loss of step-size control. In addition, the deviations from the minimal location error are small and improve for $\mu : \lambda$-ratios in the interval usually recommended.

Simulations

It remains to compare the predictions by (4.63), (4.64), and (4.65) with the results of experiments. In the experiments, $(\mu/\mu_I, 60)$-ES were used. The mutation strength and distance were aggregated over $400,000$ generations in the steady state regime for $N = 100$ and $N = 30$. The experiments were conducted using the log-normal distribution. Figure 4.13 compares the predictions with the experimental results. Figure 4.13 also depicts the approximated stationary state values (4.66) (dashed gray line). These estimates serve well to predict the experimental results for parent numbers between $\mu = 10$ and $\mu = 40$. As it can be seen, the agreement between experiment and prediction is good, in general. Note, though, the mutation strength is overestimated as a rule. As it can be seen, the dependency of the prediction quality on the search space dimensionality is relatively weak. Even

for $N = 30$ good estimates can be obtained. Only in the case of the noise strength, the increase of the dimensionality leads to a better prediction quality. The mutation strength and the location error are predicted well even for $N = 30$. However, the standard deviations are smaller in the higher dimensional search space.



a) residual location error, $N = 30$

b) residual location error, $N = 100$

c) mutation strength, $N = 30$

d) mutation strength, $N = 100$

e) noise strength, $N = 30$

f) noise strength, $N = 100$

Figure 4.13: Comparison of the predictions of the residual local error (4.65), noise strength (4.64), and mutation strength (4.63) with the results of experiments. The dotted gray lines denote the approximate stationary state values (4.66). All data points are sampled over $400,000$ generations in the steady state. The error bars indicate the size of the standard deviations. The search space dimensionality was set to $N = 100$ and $N = 30$. The noise strength was set to $\sigma_\epsilon = 1$.

4.4 Including the Fluctuation Part: A Second Order Approach

In this section, the analysis is extended to evolution equations comprising the perturbation parts. The aim is to provide a better estimate of the mean value dynamics and stationary state behavior of self-adaptive ES. As introduced in Chapter 3, the unknown distribution of the perturbation parts is approximated using an Edgeworth series expansion. The expansion is cut off after the first term assuming that higher order cumulants do not have a significant influence in the scenario under investigation. That is to say, the distribution is assumed to be sufficiently Gaussian so that the deviations from the normal distribution do not have significant effects in the mean value dynamics of evolution strategies.

4.4.1 The Evolution Equations

The analysis is started considering the evolution equations

$$R^{(g+1)} = R - \frac{\varphi_R(\sigma)}{N} - \epsilon_R(R,\sigma) \tag{4.67}$$

$$\langle \varsigma^{(g+1)} \rangle = \sigma\Big(1 + \psi(\sigma)\Big) + \epsilon_\sigma(R,\sigma). \tag{4.68}$$

First of all, the usual normalizations are introduced with $\varphi_R^* := N/R\varphi_R$, $\sigma^* := N/R\sigma$, $\epsilon_\sigma^* := \epsilon_\sigma/\sigma^*$, and $\epsilon_R^* := \epsilon_R/R$. Equations (4.67) and (4.68) change to

$$R^{(g+1)} = R\Big(1 - \frac{\varphi_R^*(\sigma^*)}{N} + \epsilon_R^*(R,\sigma^*)\Big) \tag{4.69}$$

$$\langle \varsigma^{*(g+1)} \rangle = \sigma^*\Big(\frac{1 + \psi(\sigma^*) + \epsilon_\sigma^*(R,\sigma^*)}{1 - \frac{\varphi_R^*(\sigma^*)}{N} + \epsilon_R^*(R,\sigma^*)}\Big). \tag{4.70}$$

Recall, the perturbation terms are modeled with

$$\begin{aligned} \epsilon_R^* &= \frac{D_\varphi}{R}\mathcal{N}(0,1) + \ldots = \frac{\sqrt{\varphi_R^{(2)} - \varphi_R^2}}{R}\mathcal{N}(0,1) + \ldots \\ &= \frac{1}{N}\sqrt{\varphi^{*(2)}_R - \varphi^{*\,2}_R}\mathcal{N}(0,1) + \ldots \end{aligned} \tag{4.71}$$

$$\epsilon_\sigma^* = \frac{D_\psi}{\sigma^*}\mathcal{N}(0,1) + \ldots = \sqrt{\psi^{(2)} - \psi^2}\mathcal{N}(0,1) + \ldots. \tag{4.72}$$

The inclusion of the perturbation parts changes the equations. Whereas it was sufficient in the deterministic approach just to calculate the progress rate (4.8),

$$\varphi_R^*(\sigma^*) = c_{\mu/\mu,\lambda}\sigma^* - \frac{\sigma^{*2}}{2\mu},$$

and the SAR (4.9),

$$\psi(\sigma^*) = \tau^2\Big(1/2 + e^{1,1}_{\mu,\lambda} - c_{\mu/\mu,\lambda}\sigma^*\Big),$$

the second order approach requires the second order progress rate $\varphi_R^{*(2)}$ and the second order SAR $\psi^{(2)}$. Both are obtained in the appendix (see Appendices B.1.3 and C.5). Note the following: The second order progress rate and the square of the progress rate average out. Thus for $(\mu/\mu_I, \lambda)$-ES the

evolution equation with the perturbation part approximated with a normal distribution degrades to the deterministic case. This does not occur in the case of the evolution of the mutation strength. In this case the variance must be determined. The influence of the square of the first order SAR is of order $\mathcal{O}(\tau^4)$ and only the second order SAR (C.159), p. 186,

$$\psi^{(2)} = \frac{\tau^2}{\mu} \tag{4.73}$$

will be taken into account leading finally to a linear term in τ.

4.4.2 The Mean Value Dynamics of the Mutation Strength

Before starting, let us simplify the notations setting $\varsigma^* := \langle \varsigma^{*(g+1)} \rangle$. As said before, the moments of the distribution $p(\varsigma^*)$ starting with the expectation have to be obtained. At this moment the transition densities are not needed. Before starting with the calculations, the evolution equation (4.70) is simplified which requires some assumptions. First: Assuming that $\varphi^* \ll N$ for all ς^* with positive measure, the function $1/(1-\varphi_R^*/N)$ is expanded into

$$\frac{1}{1-\frac{\varphi_R^*(\sigma^*)}{N}} = 1 + \frac{\varphi_R^*(\sigma^*)}{N}\Big(\frac{1}{1-\frac{\varphi_R^*(\sigma^*)}{N}}\Big) = 1 + \frac{\varphi_R^*(\sigma^*)}{N} + \mathcal{O}\bigg(\Big(\frac{\varphi_R^*(\sigma^*)}{N}\Big)^2\bigg). \tag{4.74}$$

Equation (4.70) changes to

$$\varsigma^* = \sigma^*\Big(1 + \psi(\sigma^*) + \sqrt{\psi^{(2)} - \psi^2}\mathcal{N}(0,1)\Big)\Big(1 + \frac{\varphi_R^*(\sigma^*)}{N}\Big). \tag{4.75}$$

Under the further conditions that $\psi\varphi_R^* \ll N$ and that the realizations of $\sqrt{\psi^{(2)} - \psi^2}\mathcal{N}(0,1)\varphi_R^*$ are generally smaller than N

$$\varsigma^* = \sigma^*\Big(1 + \psi(\sigma^*) + \sqrt{\psi^{(2)} - \psi^2}\mathcal{N}(0,1) + \frac{\varphi_R^*(\sigma^*)}{N}\Big) \tag{4.76}$$

is obtained. Using the N-independent variants, the progress rate and the self-adaptation response are given by Eqs. (4.8) and (4.9). The expectation of (4.76)

$$\mathrm{E}[\varsigma^*] = \overline{\sigma^*}\bigg(1 + \tau^2\Big(\frac{1}{2} + e_{\mu,\lambda}^{1,1}\Big)\bigg) - \overline{\sigma^{*2}}c_{\mu/\mu,\lambda}\tau^2\Big(1 - \frac{1}{N\tau^2}\Big) - \tau^2\frac{\overline{\sigma^{*3}}}{2\mu N\tau^2} \tag{4.77}$$

depends on the past values through higher order moments. As a result, the expectations of ς^{*2} and ς^{*3} are needed. It will be shown that they in turn depend on the past through higher order terms. The expectation of the square is given by

$$\mathrm{E}[\varsigma^{*2}] = \overline{\sigma^{*2}}\Big(1 + \tau^2[1 + 2e_{\mu,\lambda}^{1,1} + \frac{1}{\mu}]\Big) - 2\overline{\sigma^{*3}}\tau^2 c_{\mu/\mu,\lambda}\Big(1 - \frac{1}{N\tau^2}\Big) - \overline{\sigma^{*4}}\frac{\tau^2}{\mu}\frac{1}{N\tau^2}. \tag{4.78}$$

The expectation $\mathrm{E}[\varsigma^{*3}]$ can be approximated with

$$\begin{aligned}\mathrm{E}[\varsigma^{*3}] = {} & \overline{\sigma^{*3}}\bigg(1 + 3\frac{\tau^2}{\mu} + 3\Big(1 + \frac{\tau^2}{\mu}\Big)\tau^2\Big(\frac{1}{2} + e_{\mu,\lambda}^{1,1}\Big)\bigg) - 3\Big(1 + \frac{\tau^2}{\mu}\Big)\tau^2 c_{\mu/\mu,\lambda}\overline{\sigma^{*4}}\Big(1 - \frac{1}{N\tau^2}\Big) \\ & -3\Big(1 + \frac{\tau^2}{\mu}\Big)\tau^2\frac{\overline{\sigma^{*5}}}{2\mu N\tau^2}.\end{aligned} \tag{4.79}$$

4.4.3 The ES in the Stationary State

Let us now address the stationary state behavior. As the result, $\mathrm{E}[\varsigma^*] = \mathrm{E}[\sigma^*] = \mathrm{E}[\sigma^*_\infty]$ holds. Equations (4.77), (4.78), and (4.79) lead to the non-linear equations

$$0 = \overline{\sigma^*_\infty}\left(\frac{1}{2} + e^{1,1}_{\mu,\lambda}\right) - \overline{\sigma^*_\infty{}^2} c_{\mu/\mu,\lambda}\left(1 - \frac{1}{N\tau^2}\right) - \frac{\overline{\sigma^*_\infty{}^3}}{2\mu N\tau^2} \tag{4.80}$$

$$0 = \overline{\sigma^*_\infty{}^2}\left(1 + 2e^{1,1}_{\mu,\lambda} + \frac{1}{\mu}\right) - 2\overline{\sigma^*_\infty{}^3} c_{\mu/\mu,\lambda}\left(1 - \frac{1}{N\tau^2}\right) - \overline{\sigma^*_\infty{}^4}\frac{1}{\mu N\tau^2} \tag{4.81}$$

$$\begin{aligned} 0 = {} & \overline{\sigma^*_\infty{}^3}\left(\frac{1}{\mu} + \left(1 + \frac{\tau^2}{\mu}\right)\left(\frac{1}{2} + e^{1,1}_{\mu,\lambda}\right)\right) - \left(1 + \frac{\tau^2}{\mu}\right) c_{\mu/\mu,\lambda}\overline{\sigma^*_\infty{}^4}\left(1 - \frac{1}{N\tau^2}\right) \\ & - \left(1 + \frac{\tau^2}{\mu}\right)\frac{\overline{\sigma^*_\infty{}^5}}{2\mu N\tau^2} \end{aligned} \tag{4.82}$$

which could be solved if the invariant density of σ^*_∞ were known. Instead of determining the invariant density, a so-called *ansatz* is used. The ansatz consists in using a specific distribution to model the behavior of the mutation strength in the stationary state. In this section, a log-normal distribution in the stationary state is assumed, i.e., the moments are of the general form $\overline{\sigma^*_\infty{}^k} = S\exp(k^2t^2/2)$. The constants S and t have to be determined which is done in the next paragraph.

A Log-Normal Distribution in the Stationary State

Plugging $\overline{\sigma^*_\infty{}^k} = S\exp(k^2t^2/2)$ into Eqs. (4.80)-(4.82) leads to

$$0 = \frac{1}{2} + e^{1,1}_{\mu,\lambda} - S\mathrm{e}^{\frac{3}{2}t^2} c_{\mu/\mu,\lambda}\left(1 - \frac{1}{N\tau^2}\right) - S^2\mathrm{e}^{4t^2}\frac{1}{2\mu N\tau^2} \tag{4.83}$$

$$0 = 1 + 2e^{1,1}_{\mu,\lambda} + \frac{1}{\mu} - S\mathrm{e}^{\frac{5}{2}t^2} c_{\mu/\mu,\lambda} 2\left(1 - \frac{1}{N\tau^2}\right) - S^2\mathrm{e}^{6t^2}\frac{1}{\mu N\tau^2} \tag{4.84}$$

$$\begin{aligned} 0 = {} & \frac{1}{\mu} + \left(1 + \frac{\tau^2}{\mu}\right)\left(\frac{1}{2} + e^{1,1}_{\mu,\lambda}\right) - \left(1 + \frac{\tau^2}{\mu}\right) S\mathrm{e}^{\frac{7}{2}t^2} c_{\mu/\mu,\lambda}\left(1 - \frac{1}{N\tau^2}\right) \\ & - \left(1 + \frac{\tau^2}{\mu}\right) S^2\mathrm{e}^{8t^2}\frac{1}{2\mu N\tau^2} \end{aligned} \tag{4.85}$$

with unknown parameters S and t. Note that the equations above lead to a nonlinear system the general solution of which cannot be provided analytically. It is possible, though, to obtain numerical solutions. To this end, MATHEMATICA was used to determine the solutions of the first two equations.

Comparison with Experiments Figure 4.14 shows histogram plots of some $(\mu/\mu_I, 60)$-ES for the search space dimensionality $N = 100$. The relative frequencies were sampled over $500,000$ generations in the stationary state regime. Due to the fast convergence of the ES, the learning rate was set to $\tau = 0.01$. Also depicted are the pdfs of a Gaussian and a log-normal distribution using the sample mean and variance. As can be seen, the log-normal distribution serves relatively well as reference function for the unknown steady state distributions.

Figure 4.15 shows the stationary mutation strength obtained using (4.85) in comparison with the stationary mutation strength observed in experiments. The mutation strength is depicted as a function of the learning rate. The experiments were conducted using a $(\mu/\mu_I, 60)$-σSA-ES. Each data point was sampled over at least $100,000$ generations in the stationary state. It should be mentioned here that since the convergence velocity depends on the learning rate, the duration of the stationary phase

may be short due to the fast reduction of the distance to the optimizer to zero. As it can be seen in Fig. 4.15, the quality of the prediction depends strongly on the search space dimensionality which is due to using the N-independent formulae in the derivations.

In addition to the mutation strength, Fig. 4.15 compares the predicted stationary progress rate with the result of experiments. The predicted stationary progress rate was obtained by inserting the moments of the stationary mutation strength into (4.8). Again, there are considerable deviations in the smaller dimensional search space, but the prediction quality improves with the dimensionality.

The Influence of Fluctuations in the Second-Order Approach In the following, a closer look is taken at the obtained stationary mutation strength. Similarly to [23], Eqs. (4.83) to (4.85) are rewritten in terms of $s^*_\infty := S\,\mathrm{e}^{t^2/2}$

$$0 = \frac{1}{2} + e^{1,1}_{\mu,\lambda} - s^*_\infty \mathrm{e}^{t^2} c_{\mu/\mu,\lambda}\left(1 - \frac{1}{N\tau^2}\right) - {s^*_\infty}^2 \mathrm{e}^{3t^2}\frac{1}{2\mu N\tau^2} \tag{4.86}$$

$$0 = 1 + 2e^{1,1}_{\mu,\lambda} + \frac{1}{\mu} - s^*_\infty \mathrm{e}^{2t^2} c_{\mu/\mu,\lambda} 2\left(1 - \frac{1}{N\tau^2}\right) - {s^*_\infty}^2 \mathrm{e}^{5t^2}\frac{1}{\mu N\tau^2} \tag{4.87}$$

$$\begin{aligned} 0 = {} & \frac{1}{\mu} + \left(1 + \frac{\tau^2}{\mu}\right)\left(\frac{1}{2} + e^{1,1}_{\mu,\lambda}\right) - s^*_\infty \mathrm{e}^{3\tau^2} c_{\mu/\mu,\lambda}\left(1 - \frac{1}{N\tau^2}\right)\left(1 + \frac{\tau^2}{\mu}\right) \\ & - \frac{1 + \frac{\tau^2}{\mu}}{2\mu N\tau^2} {s^*_\infty}^2 \mathrm{e}^{7t^2}. \end{aligned} \tag{4.88}$$

Equation (4.86) can be used to give the stationary mutation strength as a function of t

$$s^*_\infty = \mu c_{\mu/\mu,\lambda} \mathrm{e}^{-2t^2}\left((1 - N\tau^2) + \sqrt{(1 - N\tau^2)^2 + \frac{\frac{1}{2} + e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}} 2N\tau^2 \mathrm{e}^{t^2}}\right). \tag{4.89}$$

The equation obtained is analogous to the case of $(1, \lambda)$-ES [21]. The mutation strength differs from the mutation strength (4.11)

$${\varsigma^*_{stat}}^{\text{det}} = \mu c_{\mu/\mu,\lambda}\left((1 - N\tau^2) + \sqrt{(1 - N\tau^2)^2 + \frac{\frac{1}{2} + e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}} 2N\tau^2}\right)$$

obtained by using the deterministic approach in two terms: One inside the root, the other a general multiplier. It is easy to see that the general influence of the multiplier $\exp(-2t^2)$ outweighs the effect by the addend $\exp(t^2)$. For this reason, Eq. (4.89) leads to lower mutation strengths than (4.11) As Beyer pointed out for $(1, \lambda)$-ES, experimentally observed mutation strengths are lower than the deterministic estimates. This can be traced back to the neglected influence of the fluctuations during the derivation of the estimate (see [21] or [23, p. 315f.]). Equation (4.89) corrects the estimate.

The progress rate remains to be considered. The expected progress rate is given by

$$\overline{\varphi^*_R(\varsigma^*)} = c_{\mu/\mu,\lambda}\overline{\varsigma^*} - \frac{\overline{\varsigma^{*2}}}{2\mu}. \tag{4.90}$$

Since $\overline{\varsigma^*}^2 \neq \overline{\varsigma^{*2}} = \mathrm{Var}[\varsigma^*] + \overline{\varsigma^*}^2$, an additional loss term, the variance, lowers the expected progress rate [21]

$$\overline{\varphi^*_R(\varsigma^*)} = c_{\mu/\mu,\lambda}\overline{\varsigma^*} - \frac{\overline{\varsigma^*}^2}{2\mu} - \frac{\mathrm{Var}[\varsigma^*]}{2\mu}. \tag{4.91}$$

Therefore, the theoretical maximal progress rate $\mu c^2_{\mu/\mu,\lambda}/2$ is not attainable [21]. The question that remains is the following: How can the fluctuations be reduced so that the ES works approximately with its optimal progress rate? In [23] several possible means were described. The remainder of the section is devoted to the question how recombination of the object vectors and mutation strengths influences the fluctuations. The analysis makes use of the aforementioned ansatz, assuming a log-normal distribution of the mutation strength in the stationary state. It should be noted that recombination does not only influence the variance but of course the expectation of ς^* and the progress rate.

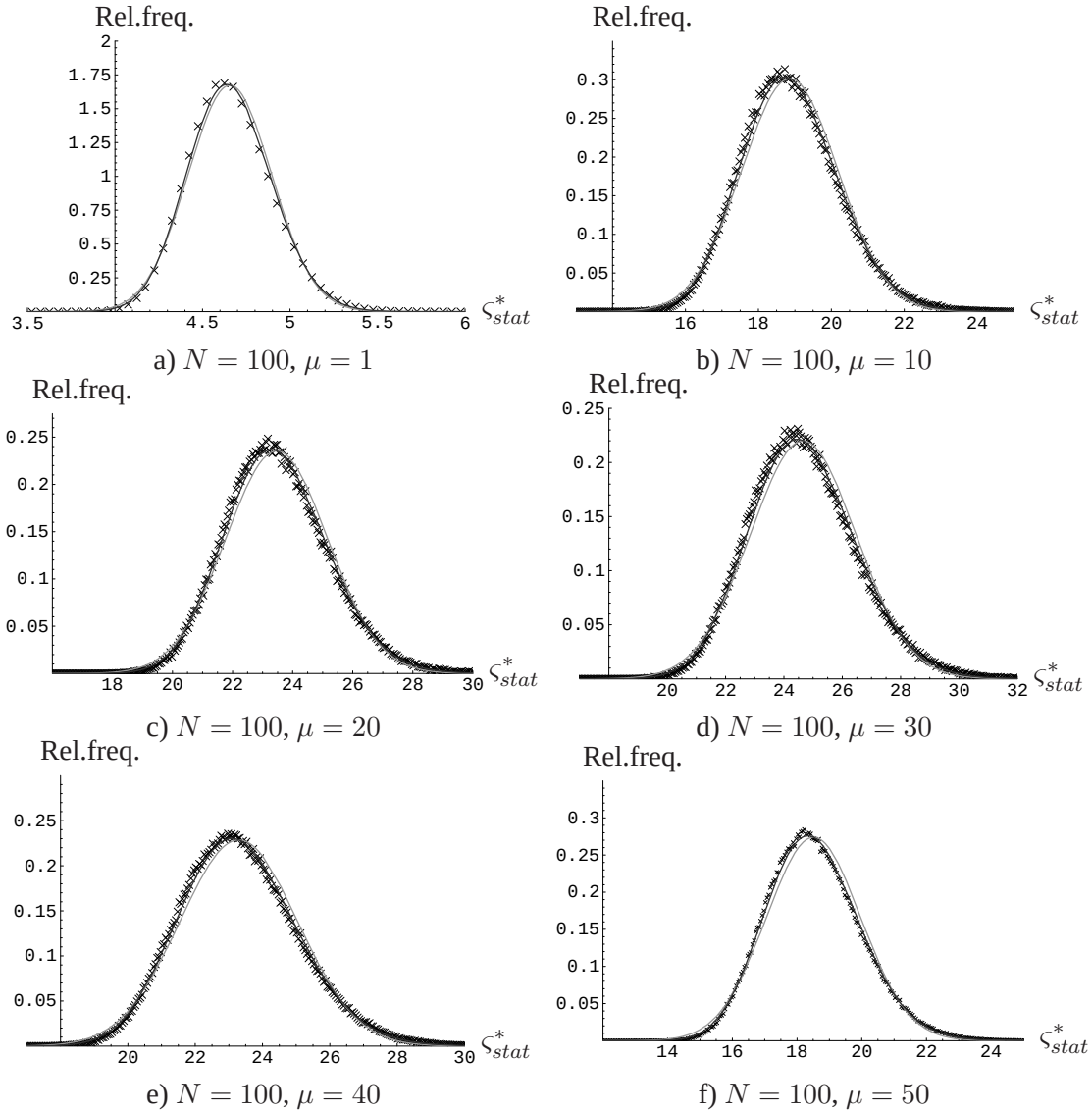
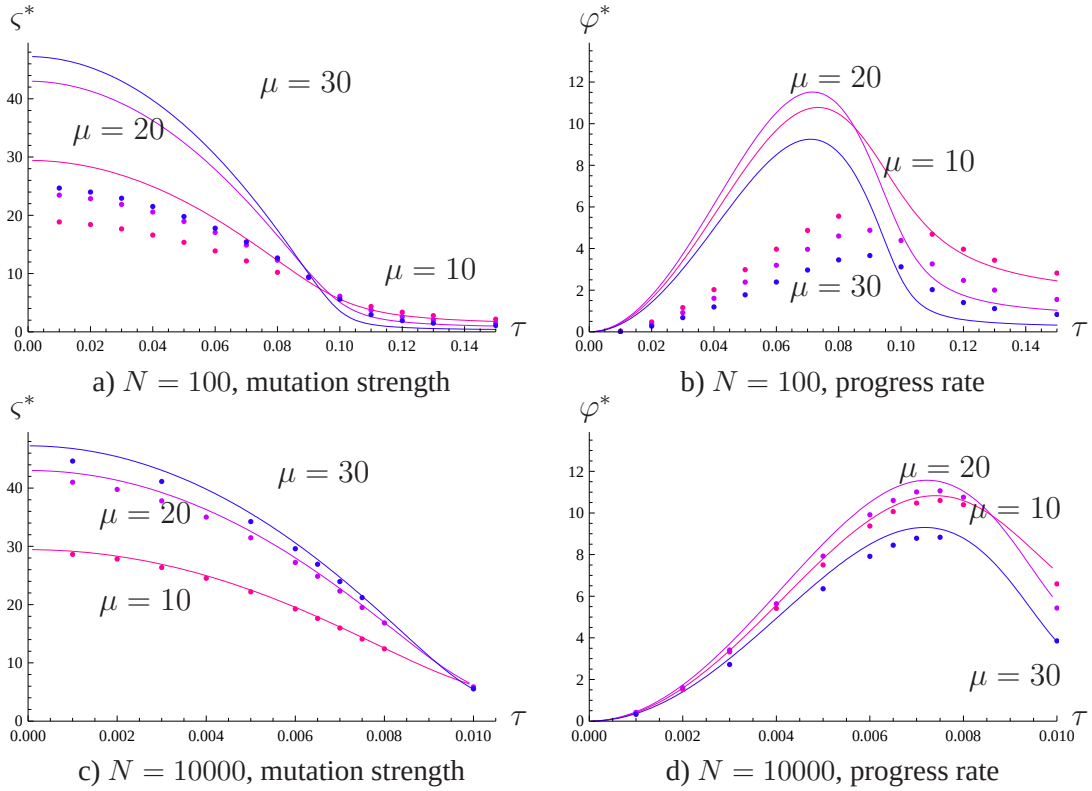


Figure 4.14: Relative frequencies of the normalized mutation strength in the stationary state. The search space dimensionality is $N = 100$. The experiments were conducted using $(\mu/\mu_I, 60)$-ES and a learning rate of 0.01. The lines indicate the density functions of log-normal distributions (black) and normal distributions (gray). The density function were obtained by inserting the experimentally found moments.



a) $N = 100$, mutation strength b) $N = 100$, progress rate

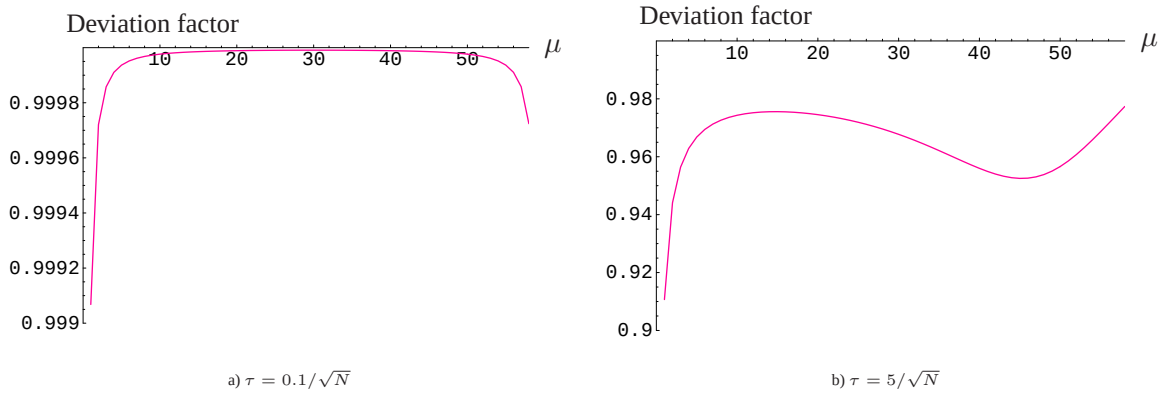
c) $N = 10000$, mutation strength d) $N = 10000$, progress rate

Figure 4.15: Stationary normalized mutation strength and progress rate as a function of τ for some $(\mu/\mu_I, 60)$-ES.

Fluctuations and Recombination Before starting, consider some results obtained numerically for two choices of the learning rate. Figure 4.16 shows how far the results from the second-order approach deviate from the those obtained using the deterministic approach. Not surprisingly, the deviations increase with the learning rate. As Fig. 4.16 reveals using recombination causes a better agreement between the two approaches. For the smaller learning rate, the main difference is between no recombination and recombination, the higher learning rate indicates an interval where the relative deviations of the first-order from the second-order approach are approximately minimal. The interval for $(\mu/\mu_I, 60)$-ES lies roughly between $\mu = 12$ and $\mu = 20$, giving a $\mu : \lambda$-ratio of approximately $0.2 - 1/3$. In the following, two special cases are considered which allow for an analytical treatment.

Limit Case of $N\tau^2 \to \infty$ Let us first consider the limit case of $N\tau \to \infty$. Starting from Eqs. (4.86) and (4.87), i.e.,

$$\begin{aligned} 0 &= \frac{1}{2} + e^{1,1}_{\mu,\lambda} - s^*_\infty \mathrm{e}^{t^2} c_{\mu/\mu,\lambda}\left(1 - \frac{1}{N\tau^2}\right) - {s^*_\infty}^2 \mathrm{e}^{3t^2} \frac{1}{2\mu N\tau^2} \\ 0 &= 1 + 2e^{1,1}_{\mu,\lambda} + \frac{1}{\mu} - s^*_\infty \mathrm{e}^{2t^2} c_{\mu/\mu,\lambda} 2\left(1 - \frac{1}{N\tau^2}\right) - {s^*_\infty}^2 \mathrm{e}^{5t^2} \frac{1}{\mu N\tau^2} \end{aligned} \tag{4.92}$$



a) $\tau = 0.1/\sqrt{N}$ b) $\tau = 5/\sqrt{N}$

Figure 4.16: The deviation from the deterministic prediction as a function of the parent number μ. The results were obtained numerically from Eqs. (4.83) and (4.84) for two choices of τ. The search space dimensionality is $N = 10,000$.

it will be shown that the system can be easily solved for $N\tau^2 \to \infty$. Computing the limit gives

$$\begin{aligned} 0 &= \frac{1}{2} + e^{1,1}_{\mu,\lambda} - s^*_\infty \mathrm{e}^{t^2} c_{\mu/\mu,\lambda} \\ 0 &= 1 + 2e^{1,1}_{\mu,\lambda} + \frac{1}{\mu} - s^*_\infty \mathrm{e}^{2t^2} 2c_{\mu/\mu,\lambda}. \end{aligned} \tag{4.93}$$

Thus, two equations describing s^*_∞ can be obtained

$$s^*_\infty = \frac{\frac{1}{2} + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}} \mathrm{e}^{-t^2} \tag{4.94}$$

$$s^*_\infty = \mathrm{e}^{-2t^2} \left(\frac{\frac{1}{2} + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}} + \frac{1}{2\mu c_{\mu/\mu,\lambda}} \right). \tag{4.95}$$

They can be used to determine the value of $\exp(-t^2)$

$$\begin{aligned} \frac{1/2 + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}} \mathrm{e}^{-t^2} &= \left(\frac{2\mu\left(1/2 + e^{1,1}_{\mu,\lambda}\right) + 1}{2\mu c_{\mu/\mu,\lambda}} \right) \mathrm{e}^{-2t^2} \\ \Rightarrow \mathrm{e}^{-t^2} &= \frac{2\mu(1/2 + e^{1,1}_{\mu,\lambda})}{2\mu(1/2 + e^{1,1}_{\mu,\lambda}) + 1}. \end{aligned} \tag{4.96}$$

In the following it is shown that recombination, i.e., switching from $\mu = 1$ to $\mu > 1$, may increase the factor (4.96). First of all, the function that appears in (4.96) is of the general form $f(x) = x/(1+x)$ which is a strictly increasing function with $f(0) = 0$ and $\lim_{x\to\infty} f(x) = 1$. That f is strictly increasing can be shown using the first derivative

$$f'(x) = \frac{1}{1+x} - \frac{x}{(1+x)^2} = \frac{1}{1+x}\left(\frac{1+x}{1+x} - \frac{x}{1+x}\right) = \frac{1}{1+x}. \tag{4.97}$$

While the progress coefficient $e^{1,1}_{\mu,\lambda}$ decreases with μ, the increase of 2μ outweighs that decrease as long as μ does not increase too far. Note, the coefficient $e^{1,1}_{\mu,\lambda}$ passes zero for $\mu = 0.5\lambda$. As numerical

comparisons show, the minimizer of (4.96) lies roughly in the region of $\mu \approx 0.2 - 1/3\lambda$. As a result, for $N\tau^2 \to \infty$, the prediction obtained using the second-order approach does not deviate far from $(1/2+e^{1,1}_{\mu,\lambda})/c_{\mu/\mu,\lambda}$ – the deterministic result. Let us shortly consider the stationary mutation strength and the progress rate. Using (4.94) and (4.96) the stationary mutation strength for $N\tau^2 \to \infty$ is obtained as

$$s^*_\infty = \frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\left(\frac{2\mu(1/2+e^{1,1}_{\mu,\lambda})}{2\mu(1/2+e^{1,1}_{\mu,\lambda}+1)}\right) \tag{4.98}$$

As stated, first factor in (4.98) equals the deterministic result, i.e., the zero of the SAR, whereas the second factor constitutes a correction factor due to taking the fluctuations into account. Plugging the mutation strength (4.98) and the inverse of (4.96) into the progress rate (B.24), $\varphi^*(\varsigma^*) = c_{\mu/\mu,\lambda}\varsigma^* - \varsigma^{*2}/(2\mu)$, leads to the stationary progress rate for $N\tau^2 \to \infty$

$$\begin{aligned}\varphi^*_\infty &= c_{\mu/\mu,\lambda}s^*_\infty - \frac{(s^*_\infty)^2\mathrm{e}^{t^2}}{2\mu} \\ &= \varphi^*\Big(\frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\Big)\left(\frac{2\mu(1/2+e^{1,1}_{\mu,\lambda})}{2\mu(1/2+e^{1,1}_{\mu,\lambda})+1}\right).\end{aligned} \tag{4.99}$$

The deterministic prediction of the progress rate is reduced by the same factor as the prediction of the mutation strength.

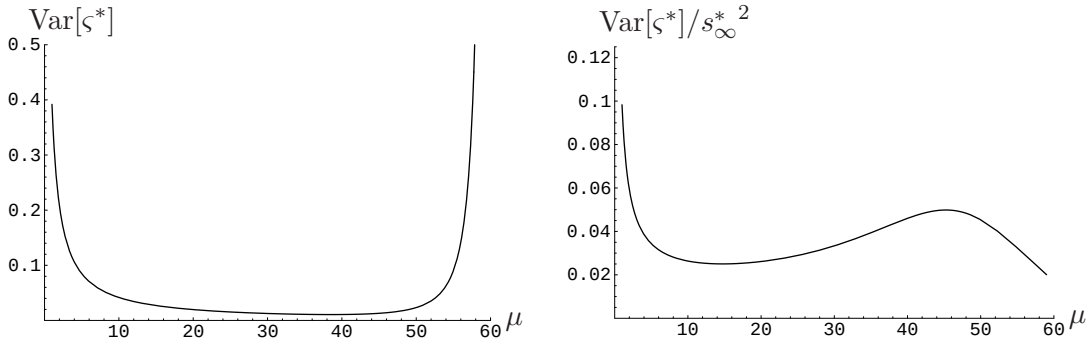
The variance $\mathrm{Var}[\varsigma^*]$, which reduces the progress rate, can be easily obtained as

$$\begin{aligned}\mathrm{Var}[\varsigma^*] = \overline{\varsigma^{*2}} - \overline{\varsigma^*}^2 &= {s^*_\infty}^2\mathrm{e}^{t^2} - {s^*_\infty}^2 = {s^*_\infty}^2\Big(e^{t^2}-1\Big) \\ &= \Big(\frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\Big)^2\Big(\frac{2\mu(1/2+e^{1,1}_{\mu,\lambda})}{2\mu(1/2+e^{1,1}_{\mu,\lambda})+1}\Big)^2\Big(\frac{2\mu(1/2+e^{1,1}_{\mu,\lambda})+1}{2\mu(1/2+e^{1,1}_{\mu,\lambda})}-1\Big) \\ &= \Big(\frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\Big)^2\frac{2\mu(1/2+e^{1,1}_{\mu,\lambda})}{(2\mu(1/2+e^{1,1}_{\mu,\lambda})+1)^2}.\end{aligned} \tag{4.100}$$

The variance depends of course on the size of s^*_∞. The absolute size of the variance reduces considerably once recombination comes into play. As Fig. 4.17a) shows, this reflects the behavior of s^*_∞ to some extent. The expectation s^*_∞ drops sharply when switching from $\mu = 1$ to $\mu > 1$. Considering the relative variance instead reveals that there is a minimizer between $\mu = 10$ and $\mu = 20$ (see Fig. 4.17b)). Therefore, the deviation of the progress rate is minimal for μ, λ-combinations that are normally recommended. Of course, again, this effect of reducing the variance is shown for $N\tau^2 \to \infty$, only.

The Case of $N\tau^2 = 1$ Let us now consider the special case of $N\tau^2 = 1$. Again, analytical solutions are easily obtained. Setting $N\tau^2 = 1$, Equations (4.86) - (4.88) describing s^*_∞ change to

$$0 = \frac{1}{2} + e^{1,1}_{\mu,\lambda} - {s^*_\infty}^2\mathrm{e}^{3t^2}\frac{1}{2\mu} \tag{4.101}$$



a) Variance b) Relative variance w.r.t. ${s^*_\infty}^2$

Figure 4.17: The variance of ς^* (4.100) in the stationary state as a function of μ for $N\tau^2 \to \infty$. Figure b) shows the variance w.r.t. ${s^*_\infty}^2$ (4.98).

$$0 = 1 + 2e^{1,1}_{\mu,\lambda} + \frac{1}{\mu} - {s^*_\infty}^2 \mathrm{e}^{5t^2} \frac{1}{\mu} \tag{4.102}$$

$$0 = \frac{1}{\mu} + \Big(1 + \frac{\tau^2}{\mu}\Big)\Big(\frac{1}{2} + e^{1,1}_{\mu,\lambda}\Big) - \frac{1 + \frac{\tau^2}{\mu}}{2\mu} {s^*_\infty}^2 \mathrm{e}^{7t^2}. \tag{4.103}$$

Only the first two equations are needed to determine s^*_∞. Rewriting Eq. (4.101) and (4.102) gives

$${s^*_\infty}^2 = 2\mu\Big(\frac{1}{2} + e^{1,1}_{\mu,\lambda}\Big)\mathrm{e}^{-3t^2} \tag{4.104}$$

$${s^*_\infty}^2 = \Big(2\mu\Big(\frac{1}{2} + e^{1,1}_{\mu,\lambda}\Big) + 1\Big)\mathrm{e}^{-5t^2}. \tag{4.105}$$

Thus, s^*_∞ can be obtained by

$$\mathrm{e}^{-t^2} = \sqrt{\frac{2\mu(1/2 + e^{1,1}_{\mu,\lambda})}{2\mu(1/2 + e^{1,1}_{\mu,\lambda}) + 1}} \tag{4.106}$$

and

$$\begin{aligned} s^*_\infty &= \sqrt{2\mu(1/2 + e^{1,1}_{\mu,\lambda})}\mathrm{e}^{-\frac{3}{2}t^2} \\ &= \sqrt{2\mu(1/2 + e^{1,1}_{\mu,\lambda})}\Big(\frac{2\mu(1/2 + e^{1,1}_{\mu,\lambda})}{2\mu(1/2 + e^{1,1}_{\mu,\lambda}) + 1}\Big)^{\frac{3}{4}} \\ &= {\varsigma^*_{stat}}^{det}\Big(\frac{2\mu(1/2 + e^{1,1}_{\mu,\lambda})}{2\mu(1/2 + e^{1,1}_{\mu,\lambda}) + 1}\Big)^{\frac{3}{4}} \end{aligned} \tag{4.107}$$

with

$${\varsigma^*_{stat}}^{det} := \sqrt{2\mu(1/2 + e^{1,1}_{\mu,\lambda})}$$

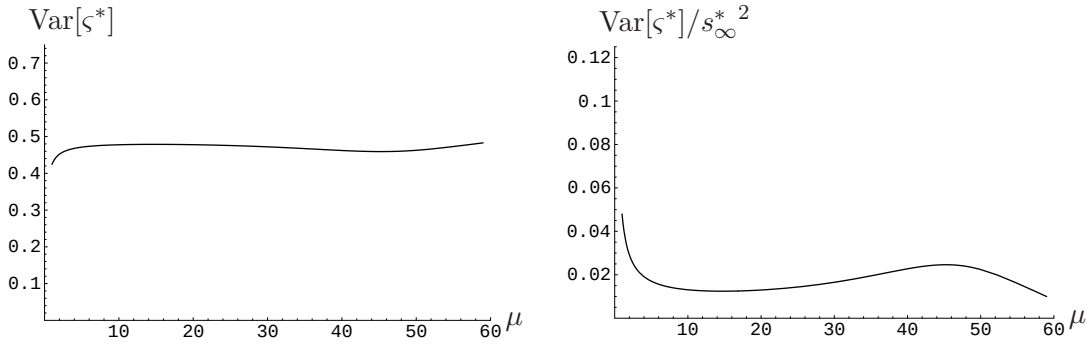
(see (4.11), p. 35). Again, the resulting mutation strength can be given as the product of the result ${\varsigma^*_{stat}}^{det}$ (obtained using the deterministic evolution equations) and a deviation term. The first claim can be verified easily by inserting $N\tau^2 = 1$ into the stationary mutation strength (4.11). The correction factor in (4.107) is a strictly increasing continuous function of the deviation term obtained in (4.98) and therefore the same conclusions apply. Again, recombination with the usual $\mu : \lambda$-ratios reduces the deviation from the deterministic result. Similarly to (4.99), the progress rate for $N\tau^2 = 1$ can be obtained as

$$\begin{aligned} \varphi^*_\infty &= c_{\mu/\mu,\lambda} s^*_\infty - \frac{{s^*_\infty}^2 \mathrm{e}^{t^2}}{2\mu} \\ &= \varphi^*\Big(\sqrt{2\mu(1/2 + e^{1,1}_{\mu,\lambda})}\Big)\Big(\frac{2\mu(1/2 + e^{1,1}_{\mu,\lambda})}{2\mu(1/2 + e^{1,1}_{\mu,\lambda}) + 1}\Big)^{\frac{3}{4}}. \end{aligned} \tag{4.108}$$

Let us now address the variance. For $N\tau^2 = 1$, the variance reads

$$\begin{aligned} \mathrm{Var}[\varsigma^*] &= {\sigma^*_\infty}^2\Big(\mathrm{e}^{t^2} - 1\Big) \\ &= 2\mu(1/2 + e^{1,1}_{\mu,\lambda})\Big(\frac{2\mu(1/2 + e^{1,1}_{\mu,\lambda})}{2\mu(1/2 + e^{1,1}_{\mu,\lambda}) + 1}\Big)^{\frac{3}{2}}\Big(\sqrt{\frac{2\mu(1/2 + e^{1,1}_{\mu,\lambda}) + 1}{2\mu(1/2 + e^{1,1}_{\mu,\lambda})}} - 1\Big). \end{aligned} \tag{4.109}$$

Figure 4.18 shows the variance (4.109) as a function of μ. As 4.18 a) indicates, recombination increases the absolute size of the variance. In contrast to $N\tau^2 \to \infty$, the dependence of the absolute size of the variance is relatively weak. Figure 4.18 a) indicates two local minima of the variance. One for the single point strategy, the other in the region of $\mu \approx 45$. If the relative variance is considered, the situation changes. Figure 4.18 b) reveals the same region of minimal relative variances as found for $N\tau^2 \to \infty$ which is not surprising regarding the similarity of both functions. Disregarding the case of $\mu \approx \lambda$, nearly optimal combinations of μ and λ can be found again in an interval of approximately $\mu = 0.25\lambda$ to $\mu = 0.35$.



a) Variance b) Relative variance w.r.t. ${s^*_\infty}^2$

Figure 4.18: The variance of ς^* in the stationary state (4.109) as a function of μ for $N\tau^2 = 1$. Figure b) depicts the variance w.r.t. ${s^*_\infty}^2$, (4.107).

A Normal Distribution in the Stationary State

As it can be seen in Fig. 4.15, deviations between predicted and measured values exist. This concerns the higher parental numbers $\mu = 20$ and $\mu = 30$. Here, the experimental values for small τ values are smaller than those calculated using (4.95). In the case of the smaller parental number $\mu = 10$ there is a better agreement between experiment and ansatz. As was pointed out in [23] the assumption of a log-normal distribution might not be valid for smaller learning rates. In an alternative attempt, the normal distribution $\mathcal{N}(m, s^2)$ was used as an alternative to model the distribution of the stationary mutation strength. Let us reconsider Equations (4.80) - (4.82) describing the stationary state

$$0 = \overline{\sigma^*_\infty}\Big(1/2 + e^{1,1}_{\mu,\lambda}\Big) - \overline{\sigma^{*\,2}_\infty} c_{\mu/\mu,\lambda}\Big(1 - \frac{1}{N\tau^2}\Big) - \frac{\overline{\sigma^{*\,3}_\infty}}{2\mu N\tau^2} \tag{4.110}$$

$$0 = \overline{\sigma^{*\,2}_\infty}\Big(1 + 2e^{1,1}_{\mu,\lambda} + \frac{1}{\mu}\Big) - \overline{\sigma^{*\,3}_\infty} 2c_{\mu/\mu,\lambda}\Big(1 - \frac{1}{N\tau^2}\Big) - \frac{\overline{\sigma^{*\,4}_\infty}}{\mu N\tau^2} \tag{4.111}$$

$$\begin{aligned} 0 = {} & \overline{\sigma^{*\,3}_\infty}\Big(\frac{1}{\mu} + \Big(1 + \frac{\tau^2}{\mu}\Big)\Big(\frac{1}{2} + e^{1,1}_{\mu,\lambda}\Big)\Big) - \Big(1 + \frac{\tau^2}{\mu}\Big) c_{\mu/\mu,\lambda} \overline{\sigma^{*\,4}_\infty}\Big(1 - \frac{1}{N\tau^2}\Big) \\ & - \Big(1 + \frac{\tau^2}{\mu}\Big) \frac{\overline{\sigma^{*\,5}_\infty}}{2\mu N\tau^2} \end{aligned} \tag{4.112}$$

Using the normal distribution $\sigma_\infty \sim \mathcal{N}(m, s^2)$ leads to

$$0 = m\Big(1/2 + e^{1,1}_{\mu,\lambda}\Big) - (s^2 + m^2) c_{\mu/\mu,\lambda}\Big(1 - \frac{1}{N\tau^2}\Big) - \frac{m(m^2 + 3s^2)}{2\mu N\tau^2} \tag{4.113}$$

$$\begin{aligned} 0 = {} & (s^2 + m^2)\Big(1 + 2e^{1,1}_{\mu,\lambda} + \frac{1}{\mu}\Big) - m(3s^2 + m^2) 2c_{\mu/\mu,\lambda}\Big(1 - \frac{1}{N\tau^2}\Big) \\ & - \frac{3s^4 + 6s^2m^2 + m^4}{\mu N\tau^2} \end{aligned} \tag{4.114}$$

$$\begin{aligned} 0 = {} & m(m^2 + 3s^2)\Big(\frac{1}{\mu} + \Big(1 + \frac{\tau^2}{\mu}\Big)\Big(\frac{1}{2} + e^{1,1}_{\mu,\lambda}\Big)\Big) \\ & - \Big(1 + \frac{\tau^2}{\mu}\Big) c_{\mu/\mu,\lambda}(3s^4 + 6s^2m^2 + m^4)\Big(1 - \frac{1}{N\tau^2}\Big) \\ & - \Big(1 + \frac{\tau^2}{\mu}\Big) \frac{m(15s^4 + 10s^2m + m^5)}{2\mu N\tau^2}. \end{aligned} \tag{4.115}$$

Again, the solutions are obtained numerically using MATHEMATICA. To this end, the solutions of the first two equations were determined. Interestingly, the results do not differ significantly from those using the log-normal distribution. The complete discussion can be found in Appendix D.3.3. The deviations between experiment and prediction are obviously not due to using a skewed distribution.

4.5 Conclusions

In this chapter, the self-adaptive behavior of ES on the sphere model was analyzed. First, ES using intermediate recombination for the object variables and the mutation strength were considered. Afterwards, self-adaptive ES on the noisy sphere were analyzed. Finally, the analysis was extended to the second-order approach for intermediate ES on the undisturbed sphere. In nearly all cases, the progress measures obtained for $N \to \infty$ were used. Therefore, the predicted and the results of experiments

deviate. It remains a point for future research to include the N-dependent versions of the progress measures into the analyses. The analysis on the sphere was mainly conducted using the deterministic evolution equations. The main drawback of this approach is revealed by considering $(1, \lambda)$-ES on the noisy sphere: This approach cannot predict the irregular behavior of the mutation strength since no perturbation parts are taken into account. The modeling assumption that the perturbation parts can be neglected is violated.

As mentioned, deviations of the predicted stationary mutation strengths from the experiments could be observed in high-dimensional search spaces for some choices of the parent numbers in the noise free case. This only occurs for comparatively small values of the learning rate. While the deviations are not high, they indicate a point for further research. On first sight, three possible explanations come to mind:

- The neglectation of higher-order moments of $[(\varsigma^* - \sigma^*)/\sigma^*]^k$ and higher-order powers of τ^2 in the derivation of the SAR.
- The distribution for the stationary state used in the ansatz followed.
- Using a normal distribution to model the perturbation terms.

The occurrence for small values of the learning rate indicate that the deviation is probably not due to neglecting the higher-order terms of τ in the derivation. A remaining cause may be that the ansatz used is not the best approximation for the stationary state distribution. Therefore, a normal distribution for the stationary state was investigated, but the results obtained could not be distinguished from the results using the log-normal distribution. Finding a better distribution remains one of the tasks for the future. Also, it might be interesting to investigate the effects of using higher order Gram-Charlier/Edgeworth series' to model the the distribution of the perturbation parts. In the following, the main results of this chapter are summarized.

In Section 4.1, a first analysis of the steady state behavior of self-adaptive $(\mu/\mu_I, \lambda)$-ES on the sphere model using the log-normal rule for mutating the mutation strength was presented.

The evolution of an ES can be described by the change of the distance to the optimizer and by the change of the mutation strength. Therefore, the progress rate and the self-adaptation response function had to be determined for the analysis. Both progress measures give the expected one-generation change of the respective parameter (which is a relative change in the case of the mutation strength).

Neglecting the stochastic perturbation parts, equations describing the evolution of the distance to the optimizer and the evolution of the normalized mutation strength were obtained. These equations can be used to characterize the system in the stationary state of the normalized mutation strength. Note, this does not entail a stationarity of the R-evolution. The formulae used are generally asymptotically correct, i.e., they hold for $N \to \infty$. Therefore, the results are only approximate for low-dimensional search spaces.

In experiments, multi-recombinative evolution strategies have been found to show a strong dependency of the stationary progress rate on the learning parameter τ. This sensitivity depends on the parental number μ and is in contrast to the behavior of the single parent $(1, \lambda)$-ES which operates on a nearly optimal level for a wider range of the learning parameter.

An explanation for this behavior can be provided by a closer look at the equations describing the stationary mutation strength and the stationary progress rate. Both are functions of the learning parameter coupled with the search space dimension.

The stationary mutation strength also depends on the maximizer of the progress rate and the ratio between the zero of the self-adaptation response and that maximum point. Similarly, the stationary

progress rate is a function of the maximal progress rate and the same ratio. If the zero of the SAR is relatively close to the maximizer of the progress rate, the stationary progress rate is robust against changes of the optimal learning rate. This is the case if there is only one parent. But while the recombination of the object parameters strongly influences the maximum point of the progress rate, the influence on the zero of the self-adaptation response is more muted (the SAR reacts to the aggregated fitness). Furthermore, increasing μ decreases the zero of the SAR at first. As a result, only if the parental number is close to one or close to the number of offspring, a more robust behavior can be expected. In the latter case, though, the ES tends to divergent behavior.

In addition, there exists an optimal normalized mutation strength and an optimal progress rate for each $(\mu/\mu_I, \lambda)$-ES. Comparing these maximally achievable progress rates, one finds a strong dependency on the relation between the number of offspring λ and the number of parents μ. As it could be shown numerically [23, p. 226], for $N \to \infty$ and λ sufficiently large, a relation μ/λ of approximately 0.27 leads to nearly maximal progress rates. Therefore, evolution strategies that adhere to this principle can exhibit high progress rates, if the mutation strength adaptation process works nearly optimal.

The performance of the ES depends on the learning parameter τ. An optimal τ choice exists even if the zero of the self-adaptation response $\varsigma^*_{\psi_0} = (1/2 + e^{1,1}_{\mu,\lambda})/c_{\mu/\mu,\lambda}$ and the maximum point of the progress rate $\varsigma^*_{\varphi_{opt}} = \mu c_{\mu/\mu,\lambda}$ differ significantly. For $N \gg 1$, τ_{opt} is given by

$$\tau_{opt} = \frac{1}{\sqrt{2N}} \sqrt{\frac{\mu c^2_{\mu/\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda} - 1/2 - e^{1,1}_{\mu,\lambda}}}. \tag{4.116}$$

The optimal learning rate scales with $1/\sqrt{2N}$. If $\mu c_{\mu/\mu,\lambda} \ll (1/2 + e^{1,1}_{\mu,\lambda})/c_{\mu/\mu,\lambda}$, the value of the second square root is close to one. This is, e.g., the case for truncation ratios of approximately 0.27 provided that λ is relatively large. This ratio is the $\mu : \lambda$-ratio recommended on the sphere [7], allowing to use $1/\sqrt{2N}$ and ensuring nearly optimal progress.

Problems arise if the learning parameter is not optimal since this may lead to progress rates that are far smaller than the possible maximum. Of course, this does not mean that the self-adaptation does not work in this case. For a wide range of the learning parameter, the mutation strengths realized will lead towards positive progress – albeit not with maximal possible speed.

Having said that, the question may be raised whether an intermediate recombination of the mutation strength exactly mirroring the recombination of the object variables might not be better replaced by a different method.

Actually, an intermediate recombination of the mutation strengths seems to be unnecessary for the fitness environment considered here. This must be taken with a grain of salt, of course, since only a deterministic approximation of the evolution equations was used and the formulae were derived for $\tau \ll 1$ or $\tau = 0$, respectively, and $N \to \infty$. Nevertheless, switching off the recombination totally and just taking the mutation strength of the best offspring is not expected to lead to a deterioration of the performance in the non-noisy case. The optimal mutation strength remains reachable, since the zero of the progress rate is still approached for $N\tau^2 \to \infty$. In addition, the zero of the SAR as the limit of the stationary progress rate for $N\tau^2 \to \infty$ is at least higher as it would be if recombination were used. The improvement might not be really significant but it indicates that for undisturbed sphere functions there appears to be no detectable positive effects stemming from the intermediate recombination of the mutation strengths. This of course, might not hold and is not expected to hold in the case of different fitness functions.

In Section 4.2, the self-adaptation of $(1, \lambda)$-ES on the noisy sphere model was investigated. To this end, the evolution of the ES over time was described by the evolution equations. First of all, the

progress measures, the self-adaptation response and the progress rate had to be obtained. Afterwards, a deterministic approach was applied, i.e., the stochastic parts of the evolution equations were neglected.

In the case of a constant noise strength σ_ϵ, three different phases of the evolution have been identified. As long as the system is still far away from the optimum, the influence of the noise can be neglected. As a result, the ES reaches a similar stationary mutation strength as in the noise free case and the same recommendations for choosing the learning parameter apply.

Approaching the optimum, however, changes the situation. Due to the increasing normalized noise, the steady state of the mutation strength is lost. The progress decreases until the ES cannot get any closer to the optimizer on average. The progress rate becomes zero. This can be used to determine the residual location error. There are two estimates that can be obtained. The first is associated with a vanishing mutation strength, the other demands stationarity of the mutation strength evolution as well – requiring the SAR to be zero. Interestingly, both estimates are very similar especially if large offspring population sizes are considered.

A remarkable observation is that the $(1, \lambda)$-ES is not able to stabilize the mutation strength although the deterministic approach predicts a locally stable non-zero mutation strength. Instead its behavior resembles a random walk where the mutation strength fluctuates between the non-zero mutation strength (4.51) and zero. A general preference of small values can be observed. Since any mutation strength between these two extremes leads nearly to the same residual location error, the estimates that were obtained lead to good predictions.

The reason for the behavior of $(1, \lambda)$-ES cannot be explained by considering the deterministic approximation. Comparing the behavior of $(1, \lambda)$-ES with that of intermediate $(\mu/\mu_I, \lambda)$-ES, one finds that the latter show a second stationary phase of the mutation strength once the system has reached the vicinity of the residual localization error. The difference in the behavior is clearly due to the missing recombination of the mutation strength. If the normalized mutation strength is considerably smaller than the normalized noise strength, the ES is virtually unable to choose the offspring on basis of the actual fitness values. Instead – concerning the mutation strength – the selection is similar to a random sampling of log-normally distributed variables.

Using intermediate recombination introduces a probabilistic preference towards an increase of the mutation strength whereas an $(1, \lambda)$-ES de- and increases the mutation strength with the same probability. Thus, $(\mu/\mu_I, \lambda)$-ES will tend to increase a small mutation strength until it is sufficiently large so that the information obtained by the fitness function is taken into account. As far as the constant noise scenario is considered, this “bias” can be regarded as a desirable property of intermediate recombination.

The $(1, \lambda)$-ES on the sphere model has a slight bias towards a decrease of the mutation strength. This explains the wandering behavior of the mutation strength. Introducing a slight counteracting bias in the σ mutation operator remedies the loss of step-size control to a certain extent.

While first insights into the mechanism of self-adaptation of ES on the noisy sphere were provided, the investigations are far from being complete. First, the considerations did not take into account the stochasticity of the evolutionary process explicitly. Especially in the large noise regime, the deterministic approximation leads to predictions which are not fully consonant with the observed dynamics. Therefore, incorporating fluctuations and solving the corresponding Chapman-Kolmogorov-Equations remains as a task for the future.

In Section 4.3, the behavior of intermediate $(\mu/\mu_I, \lambda)$-evolution strategies on the noisy sphere was investigated. To this end, the deterministic evolution equations were applied. As seen in Section 4.2, $(\mu/\mu_I, \lambda)$-ES have a slight preference for an increase of the mutation strength which is due to the intermediate recombination of the mutation strength. This bias leads to the existence of a stationary state in the case of uniform noise on the sphere which can be described using the deterministic variant

of evolution equations.

Let us sum up our findings: Intermediate recombination of the object variables and the mutation strength introduces a strong dependency on the learning rate τ during the first phase of the optimization process. Here, the ES can be assumed to be far away from the optimum and the influence of the noise can be neglected. While the sensitivity with respect to the learning rate is a drawback in comparison with the robustness of $(1, \lambda)$-ES, recombination of the object variables enables higher progress rates and a faster convergence. The learning rate can be chosen appropriately, so that the ES adapts an optimal normalized mutation strength.

In the last phase, noise overshadows the information of the fitness function. In this case, recombination is the cause of two effects: Recombination of the object variables allows smaller residual location errors, whereas recombination of the mutation strengths leads to a sufficiently stable stationary mutation strength in contrast to $(1, \lambda)$-ES.

The usual recommendation of choosing $\mu : \lambda \approx 0.27$ still applies – regardless whether self-adaptation in the noise-free case or in the case of permanent noise is considered. While this ratio results in a high sensitivity towards the size of τ, the achievable progress is optimal. Additionally in the noise scenario, this truncation ratio leads to a nearly optimal location error. Interestingly, the predicted residual location error does not deviate far from a hypothetical minimal value obtained for a zero mutation strength. Recombination improves the deviation even more.

The analysis presented here is not complete. In Section 4.3, the effects of additive uniform noise were investigated. Other noise models remain to be considered – for instance actuator noise where the noise is not added to the fitness function but to the coordinates of the object vector. Furthermore, the effects of non Gaussian noise distributions would be interesting.

The progress rate and the SAR used were obtained for $N \to \infty$. In order to capture the evolution more exactly, the N-dependent variants will have to be applied. Also, an inclusion of the perturbation parts in the evolution equations and an extension of the analysis similar to [23, p. 309] still remain. For the undisturbed sphere, Section (4.4) presented a first analysis.

In Section 4.4, the fluctuation parts were included in the analysis – approximating the unknown distribution with a normal distribution. To proceed, the variances had to be obtained. In the case of the R-evolution, the variance equals zero in the present analysis framework. Deviations from the deterministic approach only stem from the σ^*-evolution.

The task of obtaining the mean value dynamics leads to recursive equations in which the raw lower order moments depend on higher-order moments. Therefore, an ansatz has to be used setting the distribution of the stationary mutation strength equal to a reference distribution. This was done for two distributions: the log-normal distribution and a normal distribution. Concerning the stationary mutation strength, i.e., the expectation, both distributions lead to nearly the same results.

Similarly to Section 4.1, experimental results for some $(\mu/\mu_I, 60)$-ES were obtained. In contrast to Section 4.1, however, no closed general formulas could be provided. The solutions must be obtained numerically. Evaluating the stationary values as functions of the learning rate underlines the findings of Section 4.1. Again, an optimal learning rate is clearly defined. Furthermore, ES with $\mu : \lambda$-ratios close to the recommendation of 0.27 lead to the largest progress for τ-values in the vicinity of the optimal learning rate.

As said, the solutions evade an analytical treatment in general. Further analyses, therefore, were restricted to specific choices of the parameters – for example either the parent number μ or the learning rate τ.

The remainder of the subsection was concerned with the effects of recombination. For some specific values of τ, recombination with the usual $\mu : \lambda$-ratio was shown to lead approximately to the smallest deviations from the deterministic prediction and to the smallest relative variances. The

performance loss due to random fluctuations thus is nearly minimal for those ratios.

Deviations between experiments and predicted values were observed for low-dimensional search spaces. As a rule, the prediction quality improves with increasing dimensionality. Some further relatively small deviations can be observed: For small values of the learning rate, predictions and experiments deviate for ES with larger number of parents. Finding the exact cause of these deviations remains a task for future work. The same holds for an inclusion of the N-dependency of the equations in order to give more accurate predictions for low-dimensional search spaces.

5 Self-Adaptation on Ridge Functions

So far the focus was on the sphere model $f(\mathbf{y}) = g(\|\mathbf{y} - \hat{\mathbf{y}}\|)$ which depends on one parameter only: the distance to the optimizer. In this section, ridge functions are considered. They can be seen as an extension of sphere functions since they contain a linear gain part and a negative sphere-like component. General ridge functions are defined in the following way.

Definition 5. The general ridge function with axis direction $\mathbf{v}$ and parameters α and d determining the shape of the ridge is given by

$$F_{gR}(\mathbf{y}) \quad := \quad \mathbf{v}^{\mathrm{T}}\mathbf{y} - d\Big(\sqrt{(\mathbf{v}^{\mathrm{T}}\mathbf{y}\mathbf{v} - \mathbf{y})^{\mathrm{T}}(\mathbf{v}^{\mathrm{T}}\mathbf{y}\mathbf{v} - \mathbf{y})}\Big)^{\alpha} \tag{5.1}$$

with $d > 0$ and $\alpha > 0$. The vector $\mathbf{v} \in \mathbb{R}^N$ with $\|\mathbf{v}\| = 1$ is called the ridge direction. □

In this chapter a rotated version of the general ridge function is considered. In the case of the rotated ridge the ridge axis is aligned with the coordinate axis y_1 [22].

Definition 6. The rotated ridge function aligned with the coordinate axes has the form

$$F_R(\mathbf{y}) \quad = \quad y_1 - d\Big(\sum_{i=2}^{N} y_i^2\Big)^{\alpha/2}. \tag{5.2}$$

□

The parameter α determines the degree of the ridge function and the general topology of the fitness landscape. A ridge function with $\alpha = 1$ is called a sharp ridge (see. Fig. 5.1). The parameter d determines the angle by which the isofitness lines intersect with the ridge axis and therefore the "sharpness" of the function. A ridge function with $\alpha = 2$ is called a parabolic ridge (see. Fig. 5.2). Again, d determines the form of the isofitness lines. In general, if $d \to 0$, the problem degenerates to the hyperplane $F(\mathbf{y}) = y_1$, whereas for increasing d the isofitness lines appear as more and more parallel to the axis and the problem approaches a sphere model with $F(\mathbf{y}) = -d(\sum_{i=2}^{N} y_i^2)^{\alpha/2}$. The $N - 1$ terms which make up the sphere component of the ridge can be interpreted as a $(N - 1)$ dimensional distance to the axis y_1. To simplify the notation,

$$\begin{aligned} F_R(\mathbf{y}) \quad &= \quad y_1 - d\Big(\sum_{i=2}^{N} y_i^2\Big)^{\alpha/2} \\ \Rightarrow f(x, R) \quad &:= \quad x - dR^{\alpha} \end{aligned} \tag{5.3}$$

is used for the remainder of this chapter.

Ridge functions do not have a finite optimum and therefore may be considered an "ill-posed" problem for ES [9]: Since the "optimum" lies in infinity, the fitness of the ES must be steadily increased. Improvement is possible in many ways. Generally, there are two viewpoints that may be

taken [22]. First Oyman's viewpoint is taken into account [79, p.32]. He says that the "object variable for the optimum [...] reads

$$\hat{x_1} \to \infty,\ \forall i \neq 1 : \hat{x_i} = 0."$$

Note, Oyman uses $\mathbf{x}$ instead of $\mathbf{y}$ to denote the object vector. This viewpoint derives its justification from seeing ridge functions as the limit of

$$F_c(\mathbf{y}) \quad = \quad y_1 - cy_1^2 - d\Big(\sum_{i=2}^{N} y_i^2\Big)^{\alpha/2} \tag{5.4}$$

for $c \to 0$ (cf. [79, 22]). For every finite c, F_c has an optimal point at $(1/(2c), 0, \ldots, 0)^{\mathrm{T}}$. If c decreases, the position on the axis moves towards infinity.

Evolution strategies use local information. They sample the search space randomly and select the μ best offspring, i.e., the μ highest fitness values they have found. This is the foundation of the second viewpoint wich takes a more process oriented view. The ridge does not have a finite optimum. The algorithm is required to increase the fitness perpetually. This does not necessarily mean that is has to find the ridge. Although the highest fitness value is on the axis for every finite interval, the situation changes if an unbounded search space is considered. Actually, it is not even necessary to require a finite distance to the ridge. Since the search space is infinite, there are infinitely many points in arbitrary distance to the ridge for each position on the axis with exactly the same fitness. As result, the ES may diverge from the axis – as long as it increases the linear component faster than the loss components. In addition, this does not mean that the progress is slower as a rule since moving away from the axis may allow for higher step lengths.

As it will be shown, evolution strategies may actually exhibit both behaviors: Trying to converge to the axis or diverging from it – enlarging the axis-component faster than the loss components.

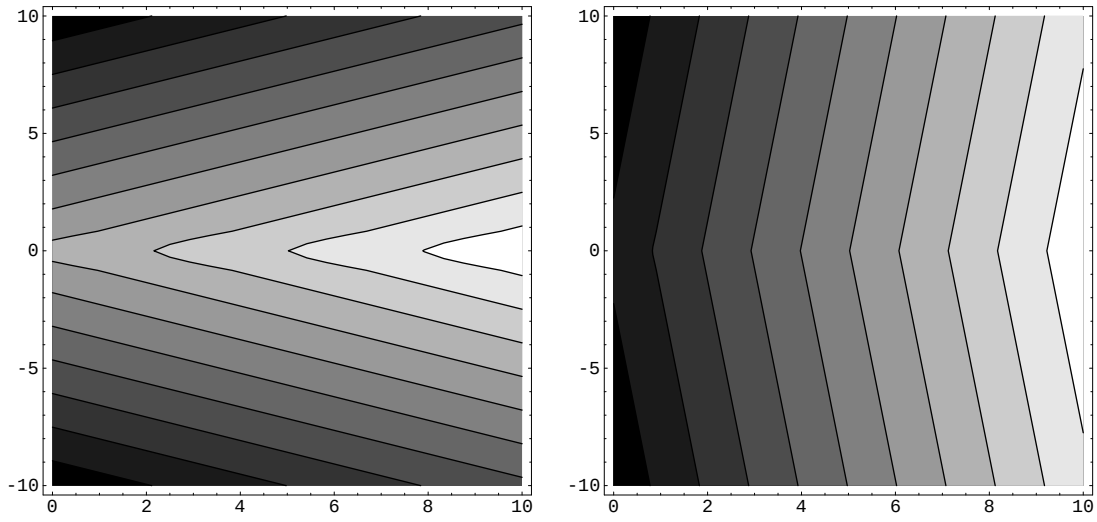
This chapter is organized as follows: First, self-adaptation on sharp ridge functions is considered. Afterwards, the parabolic ridge serves as an example for self-adaptation on ridge functions of higher degree. Finally, the case of ridge functions disturbed by noise is addressed.

5.1 Self-Adaptation in the Noise-free Case

As mentioned, this section is devoted to an analysis of the self-adaptation behavior of evolution strategies on undisturbed ridge functions. Again, the analysis makes use of the evolution equations introduced in Chapter 3. Two ridge functions serve as representatives of the function class: the sharp and the parabolic ridge.

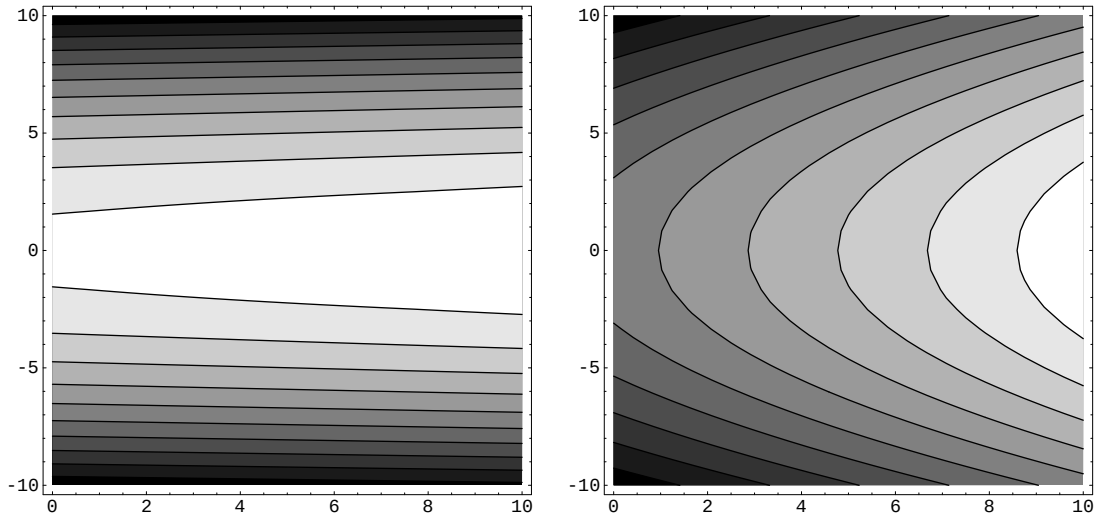
5.1.1 The Sharp Ridge: Convergence or Divergence

The sharp ridge is characterized by $\alpha = 1$ and $F(\mathbf{y}) = y_1 - d(\sum_{i=2}^{N} y_i)^{1/2}$ or $f(x, R) := x - dR$. It has been reported [57] that self-adaptive ES fail on the sharp in some cases by reducing the mutation strength so far that no significant progress is observable anymore. Since the "optimum" of the ridge lies in infinity, the ES can be said to converge prematurely. This behavior is not restricted to self-adaptation. Other adaptation schemes are also known to reduce the mutation strength prematurely – unless modifications are introduced. In the case of CSA-ES, it was found [19] that the behavior is determined by the choice of the ridge parameter d: Depending on the size of d (i.e., $d < 1$, $d > 1$), either a convergence towards the axis or a divergence $R \to \infty$ occurs. It will be shown that in the case of self-adaptation, d appears again as the decisive parameter and furthermore that the critical value of d depends on the population parameters μ and λ.



a) $d = 2$ b) $d = 0.1$

Figure 5.1: Contour Plots of the sharp ridge for $d = 2$, $d = 0.1$, and $N = 2$. The ridge axis aligns with the x-axis. Brighter grey tones indicate better fitness values.



a) $d = 2$ b) $d = 0.1$

Figure 5.2: Contour Plots of the parabolic ridge for $d = 2$, $d = 0.1$, and $N = 2$. The ridge axis aligns with the x-axis. Brighter grey tones indicate better fitness values.

The Evolution Equations

The behavior of self-adaptive ES on ridge function can be characterized by three variables: The position with respect to the axis x, the distance to the axis R, and the mutation strength $\langle \varsigma^{(g)} \rangle$. As

before in Chapter 3, the deterministic evolution equations are used in the analysis. Let $x^{(g)} := \langle x^{(g)} \rangle$ denote the x-component of the centroid of the population at generation g. Similarly $R := R^{(g)}$ denotes the distance of the centroid to the axis, whereas r is a short form for $r := R^{(g+1)}$. The parameter $\sigma^\star := N\langle \varsigma^{(g)} \rangle$ stands for the mean of the mutation strengths in generation g – normalized with respect to the search space dimensionality. Similarly, $\varsigma^\star := N\langle \varsigma^{(g+1)} \rangle$ denotes the mean in generation $g+1$ unless the dependence on the generation number shall be emphasized. As before high-dimensional search spaces are considered. This allows to identify $N-1$ with N. Accordingly, the normalized evolution equations read

$$\begin{aligned} x^{(g+1)} &= x^{(g)} + \frac{1}{N}\varphi_x^\star(\sigma^\star) \\ r &= R - \frac{1}{N}\varphi_R^\star(\sigma^\star, R) \\ \langle \varsigma^{\star(g+1)} \rangle &= \sigma^\star\Big(1 + \psi(\sigma^\star, R)\Big). \end{aligned} \tag{5.5}$$

The progress rate $\varphi_R^\star$ and SAR ψ are obtained in Appendices B.2 and C.1.2 (or C.1.3, respectively) as

$$\varphi_R^\star(\sigma^\star, R) = \frac{dc_{\mu/\mu,\lambda}}{\sqrt{1+d^2}}\sigma^\star - \frac{\sigma^{\star 2}}{2R\mu} \tag{5.6}$$

for $\tau = 0$ and $N \to \infty$ and

$$\psi(\sigma^\star) = \tau^2\left(\frac{1}{2} + e_{\mu,\lambda}^{1,1} - \frac{c_{\mu/\mu,\lambda}}{R}\sqrt{\frac{d^2}{1+d^2}}\,\sigma^\star\right) \tag{5.7}$$

for $N \to \infty$ and $\tau \ll 1$. Both performance measures are influenced by the ridge parameter d over the sine of the slope angle of the gradient vector

$$\nabla f_R(x, R) = \begin{pmatrix} 1 \\ -d \end{pmatrix} \tag{5.8}$$

with respect to the x-axis. The larger the d-value, the steeper the slope and more and more weight is put on the linear components in (5.6) and (5.7): For $d \to \infty$, both performance measures converge to their sphere model equivalent. For $d \to 0$, the optimization of the ridge is transformed into optimizing the linear function in x: Expected progress towards the axis does not occur anymore and the SAR is strictly positive.

The progress rate $\varphi_x^\star$ measuring the progress on or parallel to the axis is given by

$$\varphi_x^\star(\sigma^\star) = \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}}\sigma^\star \tag{5.9}$$

(cf. Appendix B.2) and is obtained under the same conditions as (5.6).

As the SAR (5.7) and the progress rate (5.6), (5.9) is influenced by the ridge constant d. This time, though, it is the cosine of the gradient angle that exerts its weight.

As Eqs. (5.5) - (5.7) and (5.9) show, there is no feedback of the evolution of $x^{(g)}$ on those of the other state variables whereas the change of $x^{(g)}$ is governed by the mutation strength. As consequence, the analysis is continued with considering the system in $(R^{(g+1)}, \langle \varsigma^{\star(g+1)} \rangle)^{\mathrm{T}}$

$$\begin{pmatrix} R^{(g+1)} \\ \langle \varsigma^{\star(g+1)} \rangle \end{pmatrix} = \begin{pmatrix} R - \varphi_R^\star(R, \sigma^\star)/N \\ \sigma^\star\Big(1 + \psi(R, \sigma^\star)\Big) \end{pmatrix}. \tag{5.10}$$

First of all, it should be noted that (5.10) with (5.6) and (5.7) permits negative values in contrast to described process itself. Therefore, first the zero points of the evolution equations are obtained. As can be seen, in the case of the R-evolution, the variable $R^{(g+1)}$ might be negative if

$$\begin{aligned} 0 &\geq R - \frac{1}{N}\varphi_R^\star(\sigma^\star) \\ \Rightarrow 0 &\geq R - \frac{c_{\mu/\mu,\lambda} d}{N\sqrt{1+d^2}}\sigma^\star + \frac{\sigma^{\star 2}}{2\mu R N} \end{aligned} \tag{5.11}$$

leading to the zero points

$$\sigma_{1,2}^\star = R\Big(\frac{d}{\sqrt{1+d^2}}\mu c_{\mu/\mu,\lambda} \pm \sqrt{\mu^2 c_{\mu/\mu,\lambda}^2 \frac{d^2}{1+d^2} - 2\mu N}\Big) \tag{5.12}$$

which are not defined in $\mathbb{R}$ if $N > (1/2)\mu c_{\mu/\mu,\lambda}^2 (d^2/(1+d^2))$. If the search space dimensionality is sufficiently large, the deterministic evolution equation only admits positive results. In the case of the SAR,

$$\begin{aligned} 0 &< \sigma^\star\Big(1+\psi(\sigma^\star)\Big) \\ \Rightarrow 0 &< \frac{1}{\tau^2} + \frac{1}{2} + e_{\mu/\lambda}^{1,1} - \frac{c_{\mu/\mu,\lambda}}{R}\frac{d}{\sqrt{1+d^2}}\sigma^\star \\ \Rightarrow \sigma^\star &< R\Big(\frac{\sqrt{1+d^2}}{d c_{\mu/\mu,\lambda}\tau^2} + \Big(\frac{\frac{1}{2}+e_{\mu/\lambda}^{1,1}}{c_{\mu/\mu,\lambda}}\Big)\frac{\sqrt{1+d^2}}{d}\Big) \end{aligned}$$

must hold for positive $\langle \varsigma^{\star(g+1)} \rangle$. As it can be seen, the relation between mutation strength and distance is decisive. The mutation strength must exceed the zero point of the SAR. And furthermore, it has to be considerably greater than R/τ^2. Choosing τ sufficiently small, increases the admissible region. The SAR (5.7) decreases linear with the mutation strength, though. Too large mutation strengths result in a negative answer of the evolution equation. As it is shown later on, this does not occur, actually.

Considering the deterministic difference equation system (5.10), the first question to be addressed is whether the system comes to a halt; in other words, whether stationary points exist.

Stationary Points

Stationary points are characterized by $\langle \varsigma^{\star(g+1)} \rangle = \langle \varsigma^{\star(g)} \rangle$ and $R^{(g+1)} = R^{(g)}$. Considering (5.10), the progress rate (5.6), and the SAR (5.7), a stationary state requires either a zero mutation strength or that the zero of $\varphi_R^\star$, (5.6),

$$\varsigma_{\varphi R0}^\star = 2\mu c_{\mu/\mu,\lambda}\frac{d}{\sqrt{1+d^2}}\,R \tag{5.13}$$

and the zero of ψ, (5.7),

$$\varsigma_{\psi_0}^\star = \Big(\frac{1/2 + e_{\mu,\lambda}^{1,1}}{c_{\mu/\mu,\lambda}}\Big)\frac{\sqrt{1+d^2}}{d}\,R \tag{5.14}$$

are equal. Note, both are linear functions in R. As a result, they do not intersect in general for positive distances. Only in one singular case, there are stationary points of (5.10) with a positive mutation strength: A stationary state with a non-zero mutation strength of system (5.10) exists if and only if

$$d = d_{crit} \quad = \quad \sqrt{\frac{2e^{1,1}_{\mu,\lambda} + 1}{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1}} \tag{5.15}$$

holds (see (5.13) and (5.14)). Otherwise, there is no stationary point except $\sigma^\star = 0$. In the situation of (5.15), (5.13) and (5.14) overlap as functions of R – creating a single linear function in R. For every R there is a mutation strength for which the whole system comes to a halt. As result, neither a stationary distance nor a mutation strength can be determined. The expected changes indicate that the stationary state line serves as an attractor. But where the system comes to rest depends on the position in the search space. Furthermore, the ES is subject to random perturbations which the deterministic equations neglect. Due to perturbations, the stationary state will be left. The system is expected to return to the line but to a different position than before. As a result, for d-choices close to the critical value, a meandering behavior of the ES is expected.

The parameter d_{crit} depends on the population parameters μ and λ and is largest (i.e., close to one for most choices of λ) for $\mu = 1$ or μ close to λ. The λ-dependence of d_{crit} is relatively weak which shall be illustrated exemplarily for a $(1, \lambda)$-ES. In the case of extremely small offspring population sizes, i.e., $\lambda < 3$, the critical d-value is greater than 1, going down to ≈ 0.936 around $\lambda \approx 12$ before approaching 1 again for $\lambda \to \infty$. The latter approach is extremely slow, though.

The dependence on the size of the parent population is more pronounced. Switching from $\mu = 1$ to $\mu = 2$ lowers the critical d-value about $\approx 40\%$. This trend translates to the usual $\mu : \lambda$-ratios: Compared to $\mu = 1$, recombination decreases the critical d-value as Fig. 5.3 illustrates for the case of $(\mu/\mu_I, 10)$-ES.

The Influence of d

The d_{crit}-value (5.15) is a critical point for system (5.10): For all choices $d \neq d_{crit}$, the deterministic system (5.10) does not come to a halt for strictly positive choices of the mutation strength. As already observed in various experiments, there are two opposite behaviors of the ES. Either it converges prematurely – approaching the axis and reducing the mutation strength in the process or it diverges from the axis – increasing $\sigma^\star$ and R. The size of the parameter d determines which behavior occurs: For $d > d_{crit}$ (5.15), the variables R and $\varsigma^\star$ are expected to decrease, whereas for $d < d_{crit}$ they are expected to increase.

Figure 5.3 illustrates the behavior of evolution strategies for several choices of μ. For $\mu = 1$ the results diverge with the exception of $d = 0.9$ (critical d-value 0.936). In the case of $\mu = 3$ with a critical d-value of 0.418, all runs with $d \leq 0.5$ diverge whereas for $\mu = 5$ only the runs for $d = 0.2$ diverge. The critical d-value in this case is 0.318.

The causes for these behaviors are investigated in the following. Let us start with Fig. 5.4 which shows the isoclines $\varphi^\star_R = 0$, (5.13), and $\psi = 0$, (5.14). Both are linear functions in R and influenced by the sine of the gradient's slope angle

$$\nabla f_R(x, R) \quad = \quad \begin{pmatrix} 1 \\ -d \end{pmatrix}.$$

But the influence of d on the zero of the progress rate is reciprocal to its influence on the SAR. Increasing d decreases the zero of the SAR, but increases the zero of the progress rate. Both values approach their sphere model equivalent and the influence of the linear part of the ridge is lessened. On the other hand, decreasing d lowers the zero of the progress rate since more and more weight is put

on the linear component of the ridge. The zero of the SAR increases in turn until the SAR is finally strictly positive which is required in the optimization of linear functions.

Self-Adaptation sees the fitness as a whole and thus the compromise of the linear and the sphere component. It does not generally focus on a positive lateral progress rate. It is shown later on that concerning d, the zero of the SAR behaves in the same manner as the optimizer of the quality change (expected change of the fitness). Concerning the zeros of the progress rate and the zero of the SAR, the different dependence on d can cause the zero points or the isoclines, respectively, to switch roles: For $d < d_{crit}$, (5.14) is greater than (5.13). For $d > d_{crit}$, the zero of the progress rate (5.13) dominates the zero of the SAR (5.14) as in the sphere model case.

Consider now Fig. 5.4. On the one hand, if the system (5.10) is on the line $\psi = 0$, (5.14), the evolution of the mutation strength comes to a halt. A change can only occur because of the ongoing evolution of R. On the other hand, on the line $\varphi^\star_{R0}$, there is no change in R and the system only moves due to a change in the mutation strength $\sigma^\star$.

Considering the SAR, remember that for mutation strengths smaller than the zero of the SAR, an increase is expected whereas for mutation strengths greater then the zero an expected decrease occurs. Translating that for Fig. 5.4, the area below $\psi = 0$ is characterized by a positive SAR and an expected increase of the mutation strength which is indicated by the upward arrow. The area above $\psi = 0$ is characterized instead by $(\sigma^\star, R)^{\mathrm{T}}$-combinations for which the SAR is negative and thus a decrease of the mutation strength is expected. This is indicated by the downward pointing arrow.

Similarly in the case of the progress rate, the area below $\varphi^\star_R = 0$ is characterized by $(\sigma^\star, R)$ combinations for which the progress rate is positive. Because of the definition of the progress rate $\varphi^\star_R = N\mathrm{E}[R - r]$, positive progress is connected with a decrease of the distance. Therefore, below $\varphi^\star_R = 0$ a decrease of the distance is expected (which is indicated by the left pointing arrows in Fig. 5.4). Finally, once $(\sigma^\star, R)^{\mathrm{T}}$ is above the line $\varphi^\star_R = 0$, an increase of the distance to the ridge is expected – indicated by the right pointing arrows. The figure of the isoclines can be used to illustrate the key features of the behavior of the system rather easily. First of all, recall that the choice of d decides which isocline dominates the other. If $d < d_{crit}$, the plot of $\psi = 0$, (5.14), lies above that of $\varphi^\star_R = 0$, (5.13). For $d > d_{crit}$, the opposite situation occurs. This results in different movements in the area between the two isoclines – the area system (5.10) will eventually move into as Fig. 5.4 shows:

Regardless of whether $d < d_{crit}$ or $d > d_{crit}$, the deterministic system $(\sigma^\star, R)^{\mathrm{T}}$ leaves region I_1 and I_2 via I_3 for $g \to \infty$. Region I_3 cannot be left again. If $d > d_{crit}$, system (5.10) moves towards the origin – decreasing $\varsigma^\star$ and R. If $d < d_{crit}$, system (5.10) moves towards infinity – increasing $\varsigma^\star$ and R.

Let us illustrate that by example for Fig. 5.4 a). Here, the isocline $\varphi^\star_R = 0$, (5.13), is above the isocline $\psi = 0$, (5.14). This equals the condition $d > d_{crit}$, (5.15). If the system (5.10) starts in the area below $\psi = 0$, the SAR and the progress rate are positive. As a result, the mutation strength increases and the distance decreases. The system moves towards the line $\psi = 0$. Once this is reached, the $\varsigma^\star$-evolution temporarily stops. But since the R-evolution still progresses and the distance decreases, the isocline $\psi = 0$ is crossed and the system enters the area between both isoclines. There it remains and approaches zero. Therefore, for $\varsigma^\star_{\varphi_{R0}}$, (5.13), $> \varsigma^\star_{\psi_0}$, (5.14), i.e., for $d > d_{crit}$, the system in R and $\varsigma^\star$ approaches the origin with $R \to 0$, $\varsigma^\star \to 0$ as in the case of the sphere.

The opposite behavior appears for $d < d_{crit}$, (5.15) (see Fig. 5.4 b)) and $\varsigma^\star_{\varphi_{R0}}$ (5.13) $< \varsigma^\star_{\psi_0}$ (5.14). Again, the system reaches the cone defined by $\varphi^\star_R = 0$, (5.13), and $\psi = 0$, (5.14), and cannot leave it again. But once it is inside, due to the expected increases of the mutation strength and the distance it moves into the opposite direction – going to infinity.

What does the behavior of (5.10) mean for the ES? The size of the parameter d with respect to

d_{crit}, (5.15), decides whether a premature convergence occurs. The critical size of d depends on the choice of the population parameters μ and λ. Introducing recombination, i.e., $\mu > 1$, lowers the critical d-value for $\mu \not\approx \lambda$. That is, a premature reduction of the mutation strength can be modified to some extend by using non-recombinative strategies. Summarizing, the characteristics of self-adaptive ES on the sharp ridge are the following:

1. There is no feedback of the $x^{(g)}$-evolution on the evolutions of $\langle \varsigma^{\star(g)} \rangle$ and $R^{(g)}$.
2. The evolutions of $\langle \varsigma^{\star(g)} \rangle$ and $R^{(g)}$ are coupled.
3. Because of this, the evolution of the mutation strength is kept between the zero of the progress rate $\varphi^{\star}_R$ and the SAR ψ.
4. Both variables are influenced by the constant gradient of the ridge and thus by the ridge parameter d.
5. Concerning d, the zero of the SAR follows the optimizer of the quality change – a behavior not shown by the zero of the progress rate as it is shown in the next section.
6. The size of d with respect to μ and λ decides whether the ES operates with mutation strengths that lead to a positive or negative progress rate.

The first situation connected with positive progress towards the axis results in a premature convergence whereas the latter causes the ES to show in a way the behavior required: The fitness is on average increased and increased as the next section illustrates.

Divergence: The Influence of Recombination

If the ridge parameter d is sufficiently small with respect to λ and μ, a self-adaptive ES does not converge prematurely but increases the distance to the ridge and the mutation strength. The first question that arises, though, is whether the ES has a positive quality change. If this is true it would be interesting to know whether the ES is able to travel with nearly optimal speed.

Potentially Too Small Mutation Strengths So, let $d < d_{crit}$, (5.15), and consider the expected change of the fitness from one generation to the next. This performance measure

$$\overline{\Delta Q} \quad := \quad \mathrm{E}[F(\langle \mathbf{y}^{(g+1)} \rangle) - F(\langle \mathbf{y}^{(g)} \rangle)] \tag{5.16}$$

is called the *quality change*. Using the same normalization as before, i.e., setting $\overline{\Delta Q^{\star}} = N\overline{\Delta Q}$, it can be easily given as $\overline{\Delta Q^{\star}} = \varphi^{\star}_x + d\varphi^{\star}_R$ since the sharp ridge is considered. Using the progress rates (5.6) and (5.9), the quality change reads

$$\overline{\Delta Q^{\star}} \quad = \quad \sqrt{1+d^2} c_{\mu/\mu,\lambda} \varsigma^{\star} - \frac{d}{2R\mu} \varsigma^{\star 2}. \tag{5.17}$$

Its optimizer is given by

$$\varsigma^{\star}_{opt} \quad = \quad c_{\mu/\mu,\lambda} R\mu \frac{\sqrt{1+d^2}}{d} \tag{5.18}$$

and scales with the distance to the axis. In addition, the quality change is positive for mutation strengths in the interval $]0, 2R\mu c_{\mu/\mu,\lambda}\sqrt{1+d^2}/d[$. So, first of all as long as self-adaptation leads to

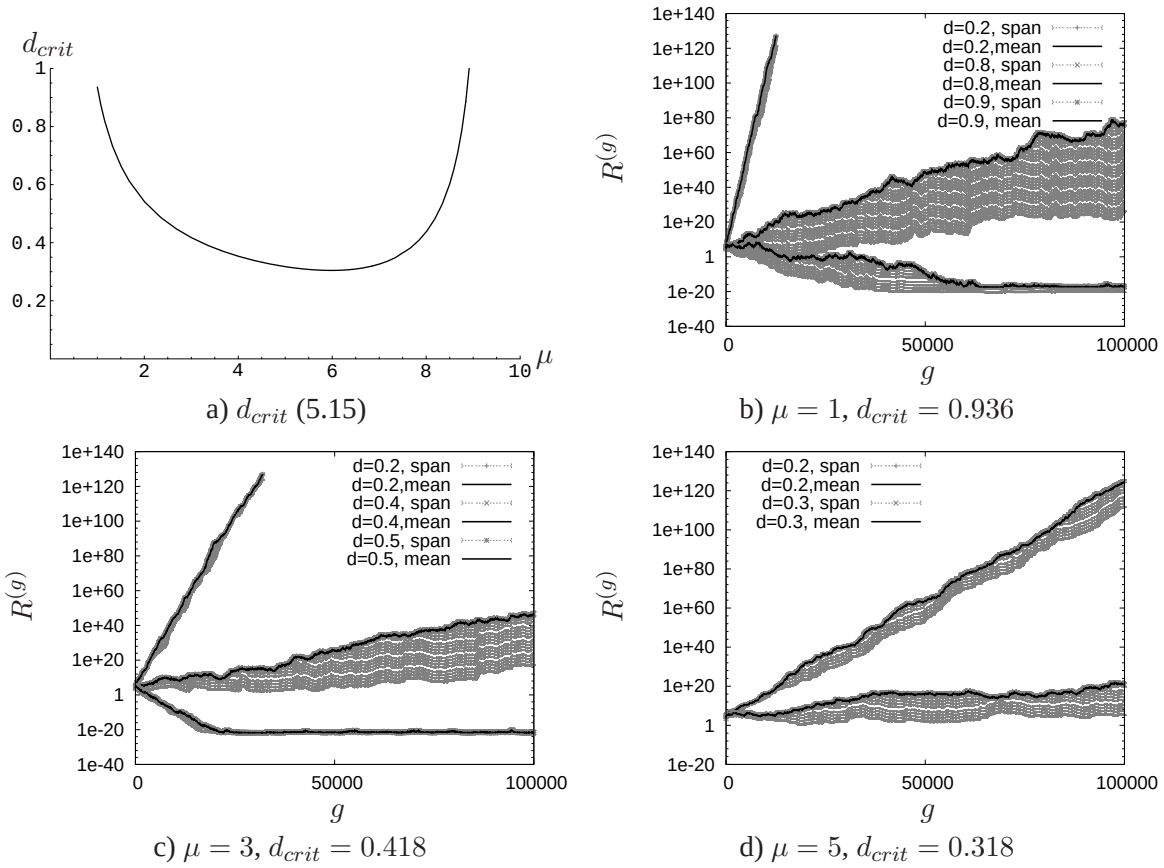


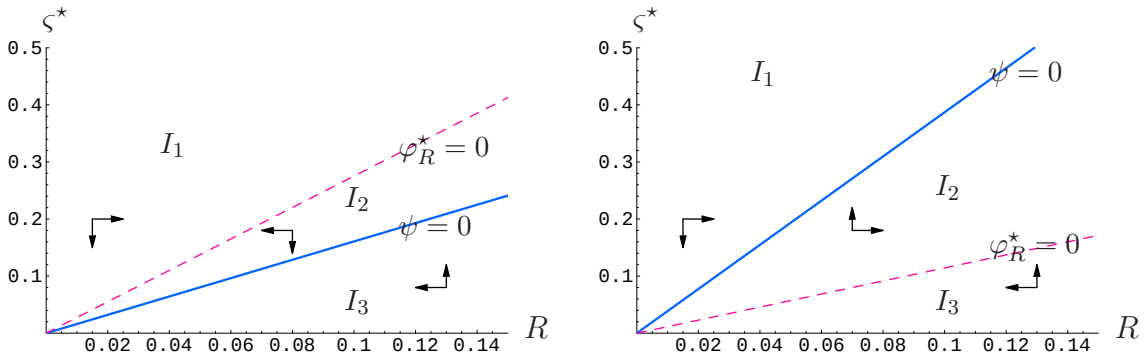
Figure 5.3: Results from $(\mu/\mu_I, 10)$-ES runs for the first $100,000$ generations for several choices of d ($N = 100$). Shown is every 20th value. Each data line is averaged over 20 runs. Also shown is the span between the minimal and maximal values.

mutation strengths inside this interval, the expected quality change is positive. That this is actually the case can be shown again by taking a look at Fig. 5.4. As the figure shows the ES – i.e., the system in $\sigma^\star$ and R – is expected to remain in region I_2 in the long run. The maximal mutation strength the ES can attain there is the SAR's zero

$$\varsigma^\star_{\psi_0} \quad = \quad R\left(\frac{1/2 + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\right)\frac{\sqrt{1+d^2}}{d}$$

(see (5.14)). It is interesting that both the zero of the SAR and the optimizer of the quality change show the same scaling behavior with respect to the gradient $\nabla f_R(x, R) = (1, -d)^{\mathrm{T}}$: Both are influenced by the reciprocal of the sine of the angle. Again, this is due to the fact that self-adaptation sees the fitness as a whole. In terms of changing d, the zero of SAR thus behaves as would be optimal for the quality change. A similar result holds for the dependence on R, of course.

Taking a closer look at (5.18) and (5.14) reveals that apart from the sine of the angle, the situation of the optimizer of the progress rate and the zero of the SAR on the sphere model reappears (cf. Section 4.1, 33ff). It can be shown by case inspection that for a long range of μ-values (except for $\mu = 1$ or $\mu \approx \lambda$) $\varsigma^\star_{\psi_0}$ is quite smaller than $\varsigma^\star_{opt}$ [74] (cf. Fig. 4.6, p. 45) and of course smaller than the



a) $\alpha = 1$, $d > d_{crit}$, (5.15) b) $\alpha = 1$, $d < d_{crit}$, (5.15)

Figure 5.4: The isoclines $\varphi^\star_R = 0$, (5.13), and $\psi = 0$, (5.14) as functions of the distance to the ridge R for $(1, 10)$-ES with $a = 1$. In a) region I_1 is characterized by $\Delta R > 0$, $\Delta\varsigma^\star < 0$, I_2 by $\Delta R < 0$, $\Delta\varsigma^\star < 0$, and I_3 by $\Delta R < 0$, $\Delta\varsigma^\star > 0$. Possible movements between the regions are $I_3 \to I_2$ and $I_1 \to I_2$. It is easy to see that I_1 and I_3 will be left eventually. The region I_2 cannot be left and the system in $\varsigma^\star$ and R approaches the origin. In b) region I_1 is characterized by $\Delta R > 0$, $\Delta\varsigma^\star < 0$, I_2 by $\Delta R > 0$, $\Delta\varsigma^\star > 0$, and I_3 by $\Delta R < 0$, $\Delta\varsigma^\star > 0$. Possible movements are $I_1 \to I_2$ and $I_3 \to I_2$, but I_2 cannot be left. The system diverges to infinity.

second zero of the quality change.

This has two effects: Self-adaptation is not expected to fail, i.e., to lead to a negative quality change. But only in the case of one parent the ES has the potential to realize mutation strengths relatively close to the optimizer – at least theoretically.

This does not necessarily exclude benefits due to recombination, though. Even if a recombinative strategy cannot reach its optimal mutation strength, the quality change associated with the mutation strength realized may be greater than that of the non-recombinative $(1, \lambda)$-ES.

Normalizing the Evolution Equations To answer the question, whether recombination on the sharp ridge is beneficial, the analysis must be extended. As it was shown, the optimal mutation strength scales with the distance to the axis. Assuming that self-adaptation works sufficiently well to adjust at least to this scaling behavior, it is postulated that $\langle\varsigma^{\star(g)}\rangle \approx cR^{(g)}$. In other words, if the normalization $\sigma^* := \sigma^\star/R$ is introduced, the existence of a stationary state of the normalized system

$$\begin{pmatrix} R^{(g+1)} \\ \langle\varsigma^{*(g+1)}\rangle \end{pmatrix} = \begin{pmatrix} R\left(1 - \frac{1}{N}\varphi^*(\sigma^*)\right) \\ \sigma^*\left(\frac{1+\psi(\sigma^*)}{1-\frac{1}{N}\varphi^*(\sigma^*)}\right) \end{pmatrix} \tag{5.19}$$

with the normalized progress rate

$$\varphi^*(\sigma^*) = \frac{dc_{\mu/\mu,\lambda}}{\sqrt{1+d^2}}\sigma^* - \frac{\sigma^{*2}}{2\mu} \tag{5.20}$$

and its second zero

$$\varsigma^*_{\varphi_R} = 2\mu c_{\mu/\mu,\lambda}\frac{d}{\sqrt{1+d^2}} \tag{5.21}$$

is assumed. Note, the mutation strength $\varsigma^* = \varsigma^*/r$ is normalized with respect to $R^{(g+1)} = R(1 - \varphi^*/N)$ thus introducing the denominator in the second line in (5.19). The equation for the mutation

strength normalized with respect to R,

$$\begin{aligned}\langle \varsigma^{*(g+1)} \rangle &= \sigma^* \Big(\frac{1+\psi(\sigma^*)}{1-\varphi_R^*(\varsigma^*)/N} \Big) \\ &= \sigma^* \Bigg(\frac{1+\tau^2(1/2+e_{\mu,\lambda}^{1,1} - \frac{d}{\sqrt{1+d^2}} c_{\mu/\mu,\lambda}\sigma^*)}{1-\frac{1}{N}(\frac{d}{\sqrt{1+d^2}} c_{\mu/\mu,\lambda}\sigma^* - \frac{\sigma^{*2}}{2\mu})} \Bigg) \end{aligned} \tag{5.22}$$

(see (5.7) and (5.20) has a stationary point with $\langle \varsigma^{*(g+1)} \rangle = \sigma^*$. Stationarity requires $\varphi_R^*(\varsigma_{st}^*) = -N\psi(\varsigma_{st}^*)$ which leads to a stationary mutation strength

$$\begin{aligned}\varsigma_{st}^* &= \sqrt{\frac{d^2}{1+d^2}} \mu c_{\mu/\mu,\lambda} \Big((1-N\tau^2) \\ &\quad + \sqrt{(1-N\tau^2)^2 + 2N\tau^2 \Big(\frac{1+d^2}{d^2} \Big) \frac{1/2+e_{\mu,\lambda}^{1,1}}{\mu c_{\mu/\mu,\lambda}^2}} \Big). \end{aligned} \tag{5.23}$$

as is illustrated in Appendix E, p. 215, Eqs. (E.3)-(E.6). The learning rate τ controls (5.23) – variating it between the zero of the progress rate and the zero of the SAR. Both are smaller than the zero of the quality change. Decreasing τ drives the stationary mutation strength towards the zero of the progress rate $\varsigma_{\varphi_{R0}^*}^* = 2\mu c_{\mu/\mu,\lambda}(d/\sqrt{1+d^2})$ (5.21), while increasing the learning rate results in the stationary mutation strength going to the zero of the SAR $\varsigma_{\psi_0}^* = (\sqrt{1+d^2}/d)(1/2+e_{\mu,\lambda}^{1,1})/c_{\mu/\mu,\lambda}$. That is to say, the maximal possible mutation strength cannot be attained in the stationary state for finite τ. Equation (5.23) is connected with a expected positive normalized quality change

$$\begin{aligned}\overline{\Delta Q_{st}^*} &= d\mu c_{\mu/\mu,\lambda}^2 \Bigg((1-N\tau^2) - \sqrt{(1-N\tau^2)^2 + 2N\tau^2 \Big(\frac{1+d^2}{d^2} \Big) \frac{1/2+e_{\mu,\lambda}^{1,1}}{\mu c_{\mu/\mu,\lambda}^2}} \Bigg) \\ &\quad \times \Bigg(1 - \frac{d^2}{2(1+d^2)} \Bigg((1-N\tau^2) \\ &\quad - \sqrt{(1-N\tau^2)^2 + 2N\tau^2 \Big(\frac{1+d^2}{d^2} \Big) \frac{1/2+e_{\mu,\lambda}^{1,1}}{\mu c_{\mu/\mu,\lambda}^2}} \Bigg) \Bigg). \end{aligned} \tag{5.24}$$

Recombination and the Stationary State The system behaves similarly as if the ES were on the sphere. Unlike to the sphere, though, a divergence of the distance occurs for $d < d_{crit}$. Furthermore, the zero of the SAR is greater than the zero of the progress rate and equals the maximal mutation strength that can be reached in the stationary state.

The question that remains concerns potential benefits from recombination – even if the actual optimal mutation strength with respect to the quality change is unattainable. This paragraph aims at shedding some light on this question. Recall that increasing the learning rate results in greater stationary mutation strengths and with it in higher quality changes. Operating with relatively large learning rates is advisable regardless of the strategy applied. But in this case, the influence of the zero of the SAR may outweigh that of the zero of the progress rate.

Let us first consider (5.23). For increasing τ-values the stationary mutation strength behaves more and more like the zero of the SAR. Concerning recombination, this is not beneficial since it is largest for $\mu = 1$ or $\mu \approx \lambda$ (cf. Section 4.1).

To an extent, the quality change (5.24) behaves differently, since there are additional dependences on μ as Fig. 5.5 illustrates. But for $N\tau^2 \to \infty$, the effects of recombination of the quality change can be easily examined by plugging (5.14) into (5.15)

$$\overline{\Delta Q^*(\varsigma^*_{\psi_0})} = \left(\frac{1+d^2}{d}\right)\left(\frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\right)\left(c_{\mu/\mu,\lambda} - \frac{1}{2\mu}\left(\frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\right)\right). \tag{5.25}$$

Interestingly, there are cases for which using recombination leads to advantages if λ is sufficiently large. But the μ which optimizes (5.25) is extremely small in relation to λ – ranging from $\mu = 1$ for very small λ values over $\mu = 2$ for $\lambda = 13$ to $\mu = 5$ for $\lambda = 100,000$.

Benefits from recombination appear for sufficiently small τ-values. As long as the stationary mutation strength behaves approximately as the zero of the progress rate, recombination increases the stationary mutation strength until $\mu \approx \lambda/2$ and the stationary quality change until $\mu \approx 1/5, \ldots, 1/3\lambda$. But the improvement by increasing τ surpasses the improvement by recombination with this ratio by far.

Figure 5.5 compares the stationary normalized mutation strength (5.23) with the results of experiments for two choices of τ. As can be seen, the larger the τ-value, the smaller the number μ for which the quality change starts to decline which is in accordance with the experiments as Fig.5.5 shows. Also visible is the influence of the learning rate τ on the prediction quality. Observed and predicted values are close together for smaller learning rates. In the case of the larger learning rate, greater deviations occur. The behavior as a function of the parent number μ is very similar, though.

5.1.2 The Parabolic Ridge: A Stationary State

Let us now consider the parabolic ridge, i.e., $\alpha = 2$ and $F(\mathbf{y}) = y_1 - d(\sum_{i=2}^{N} y_i^2) =: x - dR^2$, as a representative of ridge functions with $\alpha \geq 2$. As in the case of the sharp ridge, we start considering the deterministic system in R and $\sigma^\star$: The deterministic evolution equations in the case of the parabolic ridge are given by

$$\begin{aligned} r &= R - \frac{1}{N}\varphi^\star_R(\sigma^\star, R) \\ \langle \varsigma^{\star(g+1)} \rangle &= \sigma^\star\Big(1 + \psi(\sigma^\star, R)\Big). \end{aligned} \tag{5.26}$$

The progress rates $\varphi^\star_R$, $\varphi^\star_x$, and the SAR ψ were obtained in Appendices B.2.2 and C.1.2 as

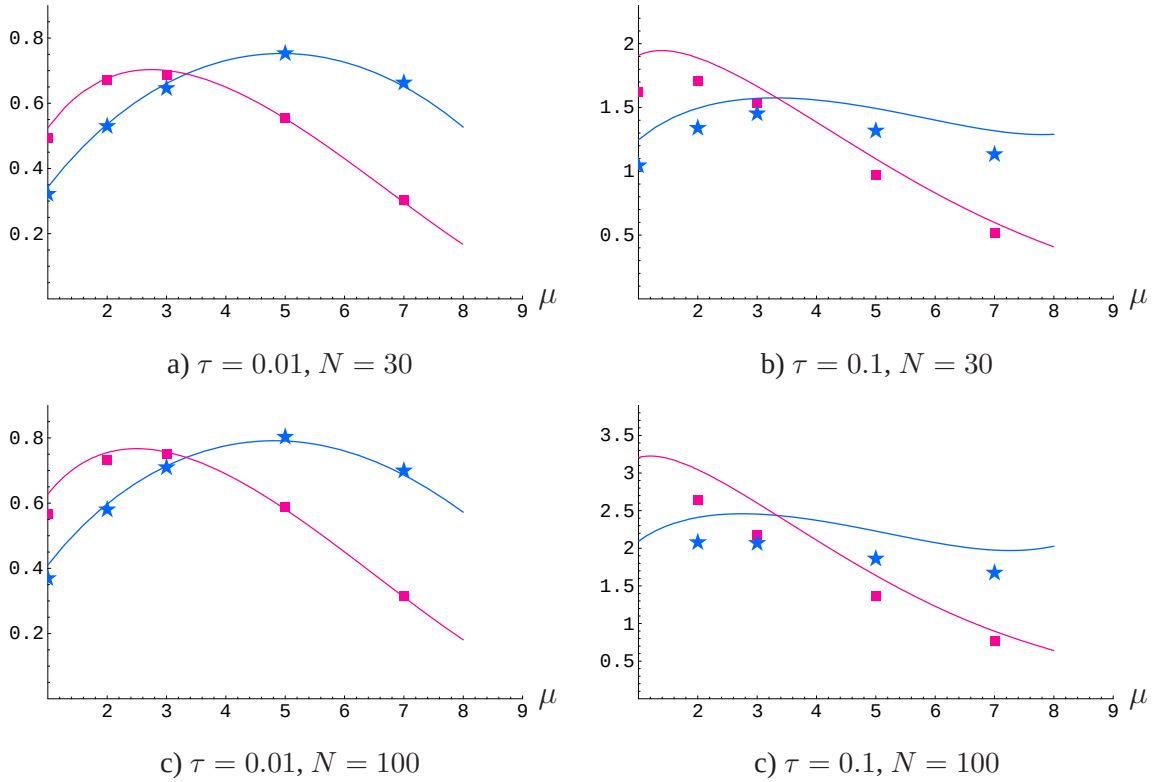
$$\varphi^\star_R(\sigma^\star, R) = \frac{d\alpha R^{\alpha-1} c_{\mu/\mu,\lambda}}{\sqrt{1 + d^2\alpha^2 R^{2\alpha-2}}}\sigma^\star - \frac{\sigma^{\star 2}}{2R\mu} \tag{5.27}$$

and

$$\varphi^\star_x(\sigma^\star, R) = \frac{c_{\mu/\mu,\lambda}}{\sqrt{1 + d^2\alpha^2 R^{2\alpha-2}}}\sigma^\star \tag{5.28}$$

for $\tau = 0$ and $N \to \infty$ and

$$\psi(\sigma^\star) = \tau^2\left(\frac{1}{2} + e^{1,1}_{\mu,\lambda} - \frac{c_{\mu/\mu,\lambda}}{R}\sqrt{\frac{d^2\alpha^2 R^{2\alpha-2}}{1 + d^2\alpha^2 R^{2\alpha-2}}}\,\sigma^\star\right). \tag{5.29}$$



a) $\tau = 0.01$, $N = 30$ b) $\tau = 0.1$, $N = 30$

c) $\tau = 0.01$, $N = 100$ c) $\tau = 0.1$, $N = 100$

Figure 5.5: The stationary mutation strength (5.23) and quality change (5.17) for some $(\mu/\mu_I, 10)$-ES with self-adaptation on the sharp ridge. Each data point was sampled over $100,000$ generations. Figs. a) and b) show the results for $N = 30$, c) and d) those for $N = 100$. The quality change is given by the red line.

First of all, note that the influence of the distance to the ridge is different compared to the case of the sharp ridge. Consider first the progress rate (5.27). In the case of the sharp ridge, the distance R only influenced the loss part of (5.27). Now, it also appears in the gain part. Similarly, the linear part of the SAR is influenced by an additional function of the distance.

The R-Dependence of the Zero Points

Let us start with the evolution of the mutation strength. The present mutation strength is increased if the value of the SAR (5.29) is positive and decreased otherwise. The SAR is a monotonously decreasing function in $\varsigma^\star$ with only one zero $\varsigma^\star_{\psi_0}$ which depends on the ridge factors over $d\alpha R^{\alpha-1}$ and furthermore on $R = R^{(g)}$

$$\begin{aligned}\varsigma^\star_{\psi_0} &= R\frac{1/2 + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\sqrt{\frac{1+\alpha^2 d^2 R^{2\alpha-2}}{\alpha^2 d^2 R^{2\alpha-2}}} \\ &= \varsigma^{\star sph}_{\psi_0}\sqrt{\frac{1+\alpha^2 d^2 R^{2\alpha-2}}{\alpha^2 d^2 R^{2\alpha-2}}}.\end{aligned} \tag{5.30}$$

The zero (5.30) only differs from the normalized (with respect to N) zero of the SAR for the sphere model

$$\varsigma^{\star\,sph}_{\psi_0} := R\frac{1/2 + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}} \tag{5.31}$$

(see [74]) by the square root which equals the reciprocal of the sine of the slope angle of the gradient. It is easy to see that

$$\begin{aligned}
\lim_{R\to\infty} \varsigma^{\star}_{\psi_0} &= \frac{1/2 + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}} \lim_{R\to\infty} \sqrt{\frac{1 + \alpha^2 d^2 R^{2\alpha-2}}{\alpha^2 d^2 R^{2\alpha-4}}} = \infty \\
\lim_{R\to 0} \varsigma^{\star}_{\psi_0} &= \begin{cases} \infty & \text{if } \alpha > 2 \\ \frac{1/2+e^{1,1}_{\mu,\lambda}}{2dc_{\mu/\mu,\lambda}} & \text{if } \alpha = 2 \\ 0 & \text{if } \alpha = 1 \end{cases}
\end{aligned} \tag{5.32}$$

holds. As one can see, if R increases, the SAR (5.29) tends to increase the zero of the SAR in turn. That is, larger and larger mutation strengths are expected to lead towards an increase. For decreasing distances, there are two different behaviors for $\alpha \geq 2$. In the case of $\alpha < 2$, the zero of the SAR goes to infinity as $R \to 0$. In the case of $\alpha = 2$, taking the limit of the zero leads to a strictly positive value. All mutation strengths greater than this limit value are expected to increase. The important point is that the ES maintains a strictly positive $\varsigma^{\star}_{\psi_0}$ – provided that $\alpha > 1$. In the case of the sharp ridge it goes to zero. These behaviors can be traced back to the local shape of the ridge, i.e., to the gradient at R, $\nabla f(x, R) = (1, -d\alpha R^{\alpha-1})^{\mathrm{T}}$. Let us reconsider the SAR (5.29)

$$\psi(\sigma^{\star}) = \tau^2\Big(1/2 + e^{1,1}_{\mu,\lambda} - c_{\mu/\mu,\lambda}\frac{d\alpha R^{\alpha-1}}{\sqrt{1 + (d\alpha R^{\alpha-1})^2}}\frac{\sigma^{\star}}{R}\Big).$$

The slope of the gradient of quadratic (or higher) ridge functions depends in stark contrast to the sharp ridge on the distance to the ridge axis. If the distance is large, the SAR resembles its sphere model equivalent. As the distance to the axis decreases, the angle between axis and gradient becomes smaller. The sine approaches zero and counteracts to some extend the normal reaction of the sphere model to increase the loss part. If $\alpha > 2$, the SAR behaves as it is required for linear functions: Every mutation strength is increased. In the case of the parabolic ridge, the reaction is different and falls short of the requirement for linear functions since only mutation strengths smaller than a distinct value are increased and otherwise decreased.

The R-evolution remains to be considered. The progress rate $\varphi^{\star}_R$ (5.27) is strictly positive in the interval $\varsigma^{\star} \in]0, 2R\mu\, c_{\mu/\mu,\lambda}\, \sqrt{(\alpha^2 d^2 R^{2\alpha-2})/(1 + \alpha^2 d^2 R^{2\alpha-2})}[$. The second zero of the progress rate (5.27) reads

$$\begin{aligned}
\varsigma^{\star}_{\varphi_{R0}} &= 2R\mu c_{\mu/\mu,\lambda}\sqrt{\frac{\alpha^2 d^2 R^{2\alpha-2}}{1 + \alpha^2 d^2 R^{2\alpha-2}}} \\
&= \varsigma^{\star\,sph}_{\varphi_R}\sqrt{\frac{\alpha^2 d^2 R^{2\alpha-2}}{1 + \alpha^2 d^2 R^{2\alpha-2}}}
\end{aligned} \tag{5.33}$$

with

$$\varsigma^{\star\,sph}_{\varphi_R} := 2R\mu c_{\mu/\mu,\lambda} \tag{5.34}$$

denoting the normalized (with respect to N) zero of the progress rate in the case of the sphere model [23]. Again the zero of the sphere model appears weighted in this case by the sine of the slope angle of the gradient and not by its reciprocal. As a result, it can be shown that the zero of the progress rate behaves in accordance with the distance to the ridge, i.e.,

$$\begin{aligned} \lim_{R\to\infty} \varsigma^\star_{\varphi_R} &= 2\mu c_{\mu/\mu,\lambda} \lim_{R\to\infty} \sqrt{\frac{\alpha^2 d^2 R^{2\alpha}}{1+\alpha^2 d^2 R^{2\alpha-2}}} = \infty \\ \lim_{R\to 0} \varsigma^\star_{\varphi_R} &= 2\mu c_{\mu/\mu,\lambda} \lim_{R\to 0} \sqrt{\frac{\alpha^2 d^2 R^{2\alpha}}{1+\alpha^2 d^2 R^{2\alpha-2}}} = 0. \end{aligned} \tag{5.35}$$

As seen, one of the first obvious differences between the sharp ridge ($\alpha = 1$) and higher-order ridge functions ($\alpha \geq 2$) is that only in the case of the latter, the SAR (5.29) eventually stops the mutation strength from following every decrease of the distance. Furthermore, only for $\alpha = 1$ both zeros (5.30) and (5.33) are linear functions in R.

A Stationary State

Figure 5.6 illustrates the behavior of the $(\sigma^\star, R)^{\mathrm{T}}$-system depicting the so-called isoclines (see, e.g., [33]) $\varphi^\star_R = 0$ and $\psi = 0$ as functions of R. The area below the isocline $\psi = 0$ is characterized by $(\sigma^\star, R)^{\mathrm{T}}$-combinations for which the SAR is positive and the mutation strength is expected to increase. Similarly, the area below $\varphi^\star_R = 0$ is characterized by a positive progress rate and thus by an expected decrease of the distance to the ridge. If R increases, the SAR tends to increase larger and larger mutation strengths. This effects in turn the R-evolution. Here, the zero of the progress rate increases as well. Mutation strengths that result in an expected decrease of the distance are increasing. On the other hand, if R decreases, the zero of the progress rate decreases as well. Mutation strengths that would increase the distance are thus also decreasing. But the potential answer of the $\sigma^\star$-evolution is either to increase an increasing range of mutation strengths or at least every mutation strength smaller than a limit. Thus, neither a convergence of $R \to 0$, i.e., a convergence to the axis, nor a divergence of $R \to \infty$ occurs.

As Fig. 5.6 shows there is a stationary state of the $(\varsigma^\star, R)^{\mathrm{T}}$-system. In the stationary state the $\varsigma^\star$- and the R-evolution come to a halt (on average) – i.e., the evolution strategy is expected to fluctuate at a certain distance to the axis. Apart from the trivial stationary state with $\varsigma^\star = 0$, the stationary state is characterized by requiring that both the SAR (5.29) and the progress rate $\varphi^\star_R$ (5.27) are zero. Therefore, the stationary states of the system (5.26) are given by

$$\begin{pmatrix} R_{st} \\ \varsigma^\star_{st} \end{pmatrix} = \begin{pmatrix} c \\ 0 \end{pmatrix} \tag{5.36}$$

with an arbitrary $c \in \mathbb{R}$ and

$$\begin{pmatrix} R_{st} \\ \varsigma^\star_{st} \end{pmatrix} = \begin{pmatrix} \frac{1}{2d}\sqrt{\frac{1/2+e^{1,1}_{\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda}-1/2-e^{1,1}_{\mu,\lambda}}} \\ \frac{\sqrt{2\mu}}{2d}\frac{1/2+e^{1,1}_{\mu,\lambda}}{\sqrt{2\mu c^2_{\mu/\mu,\lambda}-1/2-e^{1,1}_{\mu,\lambda}}} \end{pmatrix} \tag{5.37}$$

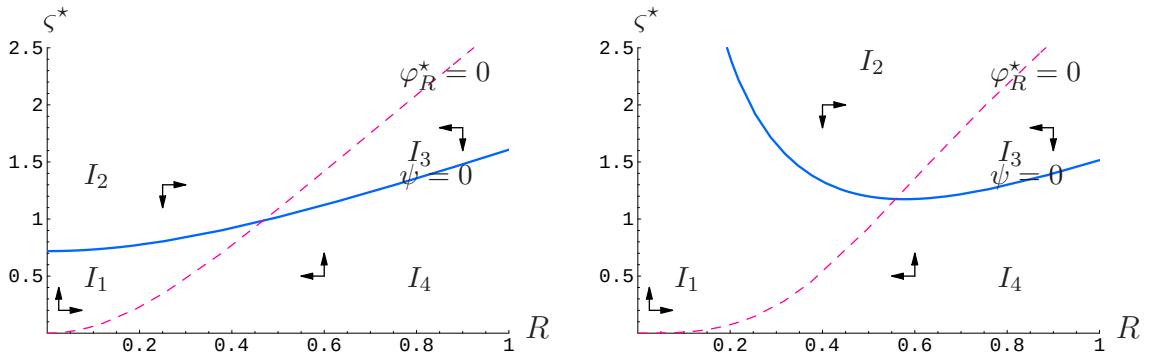


Figure 5.6: The zero points of the progress rate $\varphi^\star_R$ and ψ as functions of the distance to the ridge R for $(1, 10)$-ES with $d = 1$. Region I_1 is characterized by $\Delta R > 0$, $\Delta\varsigma^\star > 0$, I_2 by $\Delta R < 0$, $\Delta\varsigma^\star > 0$, I_3 by $\Delta R < 0$, $\Delta\varsigma^\star < 0$, and finally I_4 by $\Delta R > 0$, $\Delta\varsigma^\star < 0$. The system either leaves every region I_k again, i.e., it oscillates, or it converges to the equilibrium point.

in the case of $\alpha = 2$ and

$$\begin{pmatrix} R_{st} \\ \varsigma^\star_{st} \end{pmatrix} = \begin{pmatrix} \left(\frac{1/2 + e^{1,1}_{\mu,\lambda}}{\alpha^2 d^2 \left(2\mu c^2_{\mu/\mu,\lambda} - 1/2 - e^{1,1}_{\mu,\lambda} \right)} \right)^{1/(2\alpha-2)} \\ \sqrt{2\mu(1/2 + e^{1,1}_{\mu,\lambda})} \left(\frac{1/2 + e^{1,1}_{\mu,\lambda}}{\alpha^2 d^2 \left(2\mu c^2_{\mu/\mu,\lambda} - 1/2 - e^{1,1}_{\mu,\lambda} \right)} \right)^{1/(2\alpha-2)} \end{pmatrix} \tag{5.38}$$

for general $\alpha > 1$. The derivation can be found in Appendix E.1.2. In [19] an estimate of the stationary distance of CSA-ES was obtained for the parabolic ridge, i.e., for $\alpha = 2$, $R_{st} \propto 1/(2d)$ which also reappears in the case of σ-self-adaptation. Concerning the ridge constant d, both mechanisms show the same behavior. Again, a similarity with the situation on the noisy sphere model appears [25]. On the noisy sphere, the stationary distance scales with the standard deviation (noise strength) of additive normally distributed noise, i.e., $R_{st} \propto \sigma_\epsilon$. Therefore, $1/d$, the inversion of the weighting constant of the embedded sphere, seems similar to the noise term σ_ϵ. A further similarity is of course the stationary state of both evolutions – the evolution of R and the mutation strength $\varsigma^\star$. The stationarity of the latter has an additional effect: Due to $\psi = 0$, the learning rate τ does not have any influence in the stationary state at least if (5.29) is used. It is interesting to note a further property of the stationary state in the case of the parabolic ridge provided that $2\mu c^2_{\mu/\mu,\lambda} \gg 1/2 + e^{1,1}_{\mu,\lambda}$. This condition holds for example for sufficiently large λ and for recombinative ES with the usual ratio of $\mu : \lambda$, i.e., $\mu \not\approx \lambda$ and $\mu \not\approx 1$. In this case, the stationary distance and and mutation strength are very close to

$$\begin{pmatrix} R_{appr} \\ \sigma^\star_{appr} \end{pmatrix} = \begin{pmatrix} \frac{1}{2d} \sqrt{\frac{1/2 + e^{1,1}_{\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda}}} \\ \frac{1}{2d} \frac{1/2 + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}} \end{pmatrix}. \tag{5.39}$$

The approximate stationary distance in (5.39) is formally equal to the square root of the quotient of the zero of the SAR $(1/2 + e^{1,1}_{\mu,\lambda})/c_{\mu/\mu,\lambda}$ (4.17), p. 37, and the zero of the progress rate $2\mu c_{\mu/\mu,\lambda}$ (4.13) in the sphere model case. The difference is the appearance of the ridge constant d in (5.39).

The mutation strength in (5.39) is the zero of the SAR (5.29) obtained for $R = 0$. As far as it concerns the SAR and the evolution of the mutation strength, the situation of the stationary state does not differ much from the hypothetical case that the ES is on the axis itself having achieved the subgoal of optimizing the embedded sphere.

As was shown, in the case of the parabolic ridge, the system admits a stationary state with a strictly positive stationary mutation strength and distance to the axis. Unlike the case of the sharp ridge, neither a convergence to the axis nor a divergence of the distance occurs. The parameter d is still important since it determines the steady state distance to the ridge and with it the mutation strength.

While the progress towards the axis stops (on average), there is progress parallel to the axis. The stationary progress rate reads

$$\begin{aligned}\varphi^{\star}_{x\,st} &= \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+4d^2R^2_{st}}}\varsigma^{\star}_{st} \\ &= \sqrt{(1/2+e^{1,1}_{\mu,\lambda})(2\mu c^2_{\mu/\mu,\lambda}-1/2-e^{1,1}_{\mu,\lambda})} \\ &\quad\times\Big(\frac{1/2+e^{1,1}_{\mu,\lambda}}{\alpha^2 d^2(2\mu c^2_{\mu/\mu,\lambda}-1/2-e^{1,1}_{\mu,\lambda}))}\Big)^{1/(2\alpha-2)}\end{aligned} \tag{5.40}$$

which can be easily obtained by inserting the stationary mutation strength and distance (5.37) into the progress rate (5.28). In the case of $\alpha = 2$,

$$\varphi^{\star}_{x\,st} = \frac{1/2+e^{1,1}_{\mu,\lambda}}{2d} \tag{5.41}$$

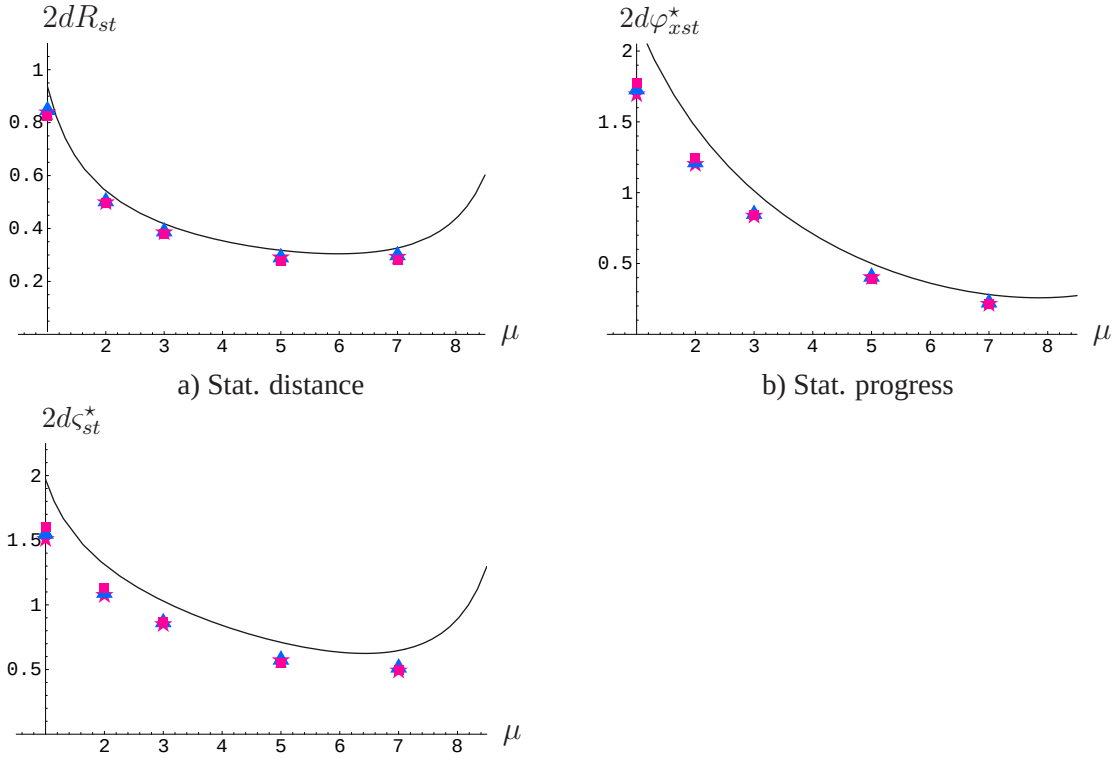
is obtained. For a more detailed derivation, the reader is referred to Appendix E.1.2. As it can be inferred from (5.41), the stationary progress depends on the population parameters μ and λ and on the ridge parameter d. Since $e^{1,1}_{\mu,\lambda} > 0$ for $\mu < \lambda/2$, $e^{1,1}_{\mu,\lambda} = 0$ for $\mu = \lambda/2$ and $e^{1,1}_{\mu,\lambda} < 0$ for $\mu > \lambda/2$, the stationary progress rate (5.41) is greater than $1/(4d)$ in the first, equals $1/(4d)$ in the second, and is smaller in the last case. It should be noted that for larger d-constants, the ES is able to get closer to the axis. In a sense, it succeeds better in fulfilling the partial aim of optimizing the sphere. However, this results finally in an overall performance loss: The larger the weighting constant of the sphere part, the smaller the progress parallel to the axis. It is interesting, to note a further characteristic of self-adaptive ES on the parabolic ridge. Due to the stationarity of the $\varsigma^{\star}$-evolution, the learning rate τ is not expected to have an influence on the progress rate. That is, the usual tuning parameter of self-adaptation cannot be used to improve the performance of the ES. Using the deterministic approach, the equations show that the learning rate may only have an influence as long as the evolution of the mutation strength has not reached a steady state.

This situation differs fundamentally from the stationary state on the undisturbed sphere. On the sphere, the evolution of the mutation strength – normalized with respect to the distance and the search space dimensionality – reached a stationary state. The ES tuned the mutation strength proportionally to the distance to the optimizer. The evolution of the distance and the evolution of the non-normalized mutation strength progressed still. Due to the non stationarity of the latter evolution, the learning rate could be used as a control parameter. Here, both evolutions come to a halt. Only as long as the ES

progresses towards the axis, the learning rate may be used to improve the performance. Once the stationary state is reached, however, the system is independent of the choice of τ. Obviously, the ES still adjust the mutation strength according to the distance. Here however, it is an adjustment to the distance to the ridge and not to the optimizer. The value of the x-component does not have any influence. Only the evolution of the distance to the axis and the evolution of the mutation strength are coupled.

The stationarity of both evolutions was already encountered in the case of self-adaptive ES on the noisy sphere. There, additive noise with a constant noise strength kept the ES from reaching the optimizer. After a transient phase, a self-adaptive ES reached a stationary state of the distance and the mutation strength. Something similar occurs on the parabolic ridge. Self-Adaptation works on the parabolic ridge in the sense that no premature convergence occurs. However, once the stationary state is reached, self-adaptation could be switched off. The mutation strength is stationary and does not reflect any movement or position in x-direction.

Figure 5.7 shows the stationary mutation strength, distance (5.37), and progress rate (5.41) as functions of the parent number μ comparing them with the results of experiments. The agreement



a) Stat. distance b) Stat. progress

c) Mutation strength

Figure 5.7: The stationary distance, mutation strength (5.37), and progress rate (5.41) for some $(\mu/\mu_I, 10)$-ES with self-adaptation on the parabolic ridge. Each data point was sampled over $200,000$ generations ($N = 30$, $N = 100$) and $900,000$ ($N = 1000$) generations. The stars indicate the results for $N = 1000$, the triangles those for $N = 100$, and the boxes those for $N = 30$.

between prediction and experiment is good, but it should be mentioned that the experimental data are lower than predicted. Interestingly, the experiments do not show significant differences between

high and low search space dimensionalities. This is somewhat surprising and up to now not fully understood.

The Influence of Recombination Equations (5.37) and (5.41) can be used to investigate the influence of recombination. As Fig. 5.7 shows, the maximal progress and the maximal mutation strength occur in the case of non-recombinative ES, i.e., for $\mu = 1$, which is confirmed by experiments. Introducing multi-parent recombination does not lead to any advantage at all. The stationary progress on the axis is influenced by the stationary mutation strength and distance and therefore by the progress rate (towards the axis) (5.27) and the SAR (5.29). In the case of the parabolic ridge, it can be given as $\varphi^{\star}_{x\,st} = (1/2+e^{1,1}_{\mu,\lambda})/(2d)$ (5.41). The effects of recombination are reflected by the progress coefficient $e^{1,1}_{\mu,\lambda}$ which stems from the SAR. All other influences have averaged out. The progress coefficient $e^{1,1}_{\mu,\lambda}$ is a monotonously decreasing function in μ for $\mu < \lambda/2$. As already pointed out, the first zero point is at $\mu = \lambda/2$. Afterwards, negative values are assumed until the coefficient approaches zero again for $\mu = \lambda$. As a result, the stationary progress (5.41) is largest for $\mu = 1$ and ES does not benefit from recombination. At first glance this is contradictory to the results obtained by Oyman [79]. He pointed out that recombination has positive effects in the case of the parabolic ridge since the distance to the ridge is decreased. This enables larger progress rates [79, p. 139]. This result was obtained for constant mutation strengths, though. Unfortunately, in the case of self-adaptive ES recombination also decreases the mutation strength mirroring the response of the zero of the SAR. The decrease of the distance fails to counteract this trend leading to a falling progress rate with μ.

Using the deterministic variant of the evolution equations, two main results can be derived: First, a stationary state other than $\varsigma^{\star} = 0$ exists which admits positive progress. Second, the ES does not benefit at all from multi-parent recombination.

5.2 Self-Adaptive ES on Noisy Ridge Functions

In this section, the analysis is extended to ridge functions that are disturbed by noise. The noise term is modeled using the standard noise model of an additive normally distributed term with zero mean and standard deviation (noise strength) σ_ϵ. As before, it is assumed that the noise strength is constant and does not depend on the position in the search space. First, the sharp ridge is addressed before an analysis of the parabolic ridge is provided.

5.2.1 Noise is Beneficial: Noise on the Sharp Ridge

As it was shown in Section 5.1.1, there are generally two types of behavior shown by evolution strategies on the sharp ridge $f(x, R) = x - dR$: Dependent on the ridge parameter d, an ES either converges prematurely or continuously enlarges the mutation strength and the distance to the ridge. But what happens if noise influences the fitness evaluations? Is d still a decisive parameter then?

The Noise Model and the Evolution Equations

To investigate the behavior of ES on the noisy sharp ridge, the standard noise model is used. Therefore, the noise is modeled using an additive normally distributed random variable with a constant (uniform) standard deviation σ_ϵ. Therefore, given the object vector $\mathbf{y}$ the apparent fitness reads

$$F_R(\mathbf{y}) \quad = \quad y_1 - d\sqrt{\sum_{i=2}^{N} y_i^2} + \epsilon$$

$$
\begin{aligned}
&= y_1 - d\sqrt{\sum_{i=2}^{N} y_i^2} + \sigma_\epsilon \mathcal{N}(0,1) \\
\Rightarrow f(x,R) &= : x - dR + \sigma_\epsilon \mathcal{N}(0,1).
\end{aligned} \tag{5.42}
$$

Again, due to the form of the undisturbed fitness function $f(x,R) = x - dR$ three variables are of interest: the x-component denoting the change parallel to the ridge axis, the lateral component R measuring the distance to the ridge, and the mutation strength σ.

To investigate the change of these three variables over time, the deterministic evolution equations

$$
x^{(g+1)} = x^{(g)} + \varphi_x^\star(\sigma^\star, \sigma_\epsilon^\star)/N \tag{5.43}
$$

$$
r = R - \varphi_R^\star(R, \sigma^\star, \sigma_\epsilon^\star)/N \tag{5.44}
$$

$$
\langle \varsigma^{\star(g+1)} \rangle = \sigma\Big(1 + \psi(R, \sigma^\star, \sigma_\epsilon^\star)\Big) \tag{5.45}
$$

are considered. Note, the normalized versions $\varphi_x^\star := N\varphi_x$, $\varphi_R^\star := N\varphi_R$, $\sigma^\star := N\sigma$, and $\sigma_\epsilon^\star := N\sigma_\epsilon$ are used again. The progress rates $\varphi_x^\star$, $\varphi_R^\star$, and the SAR ψ are obtained in Appendices B.2-C. Here, their main characteristics are shortly discussed.

The progress rate $\varphi_x^\star = N\mathrm{E}[x^{(g+1)} - x^{(g)}]$ obtained for $\tau = 0$ and $N \to \infty$ as

$$
\varphi_x^\star(\sigma^\star) = \frac{\sigma^{\star 2}}{\sqrt{(1+d^2)\sigma^{\star 2} + \sigma_\epsilon^{\star 2}}} c_{\mu/\mu,\lambda} \tag{5.46}
$$

is – as before – a quasi-linear function of the mutation strength. Again, there is no influence of x itself on its own expected change. At first sight, the progress parallel to the axis is diminished by the noise strength – but as it is shown later on the situation is more complicated.

The progress rate $\varphi_R^\star = N\mathrm{E}[R^{(g)} - R^{(g+1)}]$, i.e.,

$$
\varphi_R^\star(\sigma^\star, R) = \frac{d\sigma^{\star 2}}{\sqrt{(1+d^2)\sigma^{\star 2} + \sigma_\epsilon^{\star 2}}} c_{\mu/\mu,\lambda} - \frac{\sigma^{\star 2}}{2R\mu} \tag{5.47}
$$

consists of a gain and a quadratic loss part and can be interpreted as a function of the mutation strength. Again, (5.47) is determined for $N \to \infty$ and $\tau = 0$. The influence of the additional parameters, i.e., the ridge parameters and noise strength, enter the progress rate over the gain part. The loss part is only influenced by the parent number μ and the distance to the ridge.

The SAR $\psi = \mathrm{E}[(\langle \varsigma^{\star(g+1)} \rangle - \langle \varsigma^{\star(g)} \rangle)/\langle \varsigma^{\star(g)} \rangle]$,

$$
\begin{aligned}
\psi(\sigma^\star, R) = \tau^2 \Bigg(& 1/2 + e_{\mu,\lambda}^{1,1} \frac{(1+d^2)\sigma^{\star 2}}{(1+d^2)\sigma^{\star 2} + \sigma_\epsilon^{\star 2}} \\
& - c_{\mu/\mu,\lambda} \frac{d\sigma^{\star 2}}{R\sqrt{(1+d^2)\sigma^{\star 2} + \sigma_\epsilon^\star}^2} \Bigg)
\end{aligned} \tag{5.48}
$$

is determined under the assumption $\tau \ll 1$ and for $N \to \infty$. Noise influences the gain and the loss part, but the influence of the distance is only present in the loss part. It should be noted that the prediction quality of (5.48) deteriorates relatively fast with increasing $\sigma^\star$ for smaller values of N.

Since the required functions are given, the analysis can be started. As in the previous sections, the evolution of the x-component does not influence the evolution of R and $\sigma^\star$. The evolution parallel to the axis is governed by the evolutions of the remaining two state variables instead. Therefore, it suffices to consider the system in R and $\sigma^\star$

$$\begin{pmatrix} r \\ \langle \varsigma^{\star(g+1)} \rangle \end{pmatrix} = \begin{pmatrix} R - \varphi_R(R, \sigma^\star, \sigma_\epsilon^\star)/N \\ \sigma^\star \Big(1 + \psi(R, \sigma^\star, \sigma_\epsilon^\star)\Big) \end{pmatrix}. \tag{5.49}$$

There are two behaviors the system $(R, \sigma^\star)^{\mathrm{T}}$, (5.49), is expected to show: a convergence to a stationary state (which may be either a convergence to a point or to an orbit) or a divergence of R and $\sigma^\star$. The following part of this section is devoted to deriving conditions for divergence or convergence.

Introducing Normalizations

To make the equations easier to handle, an additional normalization for the progress rates $\varphi_x^\star$, (5.47), $\varphi_R^\star$, (5.48), the mutation strength, and the noise strength is introduced. The aim is to eliminate the distance to the ridge R in (5.46) – (5.49). Setting thus $\sigma^* := \sigma^\star/R$, $\sigma_\epsilon^* := \sigma_\epsilon^\star/R$, $\varphi_x^* := \varphi_x^\star/R$, and $\varphi_R^* := \varphi_R^\star/R$, the progress rates (5.46), (5.47), and the SAR (5.48) change to

$$\varphi_x^*(\sigma^*, \sigma_\epsilon^*) = \frac{\sigma^{*2}}{\sqrt{(1+d^2)\sigma^{*2} + \sigma_\epsilon^{*2}}} c_{\mu/\mu,\lambda}, \tag{5.50}$$

and

$$\varphi_R^*(\sigma^*, \sigma_\epsilon^*) = \frac{\sigma^{*2} d}{\sqrt{(1+d^2)\sigma^{*2} + \sigma_\epsilon^{*2}}} c_{\mu/\mu,\lambda} - \frac{\sigma^{*2}}{2\mu}, \tag{5.51}$$

and

$$\psi(\sigma^*, \sigma_\epsilon^*) = \tau^2 \left(1/2 + e_{\mu,\lambda}^{1,1} \frac{(1+d^2)\sigma^{*2}}{(1+d^2)\sigma^{*2} + \sigma_\epsilon^*} - c_{\mu,\mu,\lambda} \frac{d\sigma^{*2}}{\sqrt{(1+d^2)\sigma^{*2} + \sigma_\epsilon^{*2}}}\right). \tag{5.52}$$

The evolution equations for R and σ^* change accordingly. Note, $\langle \varsigma^{*(g+1)} \rangle := \langle \varsigma^{\star(g+1)} \rangle / r = R(1 - \varphi_R^*/N)$

$$\begin{pmatrix} r \\ \langle \varsigma^{*(g+1)} \rangle \end{pmatrix} = \begin{pmatrix} R\Big(1 - \varphi_R^*(\sigma^*, \sigma_\epsilon^*)/N\Big) \\ \sigma^* \Big(\frac{1+\psi(\sigma^*, \sigma_\epsilon^*)}{1-\varphi_R^*(\sigma^*, \sigma_\epsilon^*)/N}\Big) \end{pmatrix}. \tag{5.53}$$

As in the case of the noisy sphere (cf. Section 4.2), a third g-dependent variable appears: The normalized noise strength $\sigma_\epsilon^{*(g)}$ changes with $R^{(g)}$. However, the (direct) influence of $R^{(g)}$ can be eliminated leading to the new evolution equation

$$\sigma_\epsilon^{*(g+1)} = \frac{\sigma_\epsilon^*}{1 - \varphi_R^*(\sigma^*, \sigma_\epsilon^*)/N}. \tag{5.54}$$

Due to the normalization, the evolution of R neither influences the evolution of the mutation strength nor the evolution of the noise strength. As before in Section 4.2, its evolution is decoupled and it suffices to analyze the two-dimensional evolution equations

$$\begin{pmatrix} \sigma_\epsilon^{*(g+1)} \\ \langle \varsigma^{*(g+1)} \rangle \end{pmatrix} = \begin{pmatrix} \frac{\sigma_\epsilon^*}{1-\varphi_R^*(\sigma^*, \sigma_\epsilon^*)/N} \\ \sigma^* \Big(\frac{1+\psi(\sigma^*, \sigma_\epsilon^*)}{1-\varphi_R^*(\sigma^*, \sigma_\epsilon^*)/N}\Big) \end{pmatrix}. \tag{5.55}$$

For the remainder of this section, the evolution equations (5.55) are used.

Determining Stationary States

Following the previous approach in Section 5.1, the stationary points are determined first. Stationary points of (5.55) defined as $(\sigma_\epsilon^{*(g+1)}, \langle\varsigma^{*(g+1)}\rangle)^{\mathrm{T}} = (\sigma_\epsilon^*, \sigma^*)^{\mathrm{T}}$ can be calculated in a straightforward way. The stationary solution of the evolution equation for σ_ϵ^* in (5.55) requires the progress rate (5.51) to be zero. Therefore, the task is to find zero points of (5.51) which are also stationary points for the evolution of the mutation strength. It can be shown that the stationary state of the system (5.55) is given by

$$\begin{pmatrix} \sigma^*_{\epsilon\, stat_1} \\ \varsigma^*_{stat_1} \end{pmatrix} = \begin{pmatrix} c \\ 0 \end{pmatrix} \tag{5.56}$$

with $c \in \mathbb{R}, c \geq 0$ or by

$$\begin{pmatrix} \sigma^*_{\epsilon\, stat} \\ \varsigma^*_{stat} \end{pmatrix} = \begin{pmatrix} 2d\mu c_{\mu/\mu,\lambda} \sqrt{\frac{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1) - 2e^{1,1}_{\mu,\lambda} - 1}{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) - 2e^{1,1}_{\mu,\lambda}}} \\ \frac{2d\mu c_{\mu/\mu,\lambda}}{\sqrt{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) - 2e^{1,1}_{\mu,\lambda}}} \end{pmatrix} \tag{5.57}$$

(see Appendix E.2.1, p. 223). In the situation of (5.57), the ES does not converge to the axis. Note, the stationary mutation strength goes to zero for $d \to 0$ and to $2\mu c_{\mu/\mu,\lambda}/\sqrt{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}}$ for $d \to \infty$. The normalized noise strength behaves proportional to d: For $d \to 0$, $\sigma^*_\epsilon(d) \to 0$ and $\sigma^*_\epsilon(d) \to \infty$ for $d \to \infty$. Both variables are completely determined by the population parameter μ and λ and of course by the ridge parameter d.

The noise effectively stops the ES from approaching the ridge axis arbitrarily close. Similar to the sphere model, a stationary distance to the ridge axis can be derived

$$R_{stat} = \frac{N\sigma_\epsilon}{2d\mu c_{\mu/\mu,\lambda}} \sqrt{\frac{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) - 2e^{1,1}_{\mu,\lambda}}{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1) - 2e^{1,1}_{\mu,\lambda} - 1}}. \tag{5.58}$$

See Appendix E.2.1 for the derivation.

As it has been shown, system (5.55) comes to a halt either by a loss of step-size control in an arbitrary distance to the axis or by attaining stationary values for the mutation strength and the distance.

The question remains under which conditions this stationary state exists. As it is shown next, the weighting constant d is again decisive w.r.t. μ and λ. Let $\mu \leq \lambda/2$. The stationary state is only defined for weighting constants d which fulfill

$$d \geq d_{crit} := \sqrt{\frac{2e^{1,1}_{\mu,\lambda} + 1}{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1}}. \tag{5.59}$$

See Appendix E.2.1, p. 226f. for a derivation.

First of all, note that the same d-value as for the undisturbed sharp ridge decides over the existence of the stationary state. The reason for this is that only in the case of $d > d_{crit}$, the ES moves towards

the axis at all. In the case of $d < d_{crit}$, the distance to the axis and the mutation strength enlarge. Since the noise strength remains constant, it gradually looses its influence until the ES behaves as if it were optimizing the noise-free ridge. The constraint $\mu \leq \lambda/2$ is sufficient but not necessary. The equations generally hold unless $\mu \approx \lambda$ but a sharp boundary cannot be given. Let $\mu \leq \lambda/2$. If $d \geq d_{crit}$ holds then (5.57) is a locally stable fix-point of (5.55) for the ES considered whereas (5.56) is instable (see Appendix E.2.1).

Let us sum up our findings. For $d > d_{crit}$, ES moves towards the ridge axis as in the undisturbed case. Contrary to its behavior in the noise-free case, it converges to a stationary state that has a well-defined distance to the axis. The evolution of R comes to a halt on average and the ES travels parallel to the axis direction.

The normalized stationary progress rate behaves in the same manner as the normalized stationary mutation strength: It does not depend on the noise strength, i.e., it stays constant. This can be easily seen since the stationary progress rate can be re-expressed as

$$\varphi^*_{x\,stat} = \frac{{\sigma^*_{stat}}^2}{2\mu d} \tag{5.60}$$

since due to the stationary of the R evolution, $\sqrt{{\sigma^*_{stat}}^2 + \sigma^{*2}_{\epsilon\,stat}} = 2\mu c_{\mu/\mu,\lambda}$ holds. The normalized progress (5.60) depends on the stationary mutation strength in (5.57), the constant d and on the population parameters.

Only after switching to the non-normalized versions a dependence on the noise strength appears. This is due to the linear dependence of the stationary distance on the noise strength.

The non-normalized progress parallel to the ridge axis can be obtained by plugging (5.57) and (5.58) into (5.50) as

$$\begin{aligned}\varphi^{st}_x(\sigma_\epsilon) &= \sigma_\epsilon c_{\mu/\mu,\lambda} \sqrt{\frac{1}{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1) - 2e^{1,1}_{\mu,\lambda} - 1}} \\ &\quad \times \sqrt{\frac{1}{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) - 2e^{1,1}_{\mu,\lambda}}}\end{aligned} \tag{5.61}$$

The derivation can be found in Appendix E.2.1, p. 228. The non-normalized progress rate scales linear with the noise strength – a behavior that is also exhibited by the non-normalized mutation strength

$$\varsigma_{stat} = \frac{\sigma_\epsilon}{\sqrt{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1) - (2e^{1,1}_{\mu,\lambda} + 1)}} \tag{5.62}$$

(cf. Appendix E.2.1, p. 228). The larger the noise strength, the farer the ES stays away from the ridge axis. In turn, the mutation strength and the stationary progress increase with the distance scaling linearly with the noise strength. This is a result of the optimization behavior. If the ES is far away from the ridge axis the influence of the noise in comparison to that of R is relatively small. Provided d is relatively large, the ES starts approaching the axis but is hindered in the convergence by the noise. Higher noise strengths result in larger distances to the axis. This in turn allows larger stationary mutation strengths. Larger mutation strengths are connected with higher expected gains on the ridge axis. On the sharp ridge, noise with a constant strength effectively stops the ES from optimizing

the contained sphere model and enforces a more significant gain on the axis. This only holds for sufficiently large ridge parameters d.

If d is too small, the ridge is not being tracked and a divergence of the distance occurs. The distance R increases which lessens the (relative) influence of the noise. In this case, the ES will gradually start to behave as if it were optimizing the undisturbed sharp ridge – striding away from the axis with a negative progress rate φ_R– but with an overall positive quality change, i.e., the gain parallel to the axis surpasses the loss due to the distances increase. This case was already discussed in Section 5.1.1.

Recapitulating, note that in the case of constant noise strength the ES shows a similar behavior as in the noise free case: The choice of the ridge parameter d decides whether the ridge is tracked or not. If the ridge is not tracked the influence of the noise decreases and the ES attains a positive though not optimal quality change. If the ridge is tracked, the ES cannot converge to the ridge due to the noise. Noise is actually beneficial since it prevents premature convergence: The larger the noise, the larger the mutation strength and with it the progress in axis direction. In contrast to the former case of divergence from the axis, the ES progresses with a constant non-normalized mutation strength (on average). In short, noise on the sharp ridge either soon looses its influence or has an actual positive influence as it keeps the ES from optimizing only the sphere part.

Comparison with Experiments

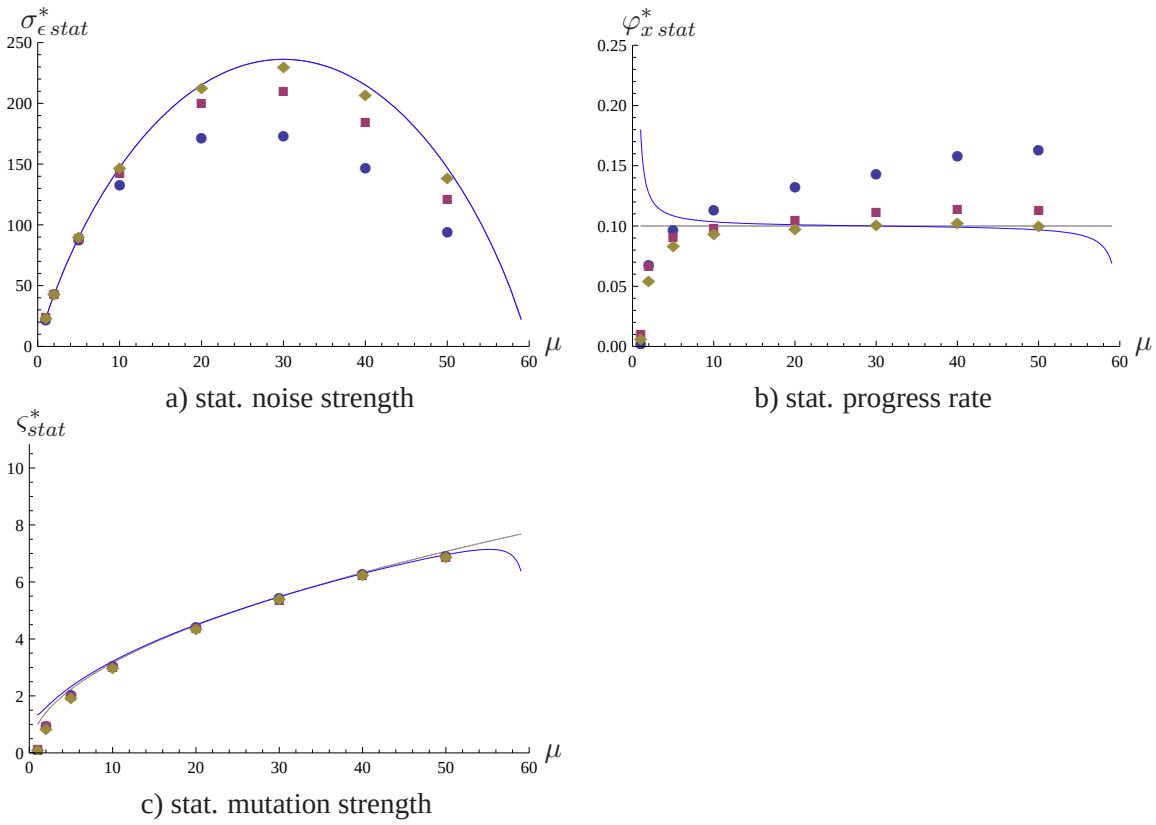
Figure 5.8 shows a comparison between the normalized stationary values (5.57) and (5.60) and experimental data for three search space dimensionalities $N = 30$, $N = 100$, and $N = 500$. The prediction quality improves with the search space dimensionality with the exception of the stationary mutation strength in (5.57). In this case, the agreement is good even for the lower dimensional search space $N = 30$ and does not improve visibly if N increases. It can be seen though that (5.57) tends to overestimate the stationary mutation strength if the parent number is relatively small. This probably causes in turn the greater deviations of (5.60) from the experimental progress rates for these μ values. While the agreement of (5.60) with the experiments is quite good for large N in general, the experimental results for $\mu = 1$ and $\mu = 2$ are far lower than predicted. Figure 5.9 compares the non-normalized values (5.58), (5.61), and (5.62) with the results of experiments. Again, the prediction quality if relatively poor for $N = 30$ and improves with the search space dimensionality. As it can be seen, the experiments for $(1, 60)$-ES result in far smaller mutation strengths than predicted.

The Effects of Recombination

The effects of recombination remain to be addressed. Figure 5.9 shows the stationary mutation strength and progress rate as functions of the parent number μ. As it reveals, switching from $\mu = 1$ to $\mu > 1$ is not beneficial. To find out why, let us start with the normalized mutation and noise strength (5.57). Provided that the size of the offspring population is not small, (5.57) shows an interesting scaling behavior with respect to μ. If $2\mu c^2_{\mu/\mu,\lambda} \gg e^{1,1}_{\mu,\lambda}$ holds, the stationary point can be approximated with

$$\begin{pmatrix} \sigma^*_{\epsilon\, appr} \\ \varsigma^*_{appr} \end{pmatrix} = \begin{pmatrix} 2d\mu c_{\mu/\mu,\lambda} \\ \sqrt{\mu} \end{pmatrix}. \tag{5.63}$$

Equation (5.63) holds for $\mu \not\approx 1$ and $\mu \not\approx \lambda$ and large λ. Interestingly, it equals the scaling behavior on the noisy sphere (4.66) with only one exception, the ridge parameter d, which appears in the case



a) stat. noise strength

b) stat. progress rate

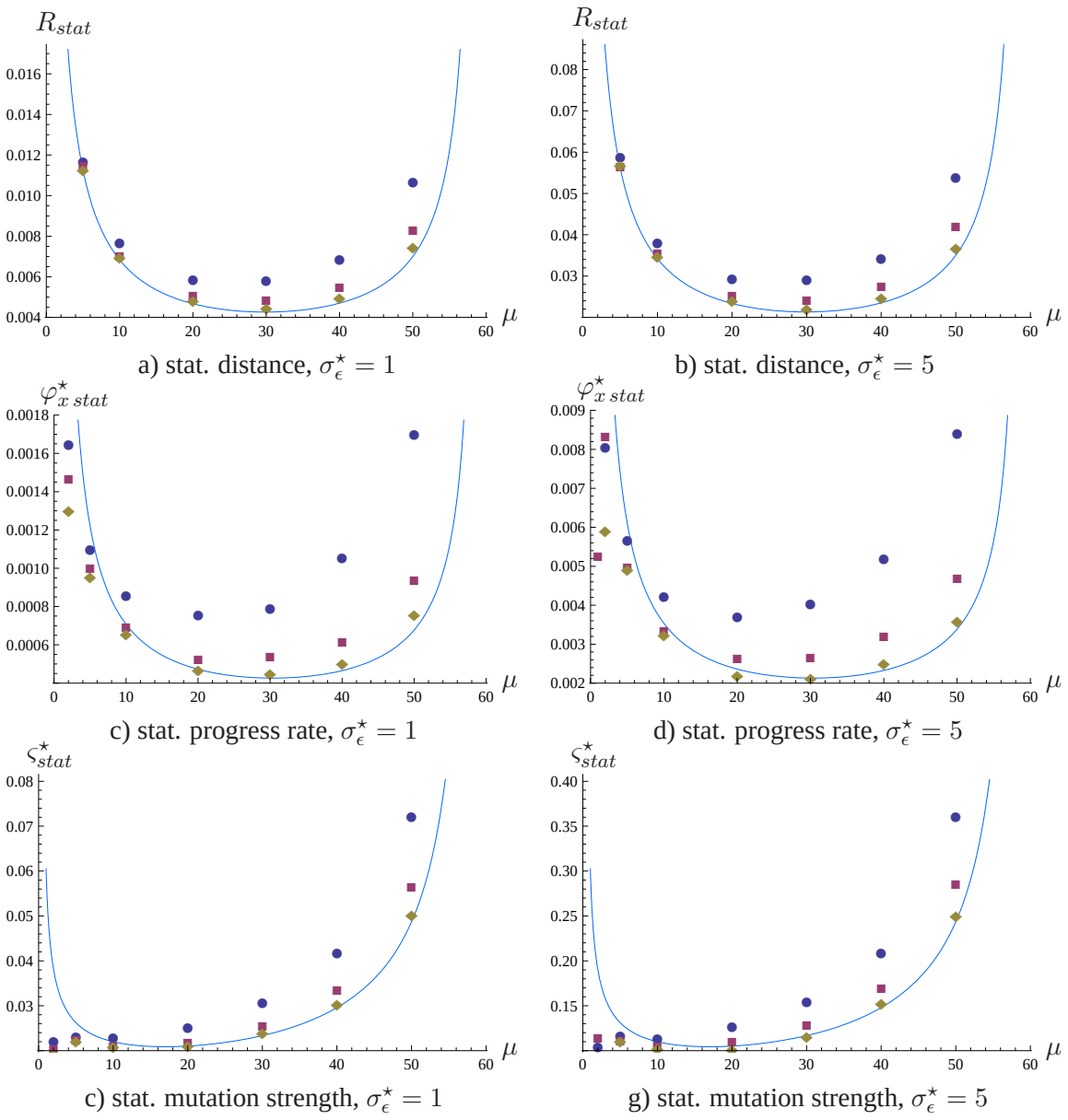
c) stat. mutation strength

Figure 5.8: The stationary noise strength , stationary mutation strength (both in (5.57)), and stationary progress rate (5.60) on the sh arp ridge as functions of μ. The results were averaged over several runs with different choices of σ_ϵ with $\sigma^\star_\epsilon = 1, 2, 3$, and 5. The search space dimensionalities are $N = 30$, $N = 100$, and $N = 500$. In the case of $N = 30$ each data point was sampled over $100,000$ generations for each noise strength and then averaged over all noise strengths, i.e., over a total of $4 \times 100,000$ generations. For $N = 100$ and $N = 500$ $4 \times 200,000$ generations were used. The results for $N = 30$ are denoted by the round points, those for $N = 100$ by the squares, whereas the results for $N = 500$ are given by the diamonds.

of the noise strength. Apparently, in this respect the ES behaves in a very similar manner on the noise sharp ridge as on the noisy sphere. The fact that the noisy sharp ridge and not the noisy sphere is to be optimized is not recognizable in the stationary mutation strength and as said concerning the stationary normalized noise only the presence of the weighting factor differentiates (5.63) from (4.66).

Recombination increases the mutation strength in (5.63) and (4.66). A similar result, though, holds for the normalized noise strength which increases with $2\mu c_{\mu/\mu,\lambda}$. This results in smaller distances to the ridge axis. With similar arguments as before, the scaling behavior of the distance to the ridge w.r.t. μ reads

$$R_{appr} = \frac{N\sigma_\epsilon}{2d\mu c_{\mu/\mu,\lambda}}. \tag{5.64}$$



a) stat. distance, $\sigma_\epsilon^\star = 1$ b) stat. distance, $\sigma_\epsilon^\star = 5$

c) stat. progress rate, $\sigma_\epsilon^\star = 1$ d) stat. progress rate, $\sigma_\epsilon^\star = 5$

c) stat. mutation strength, $\sigma_\epsilon^\star = 1$ g) stat. mutation strength, $\sigma_\epsilon^\star = 5$

Figure 5.9: The stationary distance (5.58), stationary mutation strength (5.62), and stationary progress rate (5.61) for constant noise on the sharp ridge as functions of μ. The search space dimensionalities are $N = 30$, $N = 100$, and $N = 500$. In the case of $N = 30$ each data point was sampled over $100,000$ generations, whereas for $N = 100$ and $N = 500$ $200,000$ generations were used. The results for $N = 30$ are denoted by the round points, those for $N = 100$ by the squares, whereas the results for $N = 500$ are given by the diamonds.

The approximate distance (5.64) is also the minimal possible distance that can be obtained. This can be inferred by using the stationary condition $\varphi_R^* = 0$

$$\varphi_R^* = 0 \iff \sigma^* = 0 \bigvee (1 + d^2)\sigma^{*2} + \sigma_\epsilon^{*2} = 4\mu^2 c_{\mu/\mu,\lambda}^2 = 0 \tag{5.65}$$

(see (5.50)) and letting $\sigma^* \to 0$.

While the decrease of the distance was beneficial on the sphere, it has the opposite effect on the ridge. The normalized stationary progress (5.60) reveals the problem: If $\mu \not\approx 1$ and $\mu \not\approx \lambda$, the influence of μ on the normalized progress rate is negligible. Because of $\varsigma^*_{stat} \propto \sqrt{\mu}$, the normalized progress rate does not differ much from

$$\varphi^*_{x\,appr} = \frac{1}{2d} \tag{5.66}$$

provided that λ is not small.

Since the normalized progress rate stays nearly constant, a problem is encountered in the case of the non-normalized variables. The non-normalized progress rate (5.61) scales approximately with $1/(2\mu c_{\mu/\mu,\lambda})$ and drops sharply if recombination is introduced.

The normalized noise scales with $2\mu c_{\mu/\mu,\lambda}$ and the distance therefore with $1/(2\mu c_{\mu/\mu,\lambda})$. This outperforms the increase of the normalized mutation strength with $\sqrt{\mu}$: The non-normalized mutation strength decreases with $1/(2\sqrt{\mu} c_{\mu/\mu,\lambda})$. Decreasing the mutation strength is necessary on the sphere. Since the ES is able to approach the optimizer more closely, the mutation strength must reflect this and decrease accordingly. On the ridge, though, this means that the mutation strength is decreased because the subgoal of optimizing the sphere is better realized. This does not equal a better achievement of the overall goal. Again, neither the position nor the gain in x-direction is reflected.

On first sight, recombination does not have any benefits. A caveat must be added, though. As on the sphere (cf. Section 4.2), the $(1, \lambda)$-ES looses step-size control in the stationary state – a behavior not predictable by the deterministic evolution equations. Therefore, as a rule the progress parallel to the axis halts and the ES stagnates prematurely. Recombination is therefore beneficial. The problem now consists in choosing μ sufficiently large so that the ES may stabilize the mutation strength and sufficiently small so that the progress does not decrease too far. Concerning the general behavior of the progress rate, $\mu \approx 2 - 5$ appears as a good choice – at least for the ES and noise strengths considered here.

5.2.2 Noise on the Parabolic Ridge

As it has been shown in the previous section, additive noise on the sharp ridge is actual beneficial: It keeps the ES from realizing the subgoal of optimizing the sphere. Since the ES cannot converge to the axis, it maintains a positive mutation strength. As result, there is progress in x-direction and no premature convergence occurs. The effects of noisy perturbations in the case of the remain to be addressed. Recall, the noisy parabolic ridge is defined as

$$\begin{aligned} f(\mathbf{y}) &= y_1 - d\sum_{i=2}^{N} y_i^2 + \epsilon \\ &= y_1 - d\sum_{i=2}^{N} y_i^2 + \sigma_\epsilon \mathcal{N}(0,1) \\ &:= x - dR^2 + \sigma_\epsilon \mathcal{N}(0,1). \end{aligned} \tag{5.67}$$

Again, the parameter σ_ϵ denotes the noise strength and is assumed to be constant.

The Evolution Equations and Progress Measures

As before, the first-order evolution equations without perturbation parts serve as the starting point for the analysis

$$\begin{aligned} x^{(g+1)} &= x^{(g)} + \varphi_x^\star(R,\sigma^\star,\sigma_\epsilon^\star)/N & (5.68)\\ r &= R^{(g)} - \varphi_R^\star(R,\sigma^\star,\sigma_\epsilon^\star)/N & (5.69)\\ \varsigma^\star &= \sigma^\star\Big(1+\psi(R,\sigma^\star,\sigma_\epsilon^\star)\Big). & (5.70)\end{aligned}$$

Before starting the analysis, we need the progress rates $\varphi_x^\star$ and $\varphi_R^\star$ and the SAR ψ. Their derivation can be found in Appendix B.2 and Appendix C.1.2. The progress rates

$$\varphi_x^\star(R,\sigma^\star,\sigma_\epsilon^\star) = \frac{c_{\mu/\mu,\lambda}}{\sqrt{(1+4d^2R^2)\sigma^{\star 2}+\sigma_\epsilon^{\star 2}}}\sigma^{\star 2} \quad (5.71)$$

and

$$\varphi_R^\star(R,\sigma^\star,\sigma_\epsilon^\star) = \frac{2dRc_{\mu/\mu,\lambda}}{\sqrt{(1+4d^2R^2)\sigma^{\star 2}+\sigma_\epsilon^{\star 2}}}\sigma^{\star 2} - \frac{\sigma^{\star 2}}{2R\mu} \quad (5.72)$$

are obtained for $N\to\infty$ and $\tau = 0$. The SAR

$$\psi(\sigma^\star,\sigma_\epsilon^\star) = \tau^2\Big(1/2 - \frac{2dc_{\mu/\mu,\lambda}\sigma^{\star 2}}{\sqrt{(1+4d^2R^2)\sigma^{\star 2}+\sigma_\epsilon^{\star 2}}} + e^{1,1}_{\mu,\lambda}\frac{(1+4d^2R^2)\sigma^{\star 2}}{(1+4d^2R^2)\sigma^{\star 2}+\sigma_\epsilon^{\star 2}}\Big) \quad (5.73)$$

is derived under the assumption $\tau \ll 1$ and for $N\to\infty$.

Determining Stationary Solutions

As before, the stationary state behavior of the ES is of interest. So first of all, the stationary states are determined starting with the evolution of the distance to the axis (5.69). Demanding stationarity of the R-evolution leads to an expression of the stationary mutation strength as a function of the distance

$$\begin{aligned} & 0 = \varphi_R^\star(R,\sigma^\star,\sigma_\epsilon^\star)\\ \Rightarrow\quad & 0 = \frac{2dRc_{\mu/\mu,\lambda}}{\sqrt{(1+4d^2R^2)\sigma^{\star 2}+\sigma_\epsilon^{\star 2}}}\sigma^{\star 2} - \frac{\sigma^{\star 2}}{2R\mu}\\ \Rightarrow\quad & \sigma^\star = 0 \bigvee \sigma^{\star 2} = \frac{16d^2R^4\mu^2c_{\mu/\mu,\lambda}^2}{1+4d^2R^2} - \frac{\sigma_\epsilon^{\star 2}}{1+4d^2R^2}. \end{aligned} \quad (5.74)$$

Otherwise, the distance $R^{(g)}$ to the ridge must be zero. Demanding further stationarity of the $\varsigma^\star$-evolution (5.70), either $\sigma^\star = 0$ or $\psi = 0$, i.e.,

$$0 = \tau^2\Big(1/2 - \frac{2dRc_{\mu/\mu,\lambda}\sigma^{\star 2}}{R\sqrt{(1+4d^2R^2)\sigma^{\star 2}+\sigma_\epsilon^{\star 2}}} + e^{1,1}_{\mu,\lambda}\frac{(1+4d^2R^2)\sigma^{\star 2}}{(1+4d^2R^2)\sigma^{\star 2}+\sigma_\epsilon^{\star 2}}\Big) \quad (5.75)$$

(cf. 5.73) have to hold. The mutation strength in (5.75) can be eliminated by inserting (5.74). We arrive at a third order polynomial in R^2

$$0 = R^6 - R^4 \frac{1}{4d^2}\Big(\frac{2e^{1,1}_{\mu,\lambda}+1}{4\mu c^2_{\mu/\mu,\lambda}-2e^{1,1}_{\mu,\lambda}-1}\Big) - \frac{R^2\sigma^{\star 2}_\epsilon}{8d^2\mu^2 c^2_{\mu/\mu,\lambda}}\Big(\frac{2\mu c^2_{\mu/\mu,\lambda}-e^{1,1}_{\mu,\lambda}}{4\mu c^2_{\mu/\mu,\lambda}-2e^{1,1}_{\mu,\lambda}-1}\Big) + \frac{e^{1,1}_{\mu,\lambda}\sigma^{\star 2}_\epsilon}{32d^4\mu^2 c^2_{\mu/\mu,\lambda}\Big(4\mu c^2_{\mu/\mu,\lambda}-2e^{1,1}_{\mu,\lambda}-1\Big)} \tag{5.76}$$

which can be solved analytically (see Appendix E.2.2). Since

$$\sigma^\star_{st} = \sqrt{16\Big(\frac{d^2R^4_{st}\mu^2c^2_{\mu/\mu,\lambda}}{1+4d^2R^2_{st}}\Big) - \frac{\sigma^{\star 2}_\epsilon}{1+4d^2R^2_{st}}}, \tag{5.77}$$

the solutions can be used to obtain the stationary mutation strength and with it the stationary progress parallel to the axis

$$\varphi^\star_{st} = \frac{c_{\mu/\mu,\lambda}}{\sqrt{(1+4d^2R^2_{st})\sigma^{\star\,2}_{st}+\sigma^{\star 2}_\epsilon}}\sigma^{\star\,2}_{st} = \frac{1}{4d\mu}\frac{\sigma^{\star\,2}_{st}}{R^2_{st}}. \tag{5.78}$$

The solutions of (5.76), however, are not very informative. Therefore, the influence of the noise will be discussed using Figs. 5.10-5.15.

Discussion of the Results and Comparison with Experiments

Figures 5.10-5.15 show the stationary distance, mutation strength, and progress parallel to the axis for some $(\mu/\mu_I, 60)$-ES. Also shown are the results of experiments for $N = 30$ and $N = 100$. The ridge constant d was set to $d = 5$. In the case of $N = 100$, all data points are obtained by sampling over $400,000$ generations in the stationary state whereas $200,000$ generations were sampled for $N = 30$. As long as μ is relatively small, there are deviations between experiments and predictions. This concerns especially the case of $\mu = 2$, i.e., small parent numbers. The prediction quality is generally better for larger noise strengths than for smaller. The exception is again the case of $\mu = 2$. Similarly to the case of the undisturbed parabolic ridge, increasing the search space dimensionality does not have any detectable influence on the prediction quality.

Again, it should be noted that in the case of the $(1,\lambda)$-ES, a similar problem as in the case of the sphere model appears: Once the fitness evaluations are overlaid by noise and the noise strength is too large, the $(1,\lambda)$-ES looses step-size control. The mutation strength is reduced to very small values and the progress rate drops significantly. This cannot be predicted by the deterministic evolution equations.

In the case of the distance (see Figs. 5.10 and 5.11), the prediction quality of the solution of (5.76) is good. Only for very small noise strengths, some deviations can be observed in the case of $\mu = 2$.

Greater deviations are observed in the case of the mutation strength (5.77). Especially, there are deviations for smaller parent numbers μ and small noise strengths. Equation (5.77) tends to overestimate the experimental results (see Figs. 5.13 and 5.12). Increasing μ improves the agreement provided that the noise strength is large.

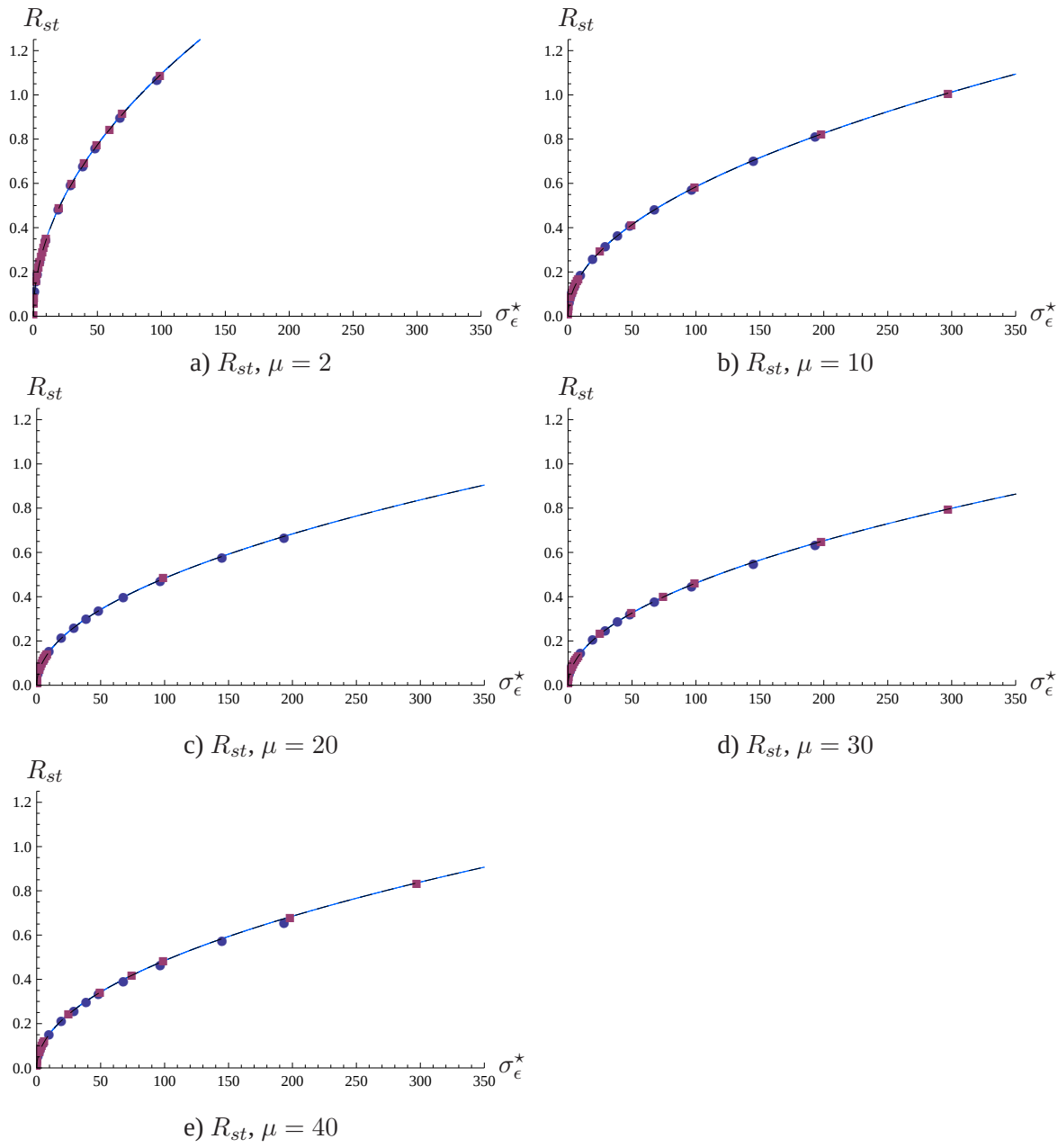
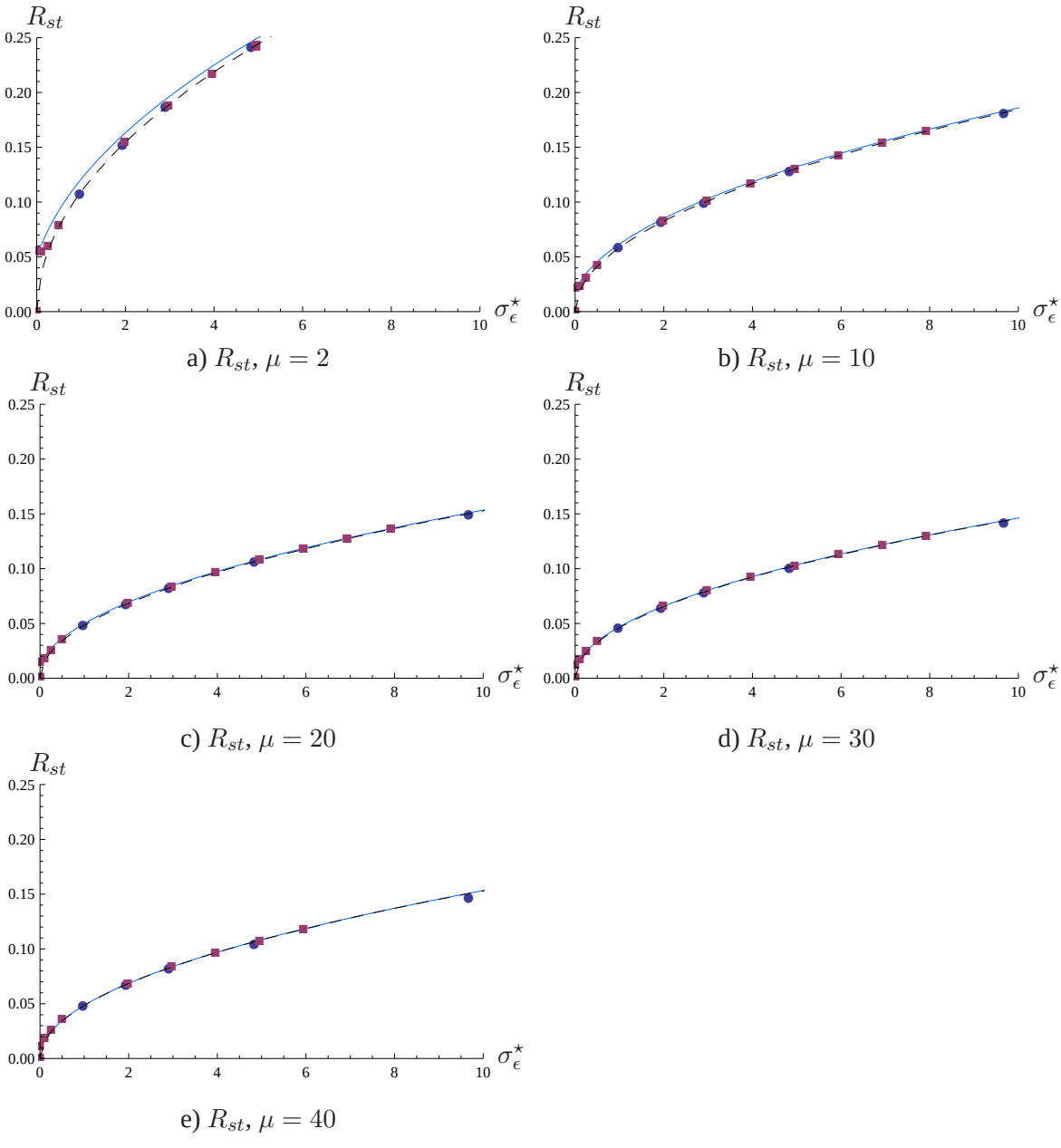


Figure 5.10: The stationary distance obtained using (5.76) (colored) in comparison to the stationary distance estimate (5.80) (black, dashed). As it can be seen, the curves are very similar and cannot be differentiated easily. The points denote the results of experiments with $(\mu/\mu_I, 60)$-ES with $d = 5$ for $N = 30$ (disks) and $N = 100$ (squares). Each data point was averaged over $400,000$ ($N = 100$) and $200,000$ ($N = 30$) generations in the stationary state.

A problem occurs in the case of the progress rate (5.78). Equation (5.78) shows a similar behavior as the experiments with respect to varying the noise strength. That is, for $\mu < \lambda/2$, the progress rate decreases with the noise and goes to a limit value. For $\mu = \lambda/2$, it remains constant. Finally, for



a) R_{st}, $\mu = 2$

b) R_{st}, $\mu = 10$

c) R_{st}, $\mu = 20$

d) R_{st}, $\mu = 30$

e) R_{st}, $\mu = 40$

Figure 5.11: The stationary distance obtained using (5.76) in comparison to the stationary mutation strength estimate (5.80). Both are shown as functions of the noise strength. The figure depicts the result for smaller noise strengths. As it can be seen, the predictions of (5.76) and (5.80) are very similar – except for the case $\mu = 2$.

$\mu > \lambda/2$ it increases with the noise approaching again a limit value. However, (5.78) overestimates the results. Furthermore, the convergence to the limit is not as fast as in the experiments (see Figs. 5.14 and 5.15). Equation (5.78) only serves well to predict the stationary state progress rate for large noise strengths. The exception is of course the case of $\mu = \lambda/2$. All examined strategies with intermediate recombination converge to very similar limits.

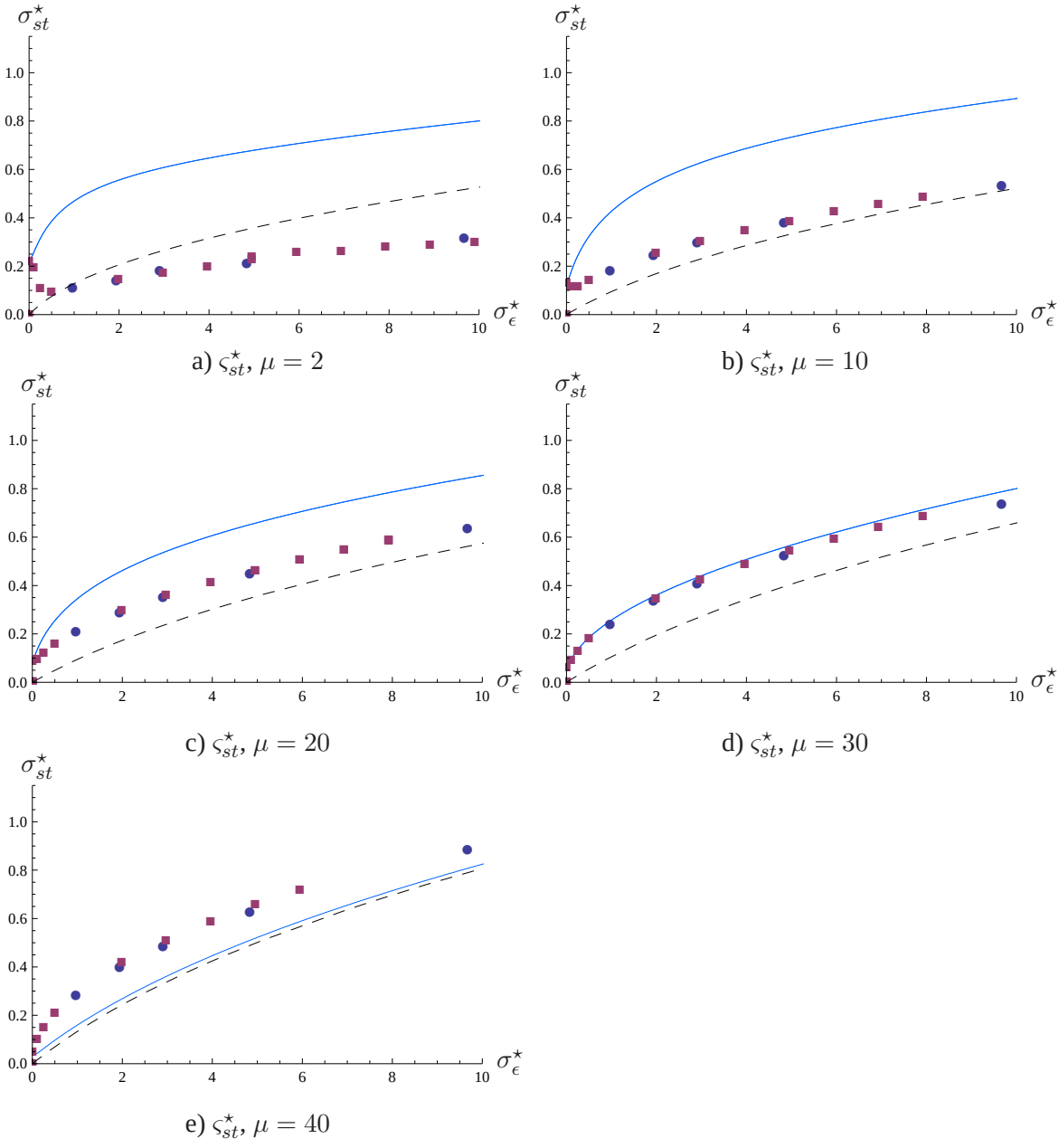
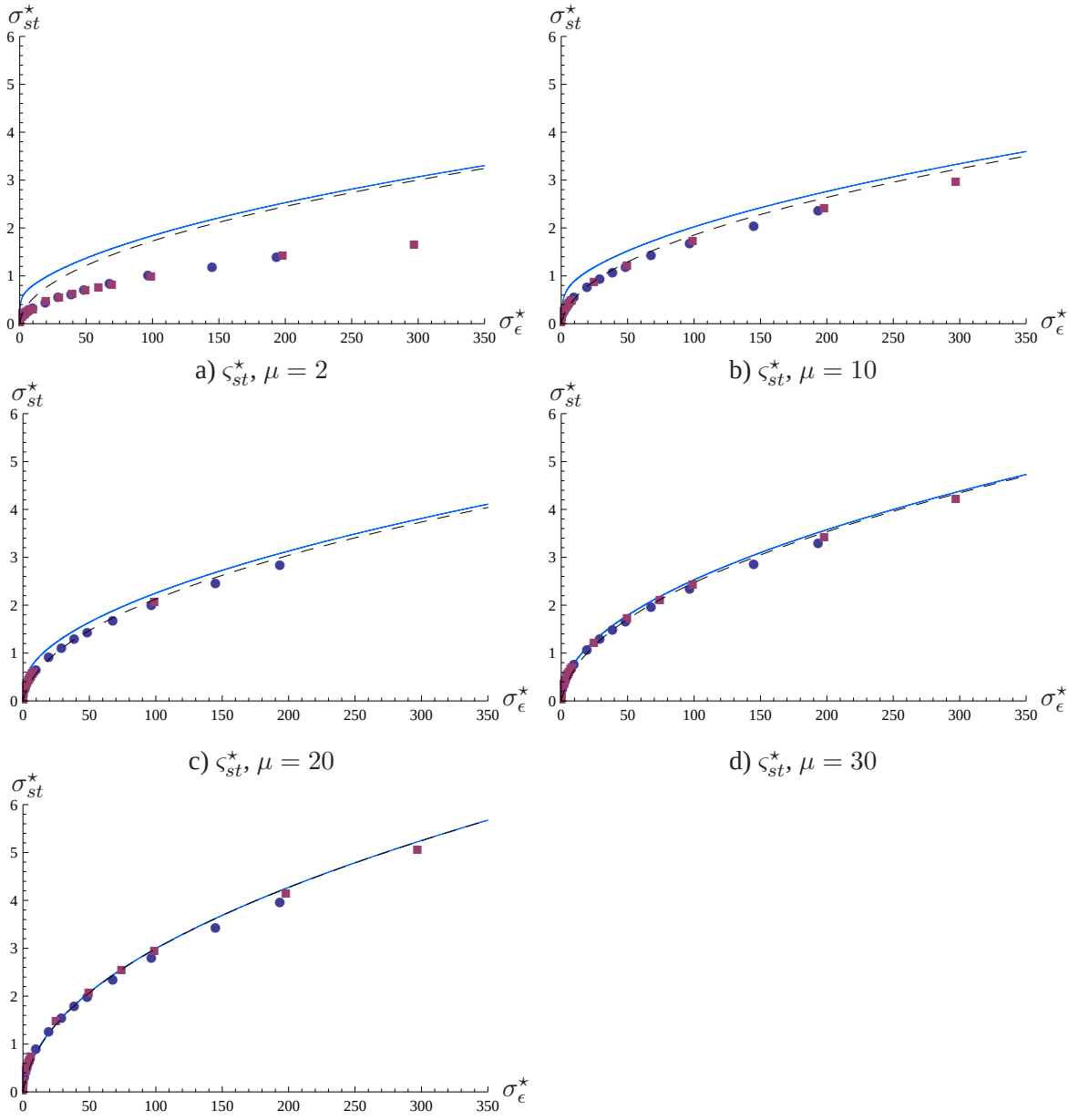


Figure 5.12: The stationary mutation strength obtained using (5.77) (dashed lines) in comparison to the stationary mutation strength estimate (5.81). As it can be seen, this scale reveals some differences between experiments and prediction.

As seen, noise generally increases the distance to the axis and the mutation strength. As the experiments showed, the transition from the zero-noise level to very small noise-levels may cause an initial decrease but this is soon overcome and the variables increase. As it can be discerned from Figs. **??** - **??**, the increase is approximately proportional to the square root of the noise for both the mutation strength and the distance. The slope of the increase is determined by the population parameters μ and λ. The stationary progress rate is influenced by the square of the ratio of the



a) $\varsigma^\star_{st}$, $\mu = 2$

b) $\varsigma^\star_{st}$, $\mu = 10$

c) $\varsigma^\star_{st}$, $\mu = 20$

d) $\varsigma^\star_{st}$, $\mu = 30$

e) $\varsigma^\star_{st}$, $\mu = 40$

Figure 5.13: The stationary mutation strength obtained using (5.77) (colored) in comparison to the stationary mutation strength estimate (5.81)(black, dashed) . The curves are similar although some deviations can be observed, especially for small noise strengths and smaller number of parents. Again, the points denote the results of experiments with $(\mu/\mu_I, 60)$-ES with $d = 5$ for $N = 30$ (disks) and $N = 100$ (squares). Each data point was averaged over $400,000$ ($N = 100$) and $200,000$ ($N = 30$) generations in the stationary state.

mutation strength and the distance. Noise would have a positive effect if the increase of the stationary mutation strength outperformed the increase of the stationary distance. However, this is not always the

case. The dependence on the noise strength appears complicated. Comparing the zero noise and the large noise regime, it can be found that noise finally lowers the initial progress rate for $\mu < \lambda/2$, but increases it for $\mu > \lambda/2$. Regardless of the noise strength, evolution strategies with parent populations with less than half the size of the offspring populations have a progress rate that is larger than that of other strategies. Large noise strengths, however, diminish this advantage.

The case of $\mu = 30 = \lambda/2$ is very interesting since the progress rate is not influenced by the noise at all. For all examined choices of $\sigma_\epsilon^\star$, it remains on the same level it had for $\sigma_\epsilon^\star = 0$. Apparently, there is an balance between the influence of distance and of the mutation strength, i.e., $\sigma_{st}^\star(\sigma_\epsilon^\star) = R_{st}(\sigma_\epsilon^\star)$. The question remains why noise increases the progress rate for intermediate ES with $\mu > \lambda/2$, but decreases it for $\mu < \lambda/2$. As (5.78) reveals, the increase of the stationary mutation strength with the noise must stronger than the increase of the distance to the axis to result finally in an increase of (5.78). Apparently, this is the case for $\mu > \lambda/2$. Unfortunately, Equations (5.76) - (5.78) lead to quite complicated solutions which cannot be easily used to answer the question.

However, another interesting behavior is shown in Fig. 5.16. Let us assume a constant noise strength for the moment and consider the stationary mutation strength, distance, and progress rate (5.76) - (5.78) as functions of the parent number μ. Concerning R_{st}, a similar behavior as in the case of the noise sphere occurs: Evolution strategies with $\mu : \lambda$-ratios around ≈ 0.5 have the smallest distances to the ridge. All other strategies are grouped around this value, with increasing distances for $\mu \to \lambda$ and $\mu \to 1$. Furthermore, the distances are approximately symmetric. This is in accordance with the behavior on the noisy sphere [25].

The behavior of the mutation strength (5.77) remains to be addressed. If the noise strength is large, the stationary mutation strength (5.77) first decreases and then increases with μ. Concerning larger values of μ, a similar increase of the mutation strength with μ was already observed on the noisy sphere. There, the non-normalized stationary mutation strength scales approximately with $1/\sqrt{4c_{\mu/\mu,\lambda}}$. As Fig. 5.16 shows (5.77) behaves similarly for large noise strengths.

The progress rate (5.77) depends on the square of ratio of the mutation and the noise strength – weighted additionally with $1/\mu$. Its behavior as a function of μ shows some similarities to the non-noisy case (5.41), p. 91, (see Fig. 5.16 c) and Fig.5.7 b), p. 92). Only in the large-noise regime the influence of the mutation strength is sufficient to lead towards a nearly constant progress rate for a wide range of μ.

Recombination is beneficial in the sense that the induced inherent bias for an increase of the mutation strength serves as a safeguard against a loss of step-size control: Strategies that make use of only one parent cannot stabilize the mutation strength. Once the ES is relatively close to the axis and the influence of the noise is too large, a loss of step-size control can be observed. Since $(1, \lambda)$-ES are prone to a loss of step-size control, recombination appears necessary. The question of how to choose the truncation ratio remains to be answered, however. For large noise strengths, the differences between the performances of different $(\mu/\mu_I, \lambda)$-ES are smoothed out. The experiments showed an even faster convergence of the progress rate towards its limit than predicted. Thus, the case of larger noise strengths appears more important than the case of smaller noise strengths. The parent number μ must be chosen so that the mutation strength and progress rate stabilize. As Figs. 5.12 to 5.15 shows, a parent number of $\mu = 2$ appears to be too small to stabilize the mutation strength sufficiently. Choosing $\mu = 10$, however, is sufficient in our scenario.

The Case of Large Noise Strengths

As said, the solutions of (5.78) are complicated. Therefore, this section aims at deriving simpler approximate solutions. The derivation is primary based on the finding of the previous section: If

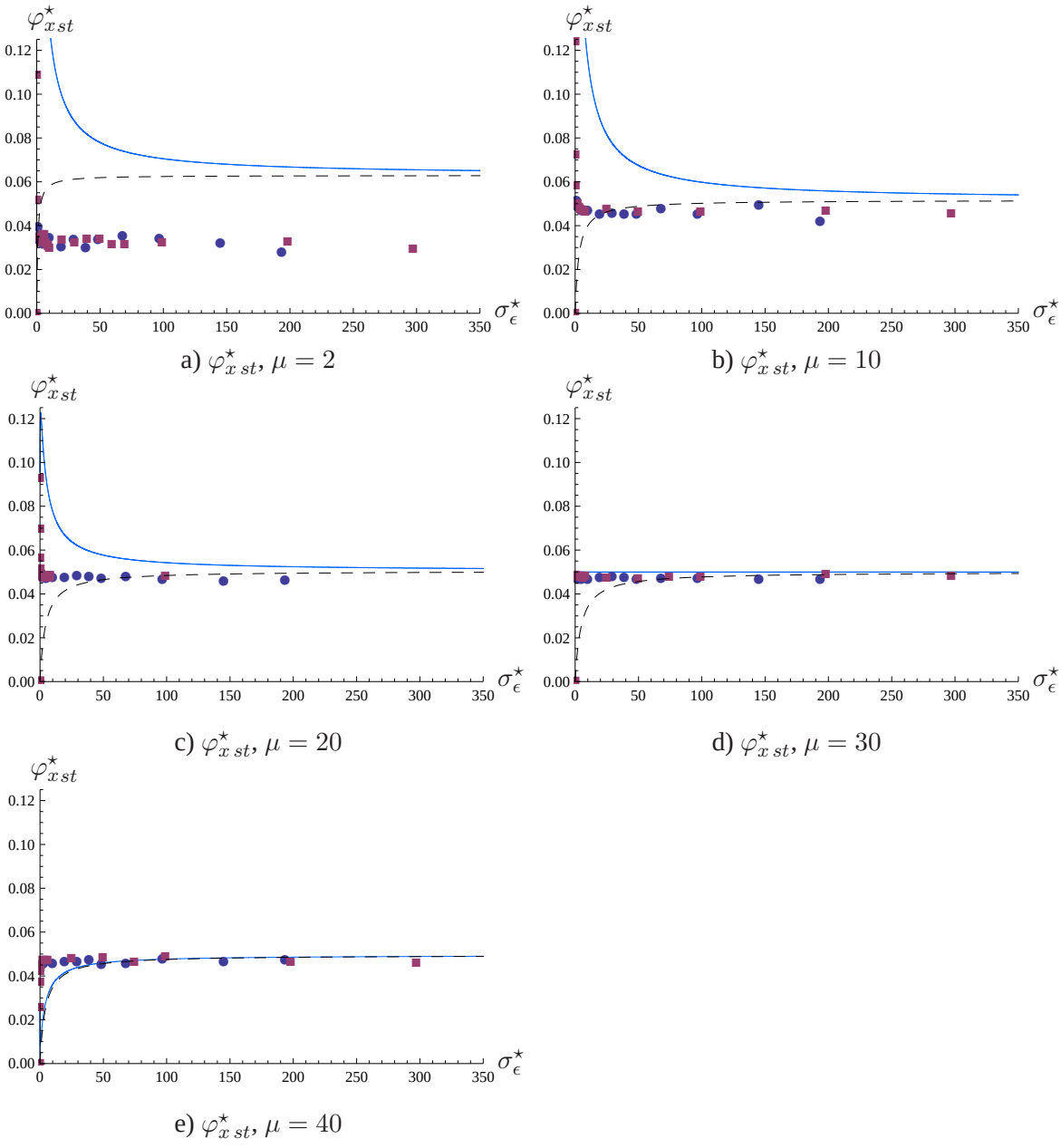


Figure 5.14: The stationary progress rate obtained using (5.79) in comparison to the estimate (5.84) (black dashed line). Again, the points denote the results of experiments with $(\mu/\mu_I, 60)$-ES with $d = 5$ for $N = 30$ (disks) and $N = 100$ (squares). Each data point was averaged over $400,000$ ($N = 100$) and $200,000$ ($N = 30$) generations in the stationary state.

the noise strength is large, a self-adaptive ES on the noisy parabolic shows a similar behavior in the stationary state as on the noisy sphere.

First of all, a minimal distance to the axis can be determined by using (5.74) and setting $\sigma^\star = 0$

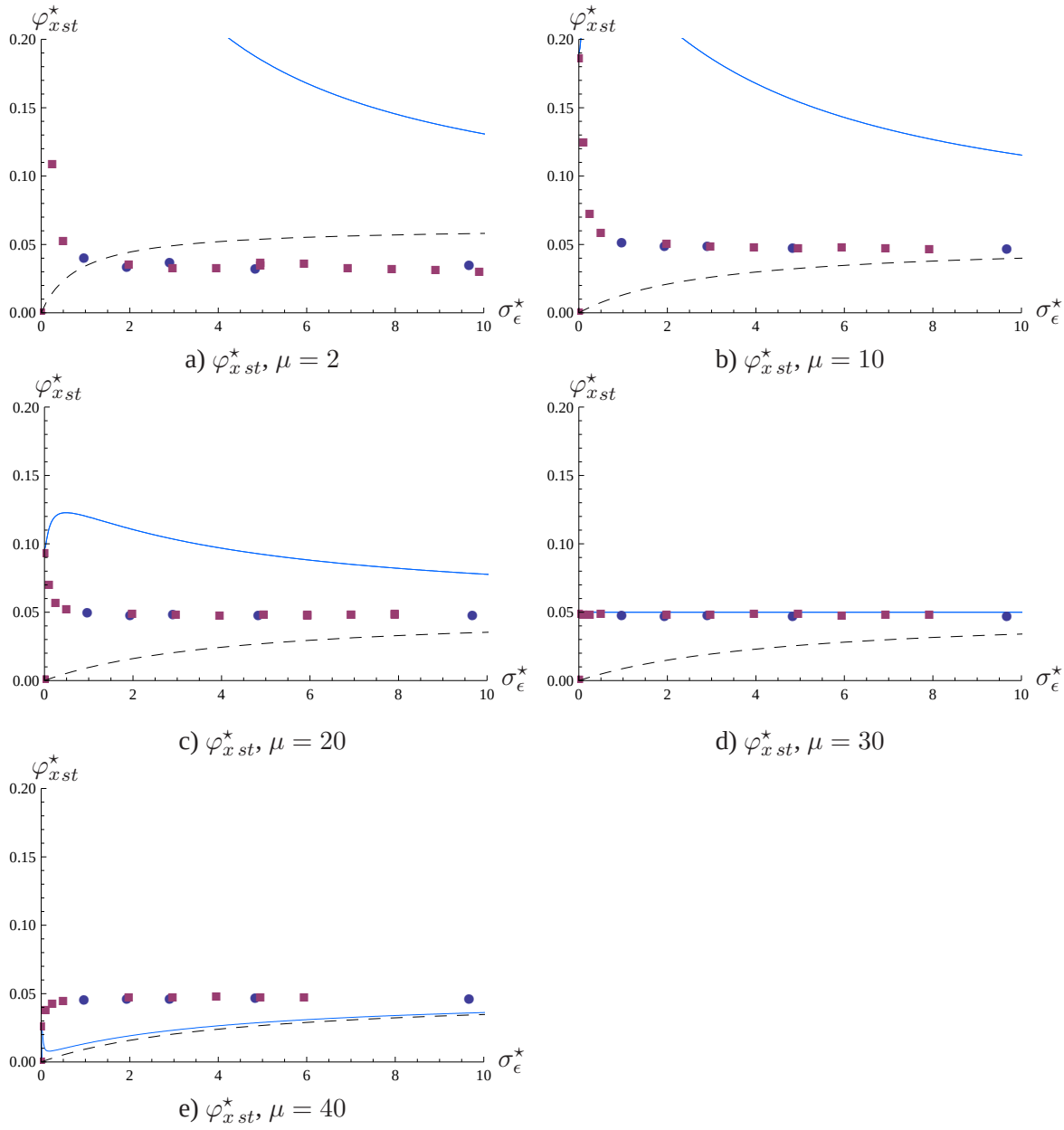


Figure 5.15: The stationary progress rate obtained using (5.79) (dashed line) in comparison to the estimate (5.80) (solid line). For $\mu < \lambda/2$, the stationary progress rate obtained using (5.79) tends to overestimate the experimental results. It should be noted that the experimental results converge far sooner than estimated.

$$R_{\min} = \sqrt{\frac{\sigma_\epsilon N}{4d\mu c_{\mu/\mu,\lambda}}} = \sqrt{\frac{\sigma^\star_\epsilon}{4d\mu c_{\mu/\mu,\lambda}}}. \tag{5.79}$$

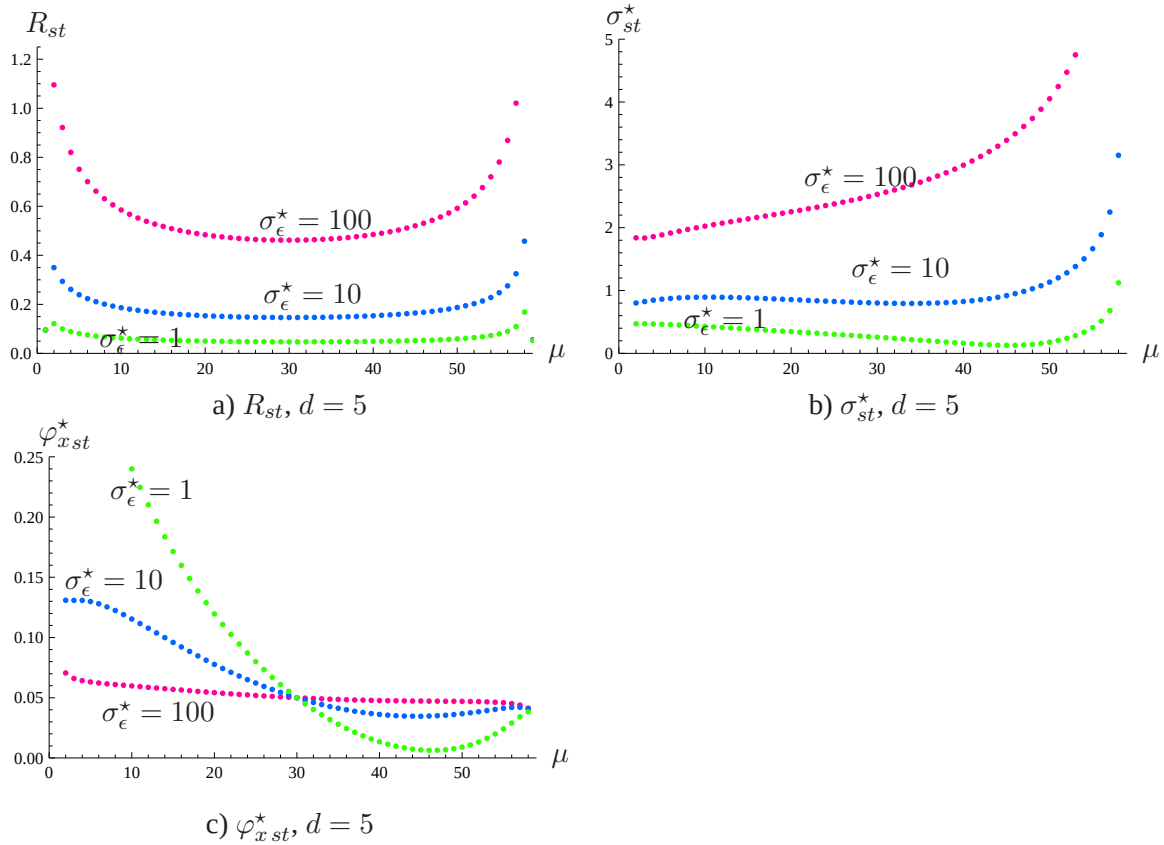


Figure 5.16: The influence of the parent number μ on the stationary distance, mutation strength and progress rate for several noise levels ($\sigma^\star_\epsilon = 1, 10, 100$).

However, (5.79) and (5.76) lead to very similar results at least for the evolution strategies examined. The influence of a non-zero mutation strength on the resulting distance is only minor. Equation (5.79) is applicable only if the noise strength is sufficiently large since it neglects the part of the stationary distance that is due to the evolution of the mutation strength.

But Equation (5.79) points to again to the very interesting characteristic: Concerning the minimal distance, the ES behaves similarly to an ES on the noisy sphere. The minimal distance mirrors the minimal distance (4.66) – apart from the weighting factor $1/d$. This is the basis for the following approach to determine easier estimates for the stationary distance and mutation strength. Since (5.79) is the minimal distance of the noisy sphere (4.66) weighted with $1/\sqrt{d}$, it is assumed that a similar relationship holds for the stationary distances (5.76) and (4.65). This leads to the estimate

$$R_{appr} = \sqrt{\frac{\sigma^\star_\epsilon}{4d\mu c_{\mu/\mu,\lambda}}} \sqrt[4]{\frac{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}}{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1}}. \tag{5.80}$$

As Figs. 5.10 and 5.11 show the deviations between estimate (5.80) and (5.76) are not high. Equation (5.80) can now be used together with (5.77) to obtain an estimate of the stationary mutation strength

$$\varsigma^\star_{appr} = \frac{\sigma^\star_\epsilon}{\sqrt{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1}\sqrt{1 + \frac{d\sigma^\star_\epsilon}{\mu c_{\mu/\mu,\lambda}}\sqrt{\frac{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}}{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1}}}}. \tag{5.81}$$

Again, (5.81) is only applicable if $\sigma_\epsilon^\star$ is large, since (5.81) predicts a zero mutation strength for zero noise strength which is not the case on the parabolic ridge. As Fig. 5.13 shows, the prediction quality of (5.81) is reasonably good: Only for small values of $\sigma_\epsilon^\star$, greater deviations occur as it was to be expected. For greater $\sigma_\epsilon^\star$ the prediction quality improves. It is interesting to note the following findings: First of all, the lines of (5.81) and (5.77) move closer together for increasing noise. Second, a similar effect occurs for an increasing parent number μ. Finally, in the case of smaller parent numbers, (5.81) even serves better as a predictor of the experiments than the results obtained by using (5.76) and (5.77) – provided that the noise is not small.

It is also interesting to note two limit behaviors of (5.81). Provided that $\sigma_\epsilon^\star$ is large, the estimate is approximately

$$\varsigma_{appr2}^\star = \sqrt{\sigma_\epsilon^\star}\frac{\mu c_{\mu/\mu,\lambda}}{\sqrt{4\mu c_{\mu/\mu,\lambda}^2 - 2e_{\mu,\lambda}^{1,1} - 1}\sqrt[4]{d^2\left(\frac{4\mu c_{\mu/\mu,\lambda}^2 - 2e_{\mu,\lambda}^{1,1}}{4\mu c_{\mu/\mu,\lambda}^2 - 2e_{\mu,\lambda}^{1,1} - 1}\right)}} \tag{5.82}$$

that is it scales approximately with $\sqrt{\sigma_\epsilon^\star}$. Provided that $2\mu c_{\mu/\mu,\lambda}^2 \gg e_{\mu,\lambda}^{1,1} + 1$, the estimate (5.81) can be approximated with

$$\varsigma_{appr3}^\star = \frac{\sigma_\epsilon^\star}{\sqrt{4\mu c_{\mu/\mu,\lambda}^2 + 4d c_{\mu/\mu,\lambda}\sigma_\epsilon^\star}}. \tag{5.83}$$

The estimates (5.80) and (5.81) can be used to obtain an estimate for the stationary progress parallel to the axis

$$\varphi_{x\ appr}^\star = \frac{c_{\mu/\mu,\lambda}\sigma_\epsilon^\star}{\sqrt{4\mu c_{\mu/\mu,\lambda}^2 - 2e_{\mu,\lambda}^{1,1} - 1}\sqrt{4\mu c_{\mu/\mu,\lambda}^2 - 2e_{\mu,\lambda}^{1,1}} + \frac{d\sigma_\epsilon^\star}{\mu c_{\mu/\mu,\lambda}}(4\mu c_{\mu/\mu,\lambda}^2 - 2e_{\mu,\lambda}^{1,1})}. \tag{5.84}$$

The prediction quality of (5.84) is only good if the noise strength is large. First of all, it fails to capture the interesting behavior of the progress rate as a function of the noise strength. Instead of showing different responses for different choices of μ, it predicts an increasing progress rate for increasing noise for all strategies considered. Equation (5.84) has a finite limit for $\sigma_\epsilon^\star \to \infty$

$$\lim_{\sigma_\epsilon^\star\to\infty} \varphi_{x\ appr}^\star = \frac{\mu c_{\mu/\mu,\lambda}^2}{d(4\mu c_{\mu/\mu,\lambda}^2 - 2e_{\mu,\lambda}^{1,1})}. \tag{5.85}$$

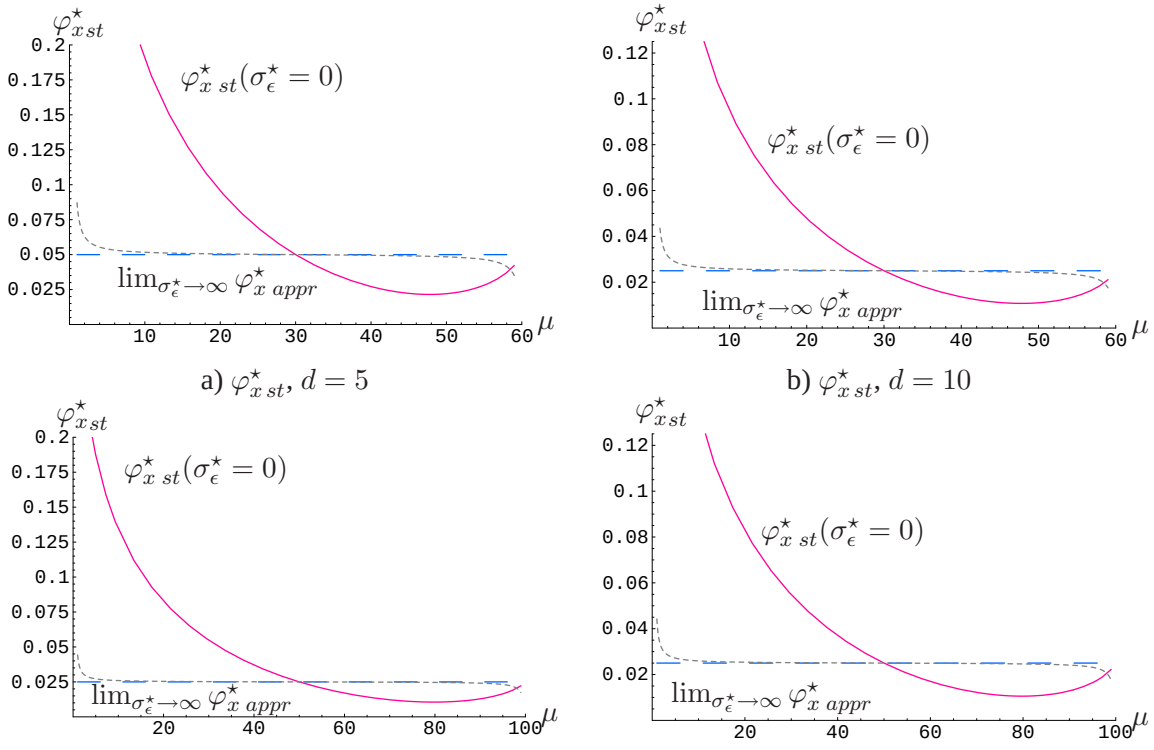
Provided that λ is large, it can be shown that (5.85) leads to nearly the same value $\approx 1/(4d)$ for a wide range of the parameter μ, i.e., as long as $\mu \not\approx 1$ or $\mu \not\approx \lambda$ and λ is relatively large.

It is interesting to compare this limit with the stationary progress rate (5.40)

$$\varphi_{x\ st}^\star = \frac{1/2 + e_{\mu,\lambda}^{1,1}}{2d}$$

obtained for zero noise. Recall, $\varphi_{x\ st}^\star > 1/(4d)$ for $\mu < \lambda/2$, $\varphi_{x\ st}^\star = 1/(4d)$ for $\mu = \lambda/2$, and $\varphi_{x\ st}^\star < 1/(4d)$ for $\mu > \lambda/2$. Figure 5.17 compares (5.40) with (5.85). The expected behavior occurs: If $\mu < \lambda/2$, (5.85) is smaller than (5.40). If μ is larger, (5.85) exceeds (5.40). The crossing point of both progress rates lies at $\mu \approx \lambda/2$. This underlines again the finding that noise does not influence the performance if $\mu \approx \lambda/2$ is chosen. It also shows that in the case of smaller noise strengths, ES with smaller parent numbers are expected to perform superiorly. The problem now consists in finding

a parent number μ that is sufficiently high to stabilize the mutation strength but sufficiently small to have a relatively high progress. As said, the expected changes cannot cover the stochastic behavior of $(1, \lambda)$-ES. Therefore, the approach using the deterministic evolution equations cannot be used to make a recommendation of how μ should be chosen. In the case of $\mu = 2$ and random selection Section 4.2 indicated a bias towards an increase. As the figures show, this bias is not sufficient for a stabilization of the mutation strength on the level needed. Thus, higher parent numbers, i.e., $\mu = 10$ should be used. Of course it is also possible to follow a similar approach as in Section 4.2 and to introduce a small bias towards an increase of the mutation strength.



a) $\varphi^\star_{x\,st}$, $d = 5$ b) $\varphi^\star_{x\,st}$, $d = 10$

c) $\varphi^\star_{x\,st}$, $d = 5$ d) $\varphi^\star_{x\,st}$, $d = 10$

Figure 5.17: Comparison of the limit (5.85) of the estimate (5.84) of the progress rate with the stationary progress rate (5.41) for the undisturbed ridge. The progress rates are shown as functions of μ. The offspring number is set to $\lambda = 60$ (Figs. a), b)) and to $\lambda = 100$ (Figs. c),d)). The value of $1/(4d)$ is indicated by the dotted blue horizontal line.

5.2.3 Self-Adaptation on Ridge Functions: Conclusions

This chapter was devoted to an analysis of self-adaptive evolution strategies on the ridge function class $f(x, R) = x - dR^\alpha$. Two types of ridge functions were considered: the sharp ridge with $\alpha = 1$ and the parabolic ridge with $\alpha = 2$. Section 5.1 was devoted to an analysis of ES on undisturbed ridge functions. In Section 5.2, the analysis was extended to allow to investigate the effects of additive normally distributed noise. All analyses used the deterministic evolution equations (see Chapter 3). Therefore, first of all, the progress measures, the progress rates in $\mathbf{R}$ and x-direction and the self-adaptation response, had to be given. In the following, the main results of the analysis are summarized.

Self-adaptive intermediate ES show very different behaviors on the sharp and parabolic ridge. In

the case of $\alpha = 2$, the ES fluctuates at a stationary distance from the ridge with a positive mutation strength (cf. Subsection 5.1.2). As a result, there is progress in axis direction and no premature convergence occurs. The fitness diverges towards infinity. However, the mutation strength stays constant on average. In the non-noisy case, recombination does not appear to have an advantage. Using the usual $\mu : \lambda$-ratios, i.e., truncation ratios between $[2/\lambda \ldots 0.5]$, the stationary mutation strength is similar to the zero of the SAR obtained for the sphere model. It shows the same response with respect to the change of μ: Recombination lowers the mutation strength. Although the distance to the axis is also reduced, this cannot counteract the resulting effect on the progress rate. Non-recombinative $(1, \lambda)$-ES have the highest progress rate.

It should be noted, though, that the SAR is also responsible for preventing a premature convergence. It strives to maintain a positive mutation strength for decreasing distances which eventually halts any convergence towards the axis. As already pointed out in [19], the case of $\alpha > 1$ closely resembles the situation in the noisy sphere model where the ES is unable to converge to the optimizer and remains on average at a certain distance to the optimizer.

In Section 5.1.2 the sharp ridge was considered. Here, no stationary state with a positive mutation strength exists – unless a normalization with the distance to the axis is introduced. On the sharp ridge, the ES either converges prematurely or enlarges the distance to the axis perpetually. Which behavior occurs depends on the size of the ridge parameter d with respect to the population parameters μ and λ. Recombination lowers this critical d-value. As result, ES with intermediate recombination show a premature convergence for smaller values of d than $(1, \lambda)$-ES.

Provided that the ES does not converge prematurely, it can be shown that the travel speed is not optimal (w.r.t. the quality change). First of all, the optimizer of the quality change cannot be obtained for finite learning rates. Self-adaptation realizes too small mutation strengths. Additionally, there may be problems with recombination. Increasing the learning rate will improve the performance of the ES. Increasing the learning rate, however, causes the stationary mutation strength (normalized w.r.t. to N and R) to behave more and more like the zero of the SAR. The zero of the SAR decreases when switching from $\mu = 1$ to $\mu > 1$. Recombination according to the truncation ratio $\mu : \lambda$ recommended on the sphere is not beneficial. Instead, apparently a fixed value of μ between $\mu = 2$ and $\mu = 5$ appears as good choice.

In Section 5.2, noisy ridge functions were investigated using the standard noise model of additive normally distributed noise. Both ridge function models behave similarly: Additive noise eventually halts the approach to the axis and stops the ES from realizing the subgoal of optimizing the embedded sphere with a finite optimum. Accordingly, in general no premature convergence occurs. Instead, evolution strategies show on average a constant progress parallel to the axis direction. Of course, considering the sharp ridge this only holds if d is sufficiently large so that the axis is approached in first case. If d is too small and the ES diverges, the effects of the noise are soon diluted until it behaves as if it were optimizing the undisturbed ridge. If d is sufficiently large, a stationary state of the distance and the mutation strength exists. In general, the following holds: The larger the noise strength, the larger the stationary distance and the mutation strength. This results in larger progress parallel to the axis direction. Additive noise is beneficial on the sharp ridge: Because of the noise the finite subgoal of optimizing the sphere cannot be realized. Since the better fulfillment of the subgoal is connected with a reduction of the mutation strength, this is an advantage. Recombination has a similar effect as in the case of the sphere model. It reduces the distance to the axis. This decrease with μ is stronger than the increase of the normalized stationary mutation strength (w.r.t. the distance and search space dimensionality). These responses eventually cause a performance degradation: The normalized progress stays constant and the non-normalized progress rate decreases. Recombination on the ridge does have a positive effect, though. The $(1, \lambda)$-ES looses step-size control similar to the

ES on the noise sphere (cf. Section 4).

On the parabolic ridge (cf. Subsection 5.2.2), noise has a similar effect in the sense that the ES stays farer away from the axis and operates with higher mutation strengths. The effect on the non-normalized progress rate is not so clearly defined: The progress rate depends on the distance to the ridge and the mutation strength. Its exact behavior depends on the population parameters μ and λ. In the case of $\mu = \lambda/2$ the distance and the mutation strength balance out: The progress rate is inert to the noise strength. For $\mu < \lambda/2$, the progress rate decreases with the noise strength whereas it increases for $\mu > \lambda/2$. The line defined by $\mu = 30$ is not crossed, though. Interestingly, this line with $\approx 1/(4d)$ serves relatively well as a predictor of the progress rate provided that λ is relatively large.In contrast to the sharp ridge where increasing the noise strength resulted in an increase of the progress rate, ES on the parabolic ridge converge to very similar limits. That is, noise cannot be used to increase or decrease the performance over a certain level. The progress rates of evolution strategies with $\mu < \lambda/2$ do not decrease significantly farther than $1/(4d)$ while the progress rates of ES with $\mu > \lambda/2$ approach $1/(4d)$ from below.

Again, evolution strategies with only one parent suffer a similar loss of step-size control as before. On approach of the axis and therefore on increase of the normalized noise strength, the mutation strength is reduced significantly. Recombination is therefore required to retain a positive mutation strength and progress.

The ES were investigated using the so-called deterministic evolution equations. These difference equations can be used to describe the expected change of the state variables from one generation to the next. The drawback of this approach is of course that the loss of step-size control is not predictable.

The analysis presented here can and should be extended in several points: First of all, the progress measures obtained for the evolution equations hold exactly only for $N \to \infty$. All results obtained using these progress measures hold only approximately in low-dimensional search spaces. Therefore, one aim should be to use progress measures obtained for finite N. Furthermore, the derivation of the SAR should be reconsidered and higher-order terms of the Taylor series development should be included (see Appendix C.1.2). In addition, an inclusion of the perturbation parts of the evolution equations would be interesting. Furthermore, a comparison with other adaptation schemes as the CSA or the $1/5$th rule is of interest.

6 Evolution Strategies and Self-Adaptation

This thesis focuses on the self-adaptation mechanism in evolution strategies. In general, an evolutionary computation is termed self-adaptive if the control of strategy parameters is left to the computation itself. In evolution strategies, self-adaptation is usually applied to the mutation operator, i.e., the mutation strength.

In Chapter 2, an overview of self-adaptation and the present state of research was given. The survey focused on explicit analyses of self-adaptation. Theoretical analyses of self-adaptive ES focus on the stochastic process generated by the evolutionary algorithm. Three main groups can be distinguished – each centering on a distinct aspect of the stochastic process: Markov chains, Martingales, and the dynamic systems approach over the evolution equations. It is interesting to note that no analysis of the mechanism of self-adaptive evolution strategies in continuous search spaces exists that does not resort to either a simplification of the system or Monte Carlo simulations.

Chapter 3 introduced the analysis approach followed in this thesis. The approach was first proposed by Beyer in [21]. In short, the state variables of the ES are described by stochastic difference equations (the evolution equations) decomposed in a deterministic and a perturbation part. The deterministic part can be identified as the expected change of the variable under consideration. The distribution of the remaining fluctuation part is unknown in general. Since it is possible to obtain some of its moments over the evolution equations, the unknown distribution is approximated with a Gram-Charlier series using the normal distribution as baseline. The further approach consists then basically of two steps: In step one, the fluctuation terms are neglected. The aim is to derive the main characteristics of the self-adaptive process. Step two extends the analysis to an inclusion of the fluctuation terms approximated with a normal distribution.

In Chapter 4, the self-adaption behavior of evolution strategies on the sphere was analyzed. In the beginning, the analysis presented in [21] was extended to intermediate $(\mu/\mu_I, \lambda)$-ES. To this end, the deterministic evolution equations were applied. An explanation was given for the experimental findings that intermediate ES show strong dependencies on the correct choice of the learning rate in contrast to $(1, \lambda)$-ES. Furthermore, an optimal learning rate valid for high-dimensional search spaces could be obtained.

In short, recombination in the case of the sphere model has the drawback that the ES is sensitive to the correct choice of the learning rate. This sensitivity can be traced back to the finding that the self-adaptation mechanism can only rely on the fitness. Thus, it cannot make use of the advantages provided by the recombination of the object parameters. Due to the recombination of the object parameters, intermediate ES could operate with higher mutation strengths which is not reflected in the self-adaptation response. On the sphere, the ES reaches a stationary state of the normalized mutation strength (normalized w.r.t. the distance and the search space dimensionality). In other words, the influences of the change of the non-normalized mutation strength (self-adaptation response) and of the change of the distance (progress rate) are balanced. The stationary state depends on the learning rate over the self-adaptation response (SAR). In general there are three decisive mutation strengths which characterize the stationary state: the zero and the optimizer of the progress rate and the zero of the SAR. The ES should strive to work with mutation strengths close to the optimizer. Which

stationary mutation strength the ES stabilizes depends on the learning rate, howver. The learning rate can be used to variate the mutation strength from its minimum to the maximum for $N\tau^2 \to \infty$. If the learning rate is is too small, the ES operates with mutation strengths close to the zero of the progress rate and if it is chosen too large, the stationary mutation strength approaches the zero of the SAR. The latter behavior is not problematic in $(1, \lambda)$-ES: The zero of the SAR and the optimizer of the progress rate are very close together. Choices of τ-values larger than optimal do not lead to a significant performance loss. In multi-recombinative $(\mu/\mu_I, \lambda)$-ES, however, the zero of the SAR is usually far smaller than the optimizer of the progress rate which accounts for the sensitivity.

Regardless of the sensitivity towards the correct choice of τ, self-adaptive ES still perform superiorly compared to $(1, \lambda)$-ES. Furthermore, the learning rate can be chosen so that the ES progresses with optimal speed (w.r.t. the progress rate). Additionally, recombination has positive effects if the fitness function evaluations are overlaid with noise. The $(1, \lambda)$-ES suffers from a loss of step size control if the noise becomes too large. It can be shown that it performs a biased random walk in the large-noise regime. Intermediate $(\mu/\mu_I, \lambda)$-ES still maintain a positive stationary mutation strength. Furthermore, recombination leads to smaller residual location errors. The smallest residual location errors are achieved by $(\mu/\mu_I, \lambda)$-ES with a parent-offspring ratio of $\mu : \lambda = 1/2$. Evolution Strategies with a ratio between 0.2 and 0.7 do not deviate far from this optimum. Therefore, the usual recommendation of choosing $\mu : \lambda \approx 0.27$ can be followed.

Finally, a second-order approach was applied in the case of intermediate ES on the undisturbed ridge. In the second-order approach the influences of the perturbation parts of the evolution equations are not neglected but modeled using a Gaussian distribution. First of all, it was seen that the results obtained do not differ significantly from those obtained using the deterministic approach – if recombination is applied. The equations derived are recursive and highly non-linear and furthermore the stationary state distribution is unknown. Therefore the ansatz introduced in [21] was followed and a log-normal distribution was used to model the unknown steady state distribution. Still, in general, the solutions can be only provided numerically. Only some exemplary cases could be analyzed analytically. For the specific learning rates, it was found that recombination leads to further benefits: The deviations due to perturbations are nearly minimal if the usual $\mu : \lambda$ ratio is chosen – at least for the learning rates considered.

Chapter 5 was devoted to evolution strategies on ridge functions. In the case of the sharp ridge, evolution strategies were found to converge prematurely in some cases. This depends on the size of the ridge function constant d with respect to the population parameters. In short, self-adaptation is torn in a way between two subgoals [79]: reduce the distance to axis or enlarge the gain along the x-axis. Concerning the improvement of the fitness, the ES neither "sees" the position on the x-axis nor the distance to the axis. The feedback is over the overall fitness change. Therefore, the quality change, i.e., the expected fitness change from generation g to $g + 1$, was considered. The optimizer of the quality change scales with the distance to the search space. If the mutation strength is normalized with respect to the distance to the ridge, its evolution equation permits a stationary state. This stationary state is also observable in experiments (see, e.g., [75]) and required if the ES should have a chance to work with nearly optimal mutation strengths with respect to the quality change. Concerning the quality change, i.e., the expected fitness change from generation g to $g + 1$, self-adaptation adjusts the stationary point correctly with respect to changes in d and R. If these parameters are changed, the stationary solution shows the same response as the optimizer of the quality change. It should be noted that in the long run a rewarding of the short-term gain may be problematic. In terms of following the optimizer of the quality change, a stationary state of the mutation strength with respect to the distance is good. But if this is coupled with an reduction of the distance to the axis which is caused by too large d-values, it means that the non-normalized mutation strength decreases and decreases until it is

too small for any significant progress: The ES converges prematurely.

It should be noted that the non-normalized evolution equations do not allow for any stationary state for neither the mutation strength nor the distance – except in the singular case with d exactly the size of the critical d-value. This is reminiscent of the finding of Lunacek and Whitley that the ridge bias in the case of $(1, \lambda)$-ES on the sharp ridge cannot be removed and the mutation strength cannot stabilize [72]. Their subsequent finding that the ES decreases the mutation strength could not be supported in general. This can be probably explained by their experimental set-up which used only d-values greater than one and therefore higher than the critical d-value.

At this point it should be noted that the self-adaptation response is influenced by the distance to the ridge in general. This holds for the sharp ridge as well as for the parabolic ridge. This is in contrast to the response to the linear gain part of the ridge function. This leads leads to a constant value in the SAR. Therefore, the SAR is inert to the position parallel to the axis.

Furthermore, recombination with the usual $\mu : \lambda$-ratio cannot be recommended. It has positive effects for small choices of τ. But it should be noted that the optimizer of the quality change is not attainable for finite τ. Increasing the learning rate turns working with the usual truncation ratio from an advantage into a disadvantage. It can be shown finally for $N\tau^2 \to \infty$, that only very small choices of $\mu > 1$ lead to a higher quality change than $\mu = 1$. This behavior is due to the response of the stationary mutation strength to changes of τ. As long as the learning rate is small, the stationary mutation strength behaves as the zero of the progress rate and increases once recombination is used. Increasing the learning rate drives the mutation strength towards the zero of the SAR which decreases if recombination with the usual $\mu : \lambda$-ratio is introduced. The learning rate increases the quality change far further than working with the best $\mu : \lambda$-ratio and smaller learning rates could. Thus, recombination with the usual truncation ratio is not recommended. It should be mentioned that increasing the learning rate causes a deterioration of the prediction quality. This can be traced back to the derivation of quality change which relied on the assumption that the changes induced by mutation are relatively small. If τ is relatively high, this may cause deviations. Generally speaking, the quality of the results is more sensitive to the choice of the learning rate in the case of the sharp ridge than in the case of the sphere model. But although the prediction quality deteriorates, experiments and prediction show the same response to recombination.

Self-Adaptive evolution strategies do not fail on the parabolic ridge. No premature convergence occurs. The evolutions of the distance to the axis and the mutation strength reach a stationary state. However, the ES still progresses parallel to the axis. The mutation strength is stationary, though, and does not reflect the x-position. Interestingly, recombination does not have positive effects on the performance of the ES. The progress rate decreases for increasing parent numbers.

At first glance this contradicts the results obtained by Oyman [79]. He pointed out that the "better fulfillment of the short-term goal" (here: achieving smaller distances to the ridge) is equivalent to a higher progress rate [79, p. 138]. But Oyman's analysis could not take the response of self-adaptive ES into account.

In the case of self-adaptation, recombination does not only decrease the stationary distance but it also decreases the stationary mutation strength. The decrease of the mutation strength outweighs that of the distance. As result, the progress rate declines. It is interesting to note that for a wide range of $\mu : \lambda$-combinations and large λ, the stationary mutation strength is very similar to the stationary mutation strength which would be obtained for a zero distance. This stationary mutation strength equals the zero point of the normalized SAR of the sphere model weighted with the ridge parameter d. As result, a similar response to changes in the parent number is observed and the evolution of the mutation strength behaves roughly as if the subgoal of optimizing the sphere component were already realized.

If the fitness functions evaluations are overlaid with noise, the results change to some extent. In this thesis, the effects of additive uniform noise were investigated using the normal distribution to model the fluctuations. As it could be observed in experiments, the $(1, \lambda)$-ES looses step-size control on approach of the ridge axis. Again, the deterministic approach cannot predict this behavior.

Recombination is therefore necessary in order to maintain a positive mutation strength and to ensure the possibility of further progress. As said before, recombination introduces a bias towards an increase of the mutation strength which serves as a safeguard.

In the case of the sharp ridge, noise has a positive influence. Considering the normalized system (w.r.t. N and R), the situation is analogous to the noisy sphere model. This leads to a result which surprises at first glance: Additive noise with a constant noise strength improves the performance of the ES. The stationary progress rate scales linearly with the noise strength. One reason for this is that the noise keeps the ES from realizing the finite subgoal. Regarding the task of optimizing the sphere part of the ridge, noise still deteriorates the performance: The larger the noise strength, the greater the location error to the axis. But for the overall goal, this is an advantage since all stationary variables scale with the noise strength and the distance to the axis. This also means that recombination with the usual truncation ratio of $\mu : \lambda = 0.27$ or similar values should not be used. As on the sphere, ES with $\mu : \lambda = 1/2$ show the smallest stationary distances to the axis and ES with truncation ratios between $1/3$ and $2/3$ come close. Concerning the progress rate (non-normalized), this decrease of the distance is too strong to be counterbalanced: Evolution strategies with these or similar truncation ratios do not show large stationary progress rates. However, recombination is necessary to prevent a loss of step-size control.

In the case of the parabolic ridge, noise increases the stationary distance to the axis and the mutation strength. Both variables influence the progress rate. Concerning the effects of the noise on the performance, three situations occur: Noise degrades the performance if the parental population is smaller than $1/2$ of the offspring population. If exactly half of the parents are used, changing the noise strength does not have any effects at all. If more parents are utilized, noise improves the performance. However, the progress rates of ES with truncation ratios smaller than $1/2$ are larger than those of ES with $\mu : \lambda > 1/2$. This advantage diminishes if the noise strength increases.

Concerning self-adaptation, two effects that may cause problems were identified. First of all, self-adaptation, i.e., the self-adaptation response, can only use aggregated information over the fitness values. In the case of the sphere model, it cannot make use of the genetic repair effect which is induced by the recombination of the object parameters.

Second, in the case of ridge functions the performance of self-adaptive depends strongly on the distance to the axis. This is often coupled with a deterioration of the performance: The better the ES succeeds in optimizing the sphere part, the more the performance decreases. Recombination with $\mu \leq \lambda/2$ generally improves this optimization result. Accordingly, recombination with the usual truncation ratio often causes a decline of the performance.

It should be noted that this behavior is in pronounced contrast to the behavior of ES using cumulative step-size adaptation [9]. As shown in [9] for the parabolic ridge, a CSA-ES achieves a progress rate of $\varphi_x = \mu c^2_{\mu/\mu,\lambda}/(2d)$ for zero noise strength. Working with the usual truncation ratio improves the performance. Furthermore, the stationary distance, $R = N/(2d)$, does not depend on μ.

The deterministic evolution equations can be used to analyze the main characteristics of the steady-state dynamics. The drawback is of course the non-capturing of the irregular dynamics of the process. As seen, the loss of step-size control $(1, \lambda)$-ES on noisy fitness functions could not be predicted. This clearly indicates a limit of the approach and requires switching to higher-order approximations. Of course, the analysis can be extended in various points. It remains to include

the N-dependent progress measures in the analysis. This concerns especially the self-adaptation response for the sharp ridge. And furthermore, other noise models, for instance actuator noise, should be investigated.

A Results from Probability Theory and Statistics

In this chapter, some results from probability and statistics are provided that are central to the analysis of self-adaptive evolution strategies using the evolution equations.

A.1 Random Variables and Distributions

First some basic definitions are given before the concepts of moments and cumulants are introduced. Afterwards, some distributions appearing often in the area of evolution strategies are described.

A.1.1 Random Variables

Let us consider a sample space Ω, i.e., the set of all possible outcomes of experiments or events ω. A random variable X is then defined as a (measurable) real-valued function on the sample space $X : \Omega \to \mathbb{R}$. The distribution function F_X defined by

$$F_X(t) \quad = \quad Pr(\{X \leq t\}) \tag{A.1}$$

is also called the *cumulative distribution function* (cdf). It is easy to see that it is a monotonously increasing and right-continuous function with $F_X(t) \to 0$ for $t \to -\infty$ and $F_X(t) \to 1$ for $t \to \infty$. If F is differentiable, its derivative p is called the *probability density function* (pdf) or shortly density function and

$$F_X(t) \quad = \quad \int_{-\infty}^{t} p(x)\,\mathrm{d}x \tag{A.2}$$

holds. The expectation of a random variable is defined by

$$\mathrm{E}[X] \quad = \quad \int_{-\infty}^{\infty} xp(x)\,\mathrm{d}x \tag{A.3}$$

whereas the variance is given by

$$\mathrm{Var}[X] \quad = \quad \mathrm{E}\Big[(X - \mathrm{E}[X])^2\Big]. \tag{A.4}$$

Expectation and variance are the special cases of the so-called moments.

A.1.2 Moments and Cumulants

Let X be a random variable with pdf $p(X)$. The kth (raw) moment is given by

$$\mu_k \quad = \quad \int_{-\infty}^{\infty} x^k p(x)\,\mathrm{d}x. \tag{A.5}$$

The central moments are taken around the mean $\mu := \mu_1$

$$m_k \quad = \quad \int_{-\infty}^{\infty} (x-\mu)^k p(x)\,\mathrm{d}x. \tag{A.6}$$

Since

$$(x-\mu)^k = (-1)^k \sum_{l=0}^{k} \binom{k}{l} (-1)^l x^l \mu^{l-k} \tag{A.7}$$

the central moments can be expressed as functions of the raw moments

$$m_k = (-1)^k \sum_{l=0}^{k} \binom{k}{l} (-1)^l \mu_l \mu^{l-k}. \tag{A.8}$$

An opposite result holds in turn of course. Note, moments do not exist for every continuous distribution. A well known example is the Cauchy-distribution with density

$$p(x) = \frac{1}{\pi a} \frac{1}{1+(x/a)^2} \tag{A.9}$$

and parameter $a > 0$ which does not have any finite moment. Moments can be defined in yet another way. The so-called moment generating function is defined by

$$\xi(t) = \int_{-\infty}^{\infty} \mathrm{e}^{tx} p(x)\,\mathrm{d}x. \tag{A.10}$$

It is easy to see that the kth raw moment is given by $\mu_k = \mathrm{d}^k/(\mathrm{d}t^k)\psi(t)|_{t=0}$. The moment generating function is similar to the characteristic function or Fourier transform of the distribution given by

$$\zeta(t) = \int_{-\infty}^{\infty} \mathrm{e}^{itx} p(x)\,\mathrm{d}x. \tag{A.11}$$

The natural logarithm of the moment-generating function is called the cumulant generating function

$$\Xi(t) = \ln(\xi(t)). \tag{A.12}$$

Similarly to the moments, the cumulant of kth order is obtained as $\kappa_k = \mathrm{d}^k/(\mathrm{d}t^k)\Xi(t)|_{t=0}$ with $\kappa_0 = 0$.

A.1.3 Distributions

In this subsection, an overview over some distributions is given which appear often in the context of evolution strategies.

Normal Distribution

The normal distribution or Gaussian is one of the most important distributions in statistics. Partly, it owns its importance to the fact the sum of random variables converges to a normally distributed random variable under relatively mild conditions. The probability density function reads

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}. \tag{A.13}$$

It depends on two parameters: the mean μ and the standard deviation σ and is a symmetric function around μ. The cumulative density function is given by

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}\sigma} \mathrm{e}^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mathrm{d}x. \tag{A.14}$$

Log-Normal Distribution

A random variable X is called log-normally distributed if its logarithm is normally distributed. The pdf

$$p(x) = \frac{1}{2x\tau\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{\ln(x)-\mu}{\tau}\right)^2} \tag{A.15}$$

is only defined for $x > 0$. The moments of the log-normal distribution are given by

$$\mu_k = \mathrm{e}^{k\mu+\frac{k^2\tau^2}{2}}. \tag{A.16}$$

χ^2-Distribution A random variable with density

$$p(x) = \begin{cases} 0, & \text{if } x \leq 0 \\ \frac{1}{2^{\frac{k}{2}}\Gamma(k/2)} x^{\frac{k-2}{2}} \mathrm{e}^{-\frac{x}{2}} & \text{if } x > 0 \end{cases} \tag{A.17}$$

is called χ^2-distributed with k degrees of freedom or χ^2_k-distributed. The Γ-function is given for $k > 0$ by

$$\Gamma(k) = \int_0^\infty y^{k-1}\mathrm{e}^{-y}\,\mathrm{d}y \tag{A.18}$$

The first two moments of the χ^2_k-distribution read $\mathrm{E}[\chi^2_k] = k$ and $\mathrm{Var}[\chi^2_k] = 2k$. The χ^2-distribution is connected with the normal distribution over the following theorem.

Theorem 1. *Let $Z_1, \ldots, Z_k$ be k standard normally distributed random variables. Then the sum of the squares $Y = Z_1^2 + \ldots + Z_k^2$ is χ^2_k-distributed.* □

The square of a single standard normally distributed variable is χ^2_1-distributed. In the case of two summands, the χ^2_2-distribution equals an exponential distribution $\gamma_{\lambda,1}(x) = \lambda\mathrm{e}^{-\lambda x}$ with parameter $\lambda = 1/2$ (see, e.g.,[80]).

A.2 Order Statistics

The presentation in this section follows [80]. Let $X_1, \ldots, X_\lambda$ denote λ random variables. For all $\omega \in \Omega$ let $X_{m:\lambda}(\omega)$ denote the mth smallest value of $X_1(\omega), \ldots, X_\lambda(\omega)$, i.e.,

$$X_{1:\lambda}(\omega) \leq X_{2:\lambda}(\omega) \leq \ldots \leq X_{\lambda:\lambda}(\omega). \tag{A.19}$$

The random variables $X_{m:\lambda}, \ldots, X_{\lambda:\lambda}$ are called order statistics with $X_{m:\lambda}$ giving the mth order statistic. Provided that all X_i are independent and identically distributed with cdf $P(x)$, the cdf of these random variables is given by

$$P_{m:\lambda}(x) = \sum_{k=m}^{\lambda} \binom{\lambda}{k} P(x)^k \Big(1 - P(x)\Big)^{\lambda-k}. \tag{A.20}$$

This can be seen easily. The proof presented is taken from [80]. Let the random variables $Y_m(x)$ denote $1_{\{X_m \leq x\}}$ and let $Y(x) := \sum_{m=1}^{\lambda} Y_m(x)$. Since $Pr(Y_m(x) = 1) = P(x)$, Y is $\mathcal{B}(\lambda, P(x))$-distributed which leads to (A.20) using

$$\{X_{m:\lambda} \leq x\} = \{m \leq Y(x) \leq \lambda\}. \tag{A.21}$$

The density function can be obtained via differentiation of (A.20) leading to

$$\begin{aligned}
p_{m:\lambda}(x) = \frac{\mathrm{d}}{\mathrm{d}x}P_{m:\lambda}(x) = & \sum_{k=m}^{\lambda} k\binom{\lambda}{k}P(x)^{k-1}\Big(1-P(x)\Big)^{\lambda-k}p(x) \\
& - \sum_{k=m}^{\lambda}(\lambda-k)\binom{\lambda}{k}P(x)^{k}\Big(1-P(x)\Big)^{\lambda-k-1}p(x) \\
= & \lambda\sum_{k=m}^{\lambda}\binom{\lambda-1}{k-1}P(x)^{k-1}\Big(1-P(x)\Big)^{\lambda-k}p(x) \\
& - \lambda\sum_{k=m}^{\lambda}\binom{\lambda-1}{k}P(x)^{k}\Big(1-P(x)\Big)^{\lambda-k-1}p(x) \\
= & \lambda\sum_{k=m-1}^{\lambda-1}\binom{\lambda-1}{k}P(x)^{k}\Big(1-P(x)\Big)^{\lambda-k-1}p(x) \\
& - \lambda\sum_{k=m}^{\lambda}\binom{\lambda-1}{k}P(x)^{k}\Big(1-P(x)\Big)^{\lambda-k-1}p(x) \\
= & \lambda\binom{\lambda-1}{m-1}P(x)^{m-1}\Big(1-P(x)\Big)^{\lambda-m}p(x) \\
& - \lambda\binom{\lambda-1}{\lambda}P(x)^{\lambda}\Big(1-P(x)\Big)^{-1}p(x) \\
= & \lambda\binom{\lambda-1}{m-1}P(x)^{m-1}\Big(1-P(x)\Big)^{\lambda-m}p(x). \qquad \text{(A.22)}
\end{aligned}$$

In the analyses, the density of the mth best offspring is required. The realization $X_{m;\lambda}(\omega)$ thus often denotes not the mth smallest but the mth highest outcome. In this case, note that the mth highest value out of λ trials is also the $(\lambda-m+1)$th smallest. The density is therefore given by

$$\begin{aligned}
p_{m;\lambda}(x) = p_{\lambda-m+1:\lambda}(x) &= \lambda\binom{\lambda-1}{\lambda-m+1-1}P(x)^{\lambda-m+1-1}\Big(1-P(x)\Big)^{\lambda-\lambda+m-1}p(x) \\
&= \lambda\binom{\lambda-1}{\lambda-m}P(x)^{\lambda-m}\Big(1-P(x)\Big)^{m-1}p(x) \\
&= \lambda\binom{\lambda-1}{m-1}P(x)^{\lambda-m}\Big(1-P(x)\Big)^{m-1}p(x). \qquad \text{(A.23)}
\end{aligned}$$

A.3 Generalized Progress Coefficients

The generalized progress coefficients are given by

$$e_{\mu,\lambda}^{\alpha,\beta} = \frac{\lambda-\mu}{\sqrt{2\pi}^{\alpha+1}}\binom{\lambda}{\mu}\int_0^\infty t^{\beta}\mathrm{e}^{-\frac{\alpha+1}{2}t^2}\Phi(t)^{\lambda-\mu-1}\Big(1-\Phi(t)\Big)^{\mu-\alpha}\,\mathrm{d}t \qquad \text{(A.24)}$$

(see [23, p. 172]). The special case $c_{\mu/\mu,\lambda} := e_{\mu,\lambda}^{1,0}$

$$c_{\mu/\mu,\lambda} = \frac{\lambda-\mu}{2\pi}\binom{\lambda}{\mu}\int_0^\infty \mathrm{e}^{-t^2}\Phi(t)^{\lambda-\mu-1}\Big(1-\Phi(t)\Big)^{\mu-1}\,\mathrm{d}t \qquad \text{(A.25)}$$

gives the expectation of the mean of the μ best of λ trials of standard normally distributed random variables.

B The Progress Rates

This chapter describes how the progress rates for the sphere model and the ridge function class can be obtained. The progress rate is a central performance measure. In the case of the sphere model, it gives the expected one-generational change of the distance to the optimizer. In the case of ridge functions, two progress rates appear since two variables are used to describe the evolution of the object variables. These are the distance to the ridge axis and the position on the axis. The progress rates are needed in Chapter 4-5 to describe the evolution of the ES. First, the progress rate for the sphere model is derived following the derivation in [6]. Afterwards, the so-called second-order progress rate is computed. The second-order progress denotes the expectation of the square of the change of the distance and is needed in the analysis if the fluctuation terms are not neglected (c.f. Sec. 4.4). Finally, the progress rates are computed for the ridge function class. First, a density function for the quality change induced by a mutation is obtained. Using this result, the progress rates for the distance to the axis and the position on the axis can be obtained. The density obtained in B.2.1 will also be used in the calculation of the self-adaptation response (SAR) in the case of ridge functions in Appendix C.

B.1 The Sphere Model

This section gives the derivation of the first- and second-order progress rate. Before these can be obtained, fitness change of an offspring must be determined. This is done in the first subsection B.1.1. Subsection B.1.2 sketches the derivation of the first-order progress rate, B.1.3 describes how the second-order progress rate can be obtained.

B.1.1 The Fitness Change of an Offspring

In this section, the fitness change due to a mutation is obtained for the sphere model. Since the fitness function f is the sphere, it is given by $f(\mathbf{y}) = g(\|\mathbf{y} - \hat{\mathbf{y}}\|)$ with $\hat{\mathbf{y}}$ the optimizer of f. The function g is a monotonously increasing or decreasing function. Without loss of generality, this subsection considers a minimization problem, i.e., g increases with the distance to the optimizer R. One of the simplest members of this function class is the quadratic sphere $g(R) = R^2$. If $\mathbf{z}$ denotes a mutation vector, the associate fitness change $Q(\mathbf{z})$ is given by

$$Q(\mathbf{z}) \quad = \quad F(\langle \mathbf{y} \rangle) - F(\langle \mathbf{y} \rangle + \mathbf{z}) = g(R) - g(r) \tag{B.1}$$

where R denotes the distance of the centroid to the optimizer whereas r stands for the distance of the mutation vector. In the general case, some approximations have to be made during the derivations. Many of these are not necessary for the quadratic sphere as will be shown later on. Provided that g is a C^{K+1}-function, the Taylor expansion to the order of K reads

$$T_g(r) \quad = \quad \sum_{k=0}^{K} \frac{\mathrm{d}^k}{\mathrm{d}r^k} g(r)|_{r=R} \frac{(r-R)^k}{k!} + \mathcal{O}((r-R)^{K+1}). \tag{B.2}$$

In the following, $r - R \ll 1$ is assumed. This allows to cut off the expansion of g after the linear term and to neglect quadratic and higher contributions

$$g(r) \;=\; g(R) + \frac{\mathrm{d}}{\mathrm{d}r} g(r)|_{r=R}(r-R) + \mathcal{O}((r-R)^2). \tag{B.3}$$

The fitness change can thus be approximated with

$$Q(\mathbf{z}) \;=\; -\frac{\mathrm{d}}{\mathrm{d}r} g(r)|_{r=R}(r-R) + \mathcal{O}((r-R)^2). \tag{B.4}$$

In following, the notation is shortened to $g'(R) := (\,\mathrm{d}/\,\mathrm{d}r) g(r)|_{r=R}$. The change of the distances must be addressed in the next step. To this end, the usual decomposition of the mutation vector is used: Each mutation vector $\mathbf{z}$ can be given as the sum of two vectors, the first, $\mathbf{z}_A$, parallel to $\mathbf{R}$ – the second, $\mathbf{z}_B$, perpendicular to $\mathbf{z}_A$. Since $r = \|\langle \mathbf{y}^{(g)} \rangle + \mathbf{z} - \hat{\mathbf{y}}\|$, we have

$$\begin{aligned} r^2 &= (R - z_A)^2 + \|\mathbf{z}_B\|^2 = R^2 - 2Rz_A + z_A^2 + \|\mathbf{z}_B\|^2 \\ &= R^2\Big(1 - \frac{2}{R} z_A + \frac{z_A^2}{R^2} + \frac{\|\mathbf{z}_B\|^2}{R^2}\Big). \end{aligned} \tag{B.5}$$

The decomposition of the mutation vector is used to obtain the difference of the distances

$$\begin{aligned} r - R &= R\sqrt{1 - \frac{2}{R} z_A + \frac{z_A^2}{R^2} + \frac{\|\mathbf{z}_B\|^2}{R^2}} - R \\ &= R\sqrt{1 - \frac{2}{R}(z_A - \frac{z_A^2}{2R} - \frac{\|\mathbf{z}_B\|^2}{2R})} - R \\ &\approx R\Big(1 - \frac{2}{R}(z_A - \frac{z_A^2}{2R} - \frac{\|\mathbf{z}_B\|^2}{2R})\Big) - R \\ &= -z_A + \frac{z_A^2}{2R} + \frac{\|\mathbf{z}_B\|^2}{2R} \end{aligned} \tag{B.6}$$

using a Taylor series expansion of the root $\sqrt{1-2x}$ and taking only the linear term. Due to the isotropy of mutations, $\mathbf{z_A}$ can be assumed to be $\sigma z_1 \mathbf{e}_1$, with $\mathbf{e}_1$ the first unit vector and z_1 a standard normally distributed random variable. The vector $\mathbf{z_B}$ consists of the remaining $N-1$ components, each also normally distributed. Assuming that the contribution of z_1^2 can be neglected, Equation (B.6) changes to

$$\begin{aligned} r - R &\approx -\sigma z_1 + \frac{\sigma^2}{2R} \sum_{i=2}^{N} z_i^2 \\ &= -\sigma z_1 + \frac{\sigma^2}{2R} \sum_{i=2}^{N} z_i^2 \end{aligned} \tag{B.7}$$

The sum is a χ^2_{N-1}-distributed random variable. As it was shown in [6] using the Central Limit Theorem, it is possible to model the sum $\sigma^2/(2R) \sum_{i=2}^{N} z_i^2$ by a normally distributed random variable with mean $(N-1)\sigma^2$ and variance $2(N-1)\sigma^4$ if N is large. Large values of N also allow to identify $N-1$ with N. The difference can therefore be approximated with

$$r - R \;\approx\; -\sigma z_1 + \frac{N\sigma^2}{2R} + \frac{\sqrt{2N}}{2R}\sigma^2 u \tag{B.8}$$

where u is a standard normally distributed random variable. This leads to

$$Q(\mathbf{z}) \approx g'(R)\sigma z_1 - \frac{g'(R)N\sigma^2}{2R} - \frac{g'(R)\sqrt{2N}}{2R}\sigma^2 u. \tag{B.9}$$

If the fitness evaluations are disturbed by additive noise, the selection is not based on the values of Q but on

$$\begin{aligned}\tilde{Q}(\mathbf{z}) &= Q(\mathbf{z}) + \epsilon \\ &\approx g'(R)\sigma z_1 - \frac{g'(R)N\sigma^2}{2R} - \frac{g'(R)\sqrt{2N}}{2R}\sigma^2 u + \epsilon.\end{aligned} \tag{B.10}$$

In this thesis, only normally distributed noise with mean zero and standard deviation σ_ϵ is considered. This results in

$$\tilde{Q}(\mathbf{z}) \approx g'(R)\sigma z_1 - \frac{g'(R)N\sigma^2}{2R} - \frac{g'(R)\sqrt{2N}}{2R}\sigma^2 u + \sigma_\epsilon z_\epsilon \tag{B.11}$$

with z_ϵ standard normally distributed. Equation (B.11) leads to the cumulative distribution function (cdf)

$$P(\tilde{q}) = \Phi\Big(\frac{\tilde{q} + \frac{g'(R)N\sigma^2}{2R}}{\sqrt{(g'(R)\sigma)^2 + \left(\frac{g'(R)\sqrt{2N}}{2R}\sigma^2\right)^2 + \sigma_\epsilon^2}}\Big) \tag{B.12}$$

and the probability density function (pdf)

$$p(\tilde{q}) = \frac{\exp\left(-\frac{1}{2}\Big(\frac{\tilde{q}+\frac{g'(R)N\sigma^2}{2R}}{\sqrt{\left(g'(R)\sigma\right)^2+\left(\frac{g'(R)\sqrt{2N}}{2R}\sigma^2\right)^2+\sigma_\epsilon^2}}\Big)^2\right)}{\sqrt{2\pi}\sqrt{(g'(R)\sigma)^2 + \left(\frac{g'(R)\sqrt{2N}}{2R}\sigma^2\right)^2 + \sigma_\epsilon^2}}. \tag{B.13}$$

Introducing the usual normalizations [6], $\tilde{Q}^* := \tilde{Q}[N/(Rg'(R))]$, $\sigma^* := \sigma(N/R)$, and $\sigma_\epsilon^*{:=}\sigma_\epsilon[N/(Rg'(R))]$, the normalized fitness change is

$$\tilde{Q}^* = \sigma^* z_1 - \frac{\sigma^{*2}}{\sqrt{2N}}u - \sigma_\epsilon^* z_\epsilon - \frac{\sigma^{*2}}{2}. \tag{B.14}$$

Equation (B.14) can be used to give the cumulative distribution function

$$P(\tilde{q}^*|\sigma^*) = \Phi\left(\frac{\tilde{q}^* + \frac{\sigma^{*2}}{2}}{\sqrt{\sigma^{*2}(1+\frac{\sigma^{*2}}{2N}) + \sigma_\epsilon^{*2}}}\right) \tag{B.15}$$

and the probability density function

$$p(\tilde{q}^*|\sigma^*) = \frac{\mathrm{e}^{-\frac{1}{2}\left(\frac{\tilde{q}^*+\frac{\sigma^{*2}}{2}}{\sqrt{\sigma^{*2}(1+\frac{\sigma^{*2}}{2N})+\sigma_\epsilon^{*2}}}\right)^2}}{\sqrt{2\pi}\sqrt{\sigma^{*2}(1+\frac{\sigma^{*2}}{2N}) + \sigma_\epsilon^{*2}}}. \tag{B.16}$$

In the case of the quadratic sphere, neither the Taylor series expansion of the function g nor the expansion of the root are necessary. The corresponding values can be obtained directly. The starting point is (B.5) which can be inserted directly into (B.1)

$$\begin{aligned} Q(\mathbf{z}) &= g(R) - g(r) = R^2 - r^2 = R^2 - R^2 + 2Rz_A - z_A^2 - \|\mathbf{z}_B\|^2 \\ &= 2R\sigma z_1 - N\sigma^2 - \sqrt{2N}\sigma^2 u \\ &= g'(R)\sigma z_1 - \frac{Ng'(R)}{2R}\sigma^2 - \frac{\sqrt{2N}}{2R}\sigma^2 u \end{aligned} \tag{B.17}$$

since $g'(R) = 2R$. The main prerequisite in the case of the quadratic sphere is a high-dimensional search space so that the χ^2_N-distributed sum in (B.5) can be approximated by a normally distributed random variable.

B.1.2 The First-Order Progress Rate

Let $\mathbf{R}^{(g)} := \langle \mathbf{y}^{(g)} \rangle - \hat{\mathbf{y}}$ and let $R := R^{(g)} = \|\mathbf{R}^{(g)}\|$ denote the distance of the centroid of the parental population to the optimizer in generation g. The notation of the mean of the μ mutation strengths $\langle \varsigma^{(g)} \rangle$ will be shortened in the following to σ in order to simplify the equations.

The progress rate is defined as the expected one-generation change of the distance

$$\varphi_R = \mathrm{E}[R - R^{(g+1)} | (R, \sigma)]. \tag{B.18}$$

and was already obtained in [6] for $\tau = 0$. This progress rate will be used in analysis. Although the progress rate has been found to depend on the learning parameter τ [51], this approach is justified by the observation that generally $\tau \propto 1/\sqrt{N}$ is chosen as a rule of thumb. Since the analysis is restricted to high-dimensional search spaces, this should allow to use the result obtained in [6].

The derivation of the progress rate relies on an appropriate decomposition of the mutation vectors which was described in the previous subsection. It is possible to decompose the centroid of the mutation vectors $\langle \mathbf{z} \rangle$ in a similar manner. Let $\langle z_A \rangle$ denote the part of the centroid pointing towards the optimizer and let $\langle \mathbf{z}_B \rangle$ denote the perpendicular components. Thus, based on

$$\begin{aligned} \varphi_R &= \mathrm{E}[R - R^{(g+1)} | \sigma, R] \\ &= R\mathrm{E}\Big[1 - \sqrt{\Big(1 - \frac{\langle z_A \rangle}{R}\Big)^2 + \frac{\|\langle \mathbf{z}_B \rangle\|^2}{R^2}} | \sigma, R\Big] \end{aligned} \tag{B.19}$$

or on the normalized equation

$$\varphi^*_R = N\mathrm{E}\Big[1 - \sqrt{\Big(1 - \frac{\langle z_A \rangle}{R}\Big)^2 + \frac{\|\langle \mathbf{z}_B \rangle\|^2}{R^2}} | \sigma^*\Big] \tag{B.20}$$

an approximate formula for φ^* can be derived by calculating the expectation of $\langle z_A \rangle$ and $\|\langle \mathbf{z}_B \rangle\|^2$ and assuming that the expectation of the square root may be approximated by these expectations. Thus, an estimate for the progress rate was obtained in [6] as

$$\varphi^*_R(\sigma^{(g)}) = N\left[1 - \sqrt{1 - \frac{2\sigma^{(g)}}{R^{(g)}}\overline{\langle z_1 \rangle} + \frac{\overline{\|\langle \mathbf{z_B} \rangle\|^2}}{(R^{(g)})^2}}\right] \tag{B.21}$$

leading to

$$\varphi^*_R(\sigma^{*(g)}) = N\left[1 - \sqrt{1 + \frac{{\sigma^{*(g)}}^2}{\mu N}}\right.$$

$$\times \sqrt{1 - \frac{2}{N}\left(\frac{1+\frac{\sigma^{*(g)^2}}{2\mu N}}{1+\frac{\sigma^{*(g)^2}}{\mu N}}\right)\frac{\sigma^{*(g)^2} c_{\mu/\mu,\lambda}}{\sqrt{\sigma_\epsilon^* + \sigma^{*(g)^2}(1+\frac{\sigma^{*(g)^2}}{2N})}}}\Bigg]. \tag{B.22}$$

Linearizing the second square root finally gives

$$\varphi_R^* = \frac{c_{\mu/\mu,\lambda}\sigma^*(1+\frac{\sigma^{*2}}{2\mu N})}{\sqrt{1+\frac{\sigma^{*2}}{\mu N}}\sqrt{1+\frac{\sigma_\epsilon^{*2}}{\sigma^{*2}}}} - N\Big[\sqrt{1+\frac{\sigma^{*2}}{\mu N}} - 1\Big]. \tag{B.23}$$

Considering $N \to \infty$ leads to the N-independent progress rate

$$\varphi_R^* = \frac{c_{\mu/\mu,\lambda}\sigma^{*2}}{\sqrt{\sigma^{*2}+\sigma_\epsilon^{*2}}} - \frac{\sigma^{*2}}{2\mu} \tag{B.24}$$

that will be used in Chapter 4. The coefficient $c_{\mu/\mu,\lambda}$ denotes a special case of the generalized progress coefficients $e_{\mu,\lambda}^{\alpha,\beta}$ ($(c_{\mu/\mu,\lambda} := e_{\mu,\lambda}^{1,0}$), see Eq. (A.25), p. 126. Equation (B.24) was derived under several assumptions. The first was to neglect the influence of the learning parameter τ. The justification for this lies in the observation that τ is generally chosen to be proportional to $1/\sqrt{N}$. In high-dimensional search spaces, the influence of the learning rate τ on the progress rate is small enough not to be considered. The second assumption is made in (B.21) where the progress rate is given as the progress of the expectation of $\langle z \rangle$. This equals assuming that the fluctuations need not be taken into account and has consequences for the second order progress rate. Under the assumption above it follows $\varphi^{*(2)} = \varphi^{*2}$ as it is shown in Section B.1.3. As a result, the variance $D_{\varphi^*}^2 = \varphi^{*(2)} - \varphi^{*2}$ is zero permitting only a first order approximation of the r-evolution equation.

B.1.3 The Second-Order Progress Rate

The second-order progress rate is needed for the evolution equations (4.74) if the second-order approximation is used and the stochastic parts are modeled using normally distributed random variables. As before in the case of the first order progress rate, only the case of $\tau = 0$ is considered. Under this restriction, the derivation of the second-order progress rate is very straightforward. As mentioned earlier, the second-order progress rate is actually a function of the first-order progress rate (B.24). Let us start with the definition

$$\begin{aligned}
\varphi^{(2)}(\langle\varsigma^{(g)}\rangle, R^{(g)}) &= \mathrm{E}\left[(R^{(g)} - R^{(g+1)})^2 | \langle\varsigma^{(g)}\rangle, R^{(g)}\right] \\
&= \mathrm{E}\left[(R^{(g)})^2 - 2R^{(g)}R^{(g+1)} + (R^{(g+1)})^2 | \langle\varsigma^{(g)}\rangle, R^{(g)}\right] \\
&= 2R^{(g)}\mathrm{E}\left[(R^{(g)} - R^{(g+1)} | \langle\varsigma^{(g)}\rangle, R^{(g)}\right] - (R^{(g)})^2 \\
&\quad + \mathrm{E}\left[(R^{(g+1)})^2 | \langle\varsigma^{(g)}\rangle, R^{(g)}\right].
\end{aligned} \tag{B.25}$$

Considering the definition of the first-order progress rate (B.18), (B.25) leads to

$$\varphi^{(2)}(\langle\varsigma^{(g)}\rangle, R^{(g)}) = 2R^{(g)}\varphi(\langle\varsigma^{(g)}\rangle, R^{(g)}) - (R^{(g)})^2 + \mathrm{E}\left[(R^{(g+1)})^2\right]. \tag{B.26}$$

If the normalization $\varphi^{*(2)} := (N/R^{(g)})^2\varphi^{(2)}$ is used, we obtain

$$\varphi^{*(2)} = N^2\mathrm{E}\left[\left(1 - \frac{R^{(g+1)}}{R^{(g)}}\right)^2\right]$$

$$
\begin{aligned}
&= N^2\mathrm{E}\left[2\left(1-\frac{R^{(g+1)}}{R^{(g)}}\right)-1\right]+N^2E\left[\left(\frac{R^{(g+1)}}{R^{(g)}}\right)^2\right]\\
&= N\varphi^* - N^2 + \frac{N^2}{(R^{(g)})^2}\mathrm{E}\left[(R^{(g+1)})^2\right]. \qquad \text{(B.27)}
\end{aligned}
$$

As mentioned before, if the estimate (B.21) is used for the first-order progress rate, Equation (B.27) leads to φ^{*2}, which can be shown by calculating the square of (B.21)

$$
\begin{aligned}
\varphi^{*2} &= N^2\left(1-2\sqrt{1-\frac{2\langle\varsigma^{(g)}\rangle}{R^{(g)}}\overline{\langle z_1\rangle}+\frac{\overline{\|\langle\mathbf{z_B}\rangle\|^2}}{(R^{(g)})^2}}+1-\frac{2\langle\varsigma^{(g)}\rangle}{R^{(g)}}\overline{\langle z_1\rangle}+\frac{\overline{\|\langle\mathbf{z_B}\rangle\|^2}}{(R^{(g)})^2}\right)\\
&= N^2\left(2-2\sqrt{1-\frac{2\langle\varsigma^{(g)}\rangle}{R^{(g)}}\overline{\langle z_1\rangle}+\frac{\overline{\|\langle\mathbf{z_B}\rangle\|^2}}{(R^{(g)})^2}}-1+1-\frac{2\langle\varsigma^{(g)}\rangle}{R^{(g)}}\overline{\langle z_1\rangle}+\frac{\overline{\|\langle\mathbf{z_B}\rangle\|^2}}{(R^{(g)})^2}\right)\\
&= 2N\varphi^* + N^2\left(-1+1-\frac{2\langle\varsigma^{(g)}\rangle}{R^{(g)}}\overline{\langle z_1\rangle}+\frac{\overline{\|\langle\mathbf{z_B}\rangle\|^2}}{(R^{(g)})^2}\right)\\
&= 2N\varphi^* + N^2\left(-1+\frac{1}{(R^{(g)})^2}\mathrm{E}[(R^{(g+1)})^2]\right) = \varphi^{*(2)}. \qquad \text{(B.28)}
\end{aligned}
$$

B.2 Ridge Functions

The section is devoted to the determination of the progress rates φ_R measuring the progress towards the axis and φ_x giving the progress parallel to the axis. The derivation makes use of a result obtained in [8]. But first of all, the density function of an offspring is needed.

B.2.1 The Fitness Change of an Offspring

Let us consider the fitness change of an offspring l based on the centroid $\langle\mathbf{y}\rangle$ of the parent population

$$
\begin{aligned}
Q := F(\mathbf{y}^l) - F(\langle\mathbf{y}\rangle) &= y_1^l - \langle y_1\rangle - d(r^\alpha - R^\alpha)\\
&=: z_x - d(r^\alpha - R^\alpha) \qquad \text{(B.29)}
\end{aligned}
$$

where $z_x := y_1^l - \langle y_1\rangle$ denotes the change in the first component of the vector, whereas $R := (\sum_{k=2}^N(\langle y_k\rangle)^2)^{1/2}$ denotes the centroid's distance to the ridge and $r := (\sum_{k=2}^N(y_k^l)^2)^{1/2}$ gives the distance of the offspring. In order to derive the cumulative density function (cdf) and probability density function (pdf) of an offspring several steps are needed:

1. Note, the rotated ridge function is used, i.e., $f(\mathbf{y}) = y_1 - d(\sum_{i=2}^N y_i^2)^{\alpha/2} =: x - dR^\alpha$. Thus, $z_x = x - \langle x\rangle$ is the change of the first component of the object vector and obeys a $\mathcal{N}(0,\sigma)$-distribution.

2. The change $r - R$ is small. Under this assumption consider the Taylor series expansion of $f(r) = r^\alpha$ around R, $T_f(r) = R^\alpha + \alpha R^{\alpha-1}(r-R) + \mathcal{O}((r-R)^2)$. Provided that the contributions of the quadratic (and higher) terms can be neglected, the fitness change simplifies to $Q = x - \langle x\rangle - d\alpha R^{\alpha-1}(r-R) + \mathcal{O}((r-R)^2)$. Note the assumption above is only necessary to treat the case of general α. In the case of $\alpha = 1$, there is no quadratic term. In the case of $\alpha = 2$, it is possible to treat r^2 directly by the usual decomposition (see below). So, in the case of the sharp and the parabolic ridge the assumption is not required.

3. Consider the $(N-1)$-dimensional system $(y_2,\ldots,y_N)^{\mathrm{T}}$. An offspring is created by adding a mutation vector $\mathbf{z}$ to the parental vector $\mathbf{R}$, i.e., $\mathbf{r} = \mathbf{R} + \mathbf{z}$. Switching to a coordinate system with origin in $\mathbf{R}$, we can decompose $\mathbf{z}$ into two parts $-z_R\mathbf{e}_R + \mathbf{h}$ with $\mathbf{e}_R := \mathbf{R}/R$ and $\mathbf{h}$ perpendicular to $\mathbf{R}$. This decomposition is similar to the decomposition in the case of the sphere model [23]. Therefore, the $\mathbf{r}$-vector can re-written as $\mathbf{r} = \mathbf{R} - z_R\mathbf{e}_R + \mathbf{h}$ and its length as $r = \|\mathbf{r}\| = \sqrt{(\mathbf{R} - z_R\mathbf{e}_R)^2 + h^2} = \sqrt{R^2 - 2Rz_r + z_r^2 + h^2}$.

4. The distributions of the components of the $\mathbf{r}$-vector remains to be addressed. Due to the isotropy of the mutations used, the component z_R will be assumed to be the second component of the object vector $\mathbf{y}$. It is therefore $\mathcal{N}(0,\sigma)$-distributed. Its square is χ_1^2-distributed. The remaining sum $h^2 = \sum_{i=3}^{N} y_i^2$ consists of the squares of $N-2$ normally distributed random variables and is χ_{N-2}^2-distributed. A χ_{N-2}^2-distribution may be modeled using a normal distribution provided that N is large (see Appendix B.1.1). Considering large N allows additional to substitute $N-2$ with N. Accordingly, it is assumed in the following that h^2 is $\mathcal{N}(N\sigma^2, \sqrt{2N}\sigma^2)$ distributed.

5. Consider the square root

$$f(z_R, h_R) = \sqrt{(R - z_R)^2 + h^2} = \sqrt{R^2 - 2Rz_R + z_R^2 + h^2} \tag{B.30}$$

which can be rewritten as

$$\begin{aligned} f(z_R, h_R) &= R\sqrt{1 - \frac{2}{R}z_R + \frac{z_R^2}{R^2} + \frac{h^2}{R^2}} \\ &= R\sqrt{1 - 2\Big(\frac{z_R}{R} - \frac{z_R^2}{2R^2} - \frac{h^2}{2R^2}\Big)}. \end{aligned} \tag{B.31}$$

Provided that $z_R \ll R$, $h \ll R$ hold, the root can be expanded into a Taylor series around zero and cut off after the very first term giving $f(z_R, h_R) = R(1 - z_R/R + z_R^2/(2R^2) + h^2/(2R^2))$. Provided that $z_R^2/(2R^2) \ll 1$, the term may be neglected.

6. Let us treat the case of $\alpha = 2$ separately. Here, we have $r^2 = (R - z_R)^2 + h^2 = 2R^2(1 - z_R/R + h^2/(2R^2))$. Neither, the smallness assumption of $r - R$ in 2. nor the assumptions in 5., $z_R \ll R$ and $h \ll R$, are required at this point.

As already pointed out in [19] the resulting fitness change

$$\begin{aligned} Q &= z_x - d\alpha R^{\alpha-1}\Big(R(1 - \frac{z_R}{R} + \frac{h^2}{2R^2}) - R\Big) \\ &= z_x + d\alpha R^{\alpha-1}\Big(z_R - \frac{h^2}{2R}\Big) \end{aligned} \tag{B.32}$$

is very similar to that of a noisy sphere with z_x in the role of the noise term. The cumulative density function (cdf) and the probability density function (pdf) of an offspring can now be easily given as

$$P_Q(Q) = \Phi\Bigg(\frac{Q + q\frac{N}{2R}\sigma^2}{\sqrt{\sigma^2(1+q^2) + q^2\frac{N}{2R^2}\sigma^4}}\Bigg) \tag{B.33}$$

and

$$p_Q(Q) = \frac{\exp\Big(-\frac{1}{2}\Big(\frac{Q+q\frac{N}{2R}\sigma^2}{\sqrt{\sigma^2(1+q^2)+q^2\frac{N}{2R^2}\sigma^2}}\Big)^2\Big)}{\sqrt{2\pi}\sqrt{\sigma^2(1+q^2) + q^2\frac{N}{2R^2}\sigma^4}} \tag{B.34}$$

with

$$q \;:=\; d\alpha R^{\alpha-1}. \tag{B.35}$$

Introducing the normalizations $Q^\star = QN$ and $\sigma^\star = \sigma N$, the pdf and cdf change to

$$\begin{aligned} P_Q(Q^\star) &= \Phi\Bigg(N\frac{\frac{Q^\star}{N} + q\frac{\sigma^{\star 2}}{2RN}}{\sqrt{\sigma^{\star 2}(1+q^2) + q^2\frac{\sigma^{\star 4}}{2N^2R^2}}}\Bigg) \\ &= \Phi\Bigg(\frac{Q^\star + q\frac{\sigma^{\star 2}}{2R}}{\sqrt{\sigma^{\star 2}(1+q^2) + q^2\frac{\sigma^{\star 4}}{2N^2R^2}}}\Bigg) \end{aligned} \tag{B.36}$$

and

$$p_Q(Q^\star) \;=\; N\frac{\exp\Bigg(-\frac{1}{2}\Big(\frac{Q^\star + q\frac{\sigma^{\star 2}}{2R}}{\sqrt{\sigma^{\star 2}(1+q^2) + q^2\frac{\sigma^{\star 4}}{2N^2R^2}}}\Big)^2\Bigg)}{\sqrt{2\pi}\sqrt{\sigma^{\star 2}(1+q^2) + q^2\frac{\sigma^{\star 2}}{2N^2R^2}}} \tag{B.37}$$

The expression $p(Q)\,\mathrm{d}Q$ is equal to $(1/N)\;p(Q^\star)\,\mathrm{d}Q^\star$. For $N \to \infty$, some components in (B.37) stemming from the distance's perpendicular part vanish leading to

$$P_Q(Q^\star) \;=\; \Phi\Big(\frac{Q^\star + q\frac{\sigma^{\star 2}}{2R}}{\sqrt{\sigma^{\star 2}(1+q^2)}}\Big) \tag{B.38}$$

and

$$p^\star_Q(Q^\star) \;=\; \frac{\exp\Bigg(-\frac{1}{2}\Big(\frac{Q^\star + q\frac{\sigma^{\star 2}}{2R}}{\sqrt{\sigma^{\star 2}(1+q^2)}}\Big)^2\Bigg)}{\sqrt{2\pi}\sqrt{\sigma^{\star 2}(1+q^2)}} \tag{B.39}$$

which will be used in the determination of the SAR in C.1.2. Equations (B.38) and (B.39) can be easily adapted to the case of noisy fitness evaluations. Using the standard model of additive normally distributed noise z_ϵ with zero mean and standard deviation σ_ϵ, it is easy to see that the fitness change of an offspring (B.32) changes to

$$Q \;=\; z_x + d\alpha R^{\alpha-1}\Big(z_R - \frac{h^2}{2R^2}\Big) + z_\epsilon. \tag{B.40}$$

The cdf and pdf of Q are obtained as

$$P_Q(Q) \;=\; \Phi\Bigg(\frac{Q + q\frac{N}{2R}\sigma^2}{\sqrt{\sigma^2(1+q^2) + q^2\frac{N}{2R^2}\sigma^4 + \sigma_\epsilon^2}}\Bigg) \tag{B.41}$$

and

$$p_Q(Q) \;=\; \frac{1}{\sqrt{2\pi}\sqrt{\sigma^2(1+q^2) + q^2\frac{N}{2R^2}\sigma^4 + \sigma_\epsilon^2}}\mathrm{e}^{-\frac{1}{2}\left(\frac{Q + q\frac{N}{2R}\sigma^2}{\sqrt{\sigma^2(1+q^2) + q^2\frac{N}{2R^2}\sigma^4 + \sigma_\epsilon^2}}\right)^2}. \tag{B.42}$$

B.2.2 The Progress Rates

The following lemma is taken directly from [8, p.6]:

Lemma 1. Let $Y_1, Y_2, \ldots, Y_\lambda$ be λ independent standard normally distributed random variables and let $Z_1, Z_2, \ldots, Z_\lambda$ be λ independent normally distributed random variables with zero mean and variance θ^2. Then, defining $X_l = Y_l + Z_l$ for $l = 1, \ldots, \lambda$ and ordering the sample members by nondecreasing values of the X variates, the expected value of the arithmetic mean of those μ of the Y_l with the largest associated values of X_l is

$$\overline{\langle Y \rangle} = \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+\theta^2}}. \tag{B.43}$$

The progress coefficient in (B.43) is given by Eq. (A.25), p. 126. Lemma 1 can be used to determine the progress rates. Note, the same decomposition as in Appendix B.2.1 applies: The fitness change of an offspring is given by

$$Q = z_x + q z_R - \frac{q}{2R} h^2 + z_\epsilon \tag{B.44}$$

with $q := d\alpha R^{\alpha-1}$ (B.35). The random variables z_x and z_R are normally distributed with mean zero and standard deviation σ. Similarly, the random variable h^2 may be assumed to be normally distributed with mean $N\sigma^2$ and standard deviation $\sqrt{2N}\sigma^2$ if N is large. The noise terms z_ϵ also follows a normal distribution with zero mean and standard deviation σ_ϵ. In the following, we will switch to standard normally distributed random variables u_*:

$$Q = \sigma u_x + q\sigma u_R + \sigma_\epsilon u_\epsilon - \frac{q}{2R}\sqrt{2N}\sigma^2 u_{h^2} - \frac{q}{2R} N\sigma^2. \tag{B.45}$$

Let us start with the axial progress

$$\varphi_x = \mathrm{E}[\langle x^{(g+1)} \rangle - \langle x^{(g)} \rangle] = \mathrm{E}[\langle z_x \rangle] = \sigma \mathrm{E}[\langle u_x \rangle]. \tag{B.46}$$

The expectation can be determined using Lemma 1. Note, the addend $[q/(2R)]N\sigma^2$ in (B.45) does not influence the selection since it is the same for all offspring. The corresponding normally distributed variables Z_l of Lemma 1 are defined by

$$Z_l = \frac{q}{\sigma}\sigma u_R + \frac{\sigma_\epsilon}{\sigma} u_\epsilon - \frac{q}{2\sigma R}\sqrt{2N}\sigma^2 u_{h^2} = \sqrt{q^2\left(1 + \frac{N}{2R^2}\sigma^2\right) + \frac{\sigma_\epsilon^2}{\sigma^2}}\,\mathcal{N}_l(0,1) \tag{B.47}$$

where $\mathcal{N}_l(0,1)$ denotes a standard normally distributed random variable. Note, the sum of two normally distributed random variables is again a normally distributed random variable. Therefore, Lemma 1 gives

$$\varphi_x = \frac{c_{\mu/\mu,\lambda}\sigma^2}{\sqrt{\sigma^2(1+q^2) + q^2\frac{N}{2R^2}\sigma^4 + \sigma_\epsilon^2}}. \tag{B.48}$$

Introducing the normalizations $\varphi_x^\star := N\varphi_x$, $\sigma^\star := N\sigma$, and $\sigma_\epsilon^\star := N\sigma_\epsilon$, (B.48) changes to

$$\varphi_x^\star = \frac{c_{\mu/\mu,\lambda}\sigma^{\star 2}}{\sqrt{\sigma^{\star 2}(1+q^2) + \frac{q^2}{2R^2 N}\sigma^{\star 2} + \sigma_\epsilon^{\star 2}}}. \tag{B.49}$$

Letting $N \to \infty$ leads to the progress rate

$$\varphi_x^\star = \frac{c_{\mu/\mu,\lambda}\sigma^{\star 2}}{\sqrt{\sigma^{\star 2}(1+q^2)+\sigma_\epsilon^{\star 2}}} \tag{B.50}$$

which will be used in the calculations in Chapter 5.

The progress (not normalized and normalized) towards the axis is defined as

$$\begin{aligned}\varphi_R &:= \mathrm{E}[R-r] = R\mathrm{E}\Big[1-\sqrt{\Big(1-\frac{\langle z_R\rangle}{R}\Big)^2+\frac{\langle \mathbf{h}\rangle^2}{R^2}}\Big] \\ \varphi_R^\star &:= N\mathrm{E}[R-r] = RN\mathrm{E}\Big[1-\sqrt{\Big(1-\frac{\langle z_R\rangle}{R}\Big)^2+\frac{\langle \mathbf{h}\rangle^2}{R^2}}\Big].\end{aligned} \tag{B.51}$$

To continue, we use the results obtained in [23] and [8]:

1. It was shown in [23, p. 209] that

$$\varphi_R^\star = NR\left(1-\sqrt{\Big(1-\frac{\overline{\langle z_R\rangle}}{R}\Big)^2+\frac{\overline{\langle \mathbf{h}\rangle^2}}{R^2}}\right)+\mathcal{O}\Big(\frac{1}{\sqrt{N}}\Big). \tag{B.52}$$

2. To determine the expectation of the central component Lemma 1 can be used. The determination is completely analogous to the determination of $\mathrm{E}[\langle z_x\rangle]$. Only the roles of z_R and z_x are reversed

$$\overline{\langle z_R\rangle} = \frac{c_{\mu/\mu,\lambda}q\sigma^2}{\sqrt{\sigma^2(1+q^2)+q^2\frac{N}{2R^2}\sigma^4+\sigma_\epsilon^2}} \tag{B.53}$$

$$= \frac{c_{\mu/\mu,\lambda}q\sigma^{\star 2}}{N\sqrt{\sigma^{\star 2}(1+q^2)+q^2\frac{N}{2R^2}\sigma^{\star 4}+\sigma_\epsilon^{\star 2}}}. \tag{B.54}$$

3. In the case of the lateral component, the expectation over the square of the sum of μ vectors must be taken. Since the random vectors $\mathbf{h}_{m;\lambda}$ are independent [23], $\mathrm{E}[\mathbf{h}_{m;\lambda}^{\mathrm{T}}\mathbf{h}_{l;\lambda}] = 0$ holds for $m \neq l$. The expectation

$$\overline{\langle \mathbf{h}\rangle^2} = \frac{\overline{\langle \mathbf{h}^2\rangle}}{\mu} \tag{B.55}$$

remains. Remember, the random variable h^2 of each offspring is also a normally distributed random variable with mean $N\sigma^2$ and standard deviation $\sqrt{2N}\sigma^2$

$$\frac{\overline{\langle \mathbf{h}^2\rangle}}{\mu} = \frac{\sqrt{2N}}{\mu}\sigma^2\overline{\langle u_{h^2}\rangle}+\frac{N}{\mu}\sigma^2. \tag{B.56}$$

Let us now consider $\overline{\langle u_{h^2}\rangle}$. Using (B.45), the corresponding Z_l of Lemma 1 read

$$\begin{aligned}Z_l &= \frac{\sigma}{\frac{q}{2R}\sqrt{2N}\sigma^2}u_x+\frac{q\sigma}{\frac{q}{2R}\sqrt{2N}\sigma^2}u_z+\frac{\sigma_\epsilon}{\frac{q}{2R}\sqrt{2N}\sigma^2}u_\epsilon \\ &= \frac{\sqrt{\sigma^2(1+q^2)+\sigma_\epsilon^2}}{\sqrt{2N}\sigma^2\frac{q}{2R}}\mathcal{N}_l(0,1).\end{aligned} \tag{B.57}$$

Taking note of the sign in (B.45), this leads to

$$\overline{\langle u_{h^2}\rangle} = -\frac{\frac{q}{2R}\sqrt{2N}c_{\mu/\mu,\lambda}\sigma^2}{\sqrt{\sigma^2(1+q^2)+q^2\frac{N}{2R^2}\sigma^4+\sigma_\epsilon^2}}. \tag{B.58}$$

By plugging (B.58) into (B.56),

$$\frac{\overline{\langle \mathbf{h}^2\rangle}}{\mu} = -\frac{c_{\mu/\mu,\lambda}q\frac{N}{R}\sigma^4}{\mu\sqrt{\sigma^2(1+q^2)+q^2\frac{N}{2R^2}\sigma^4+\sigma_\epsilon^2}}+\frac{N}{\mu}\sigma^2 \tag{B.59}$$

is obtained. Introducing again the normalizations $\sigma^\star := N\sigma$ and $\sigma_\epsilon^\star := N\sigma_\epsilon$, (B.59) changes to

$$\frac{\overline{\langle \mathbf{h}^2\rangle}}{\mu} = -\frac{c_{\mu/\mu,\lambda}\frac{q}{RN^2}\sigma^{\star 4}}{\mu\sqrt{\sigma^{\star 2}(1+q^2)+q^2\frac{\sigma^{\star 4}}{2R^2N}+\sigma_\epsilon^{\star 2}}}+\frac{\sigma^{\star 2}}{\mu N} \tag{B.60}$$

and (B.54) becomes

$$\overline{\langle z_R\rangle} = \frac{c_{\mu/\mu,\lambda}q\frac{\sigma^{\star 2}}{N}}{\sqrt{\sigma^{\star 2}(1+q^2)+\frac{q^2}{2NR^2}\sigma^{\star 4}+\sigma_\epsilon^{\star 2}}}. \tag{B.61}$$

The results (B.60) and (B.61) are then inserted into the lateral progress rate (B.52).

4. Using Taylor series expansions (see [23, p.215]) for (B.52) and the resulting expressions, it can be shown that for $N\to\infty$

$$\varphi_R^\star = \frac{q\sigma^{\star 2}}{\sqrt{\sigma^{\star 2}(1+q^2)+\sigma_\epsilon^{\star 2}}}c_{\mu/\mu,\lambda}-\frac{\sigma^{\star 2}}{2R\mu} \tag{B.62}$$

with $q = d\alpha R^{\alpha-1}$ (B.35) is obtained. The calculations are straightforward. Inserting (B.60) and (B.61) into (B.52) leads to the following argument of the root

$$\begin{aligned}\Big(1-\frac{\overline{\langle z_R\rangle}}{R}\Big)^2+\frac{\overline{\langle \mathbf{h}^2\rangle}}{\mu R^2} &= \Big(1-\frac{qc_{\mu/\mu,\lambda}\sigma^{\star 2}}{RN\sqrt{\sigma^{\star 2}(1+q^2)+q^2\frac{\sigma^{\star 4}}{2N}+\sigma_\epsilon^{\star 2}}}\Big)^2\\ &\quad -\frac{qc_{\mu/\mu,\lambda}\sigma^{\star 4}}{N^2R^2\mu\sqrt{\sigma^{\star 2}(1+q^2)+q^2\frac{\sigma^{\star 4}}{2N}+\sigma_\epsilon^{\star 2}}}+\frac{\sigma^{\star 2}}{\mu NR^2}.\end{aligned}$$

Performing the multiplication and reordering the result into an expression of the form $1-2x$ gives

$$\begin{aligned}\Big(1-\frac{\overline{\langle z_R\rangle}}{R}\Big)^2+\frac{\overline{\langle \mathbf{h}^2\rangle}}{\mu R^2} &= 1-2\frac{qc_{\mu/\mu,\lambda}\sigma^{\star 2}}{RN\sqrt{\sigma^{\star 2}(1+q^2)+q^2\frac{\sigma^{\star 4}}{2N}+\sigma_\epsilon^{\star 2}}}\\ &\quad +\frac{q^2c_{\mu/\mu,\lambda}^2\sigma^{\star 4}}{R^2N^2\Big(\sigma^{\star 2}(1+q^2)+q^2\frac{\sigma^{\star 4}}{2N}+\sigma_\epsilon^{\star 2}\Big)}\end{aligned}$$

$$
\begin{aligned}
& -\frac{qc_{\mu/\mu,\lambda}\sigma^{\star 4}}{N^2R^2\mu\sqrt{\sigma^{\star 2}(1+q^2)+q^2\frac{\sigma^{\star 4}}{2N}+\sigma_\epsilon^{\star 2}}}+\frac{\sigma^{\star 2}}{\mu NR^2} \\
= \; & 1-2\Bigg(\frac{qc_{\mu/\mu,\lambda}\sigma^{\star 2}}{RN\sqrt{\sigma^{\star 2}(1+q^2)+q^2\frac{\sigma^{\star 4}}{2N}+\sigma_\epsilon^{\star 2}}} \\
& -\frac{q^2c_{\mu/\mu,\lambda}^2\sigma^{\star 4}}{2R^2N^2\left(\sigma^{\star 2}(1+q^2)+q^2\frac{\sigma^{\star 4}}{2N}+\sigma_\epsilon^{\star 2}\right)} \\
& +\frac{qc_{\mu/\mu,\lambda}\sigma^{\star 4}}{2N^2R^2\mu\sqrt{\sigma^{\star 2}(1+q^2)+q^2\frac{\sigma^{\star 4}}{2N}+\sigma_\epsilon^{\star 2}}}-\frac{\sigma^{\star 2}}{2\mu NR^2}\Bigg)
\end{aligned}
\tag{B.63}
$$

The next step consists of expanding $\sqrt{1-2x}$ into its Taylor series around zero and taking only the linear term in x. Thus, the approximation is only valid for small values of x. Regarding (B.63), $N \ll 1$ must hold and the resulting error term is of order $1/N$. Thus, the order of the previous error term $1/\sqrt{N}$ still applies. The first derivative of $f(x)=\sqrt{1-2x}$ is given by $f'(x)=-1/\sqrt{1-2x}$. The progress rate (B.52) changes to

$$
\begin{aligned}
\varphi_R^\star &= NR\Bigg(1-1+\frac{qc_{\mu/\mu,\lambda}\sigma^{\star 2}}{RN\sqrt{\sigma^{\star 2}(1+q^2)+q^2\frac{\sigma^{\star 4}}{2N}+\sigma_\epsilon^{\star 2}}} \\
&\quad -\frac{q^2c_{\mu/\mu,\lambda}^2\sigma^{\star 4}}{2R^2N^2\left(\sigma^{\star 2}(1+q^2)+q^2\frac{\sigma^{\star 4}}{2N}+\sigma_\epsilon^{\star 2}\right)} \\
&\quad +\frac{qc_{\mu/\mu,\lambda}\sigma^{\star 4}}{2N^2R^2\mu\sqrt{\sigma^{\star 2}(1+q^2)+q^2\frac{\sigma^{\star 4}}{2N}+\sigma_\epsilon^{\star 2}}}-\frac{\sigma^{\star 2}}{2\mu NR^2}\Bigg)+\mathcal{O}(\frac{1}{\sqrt{N}}) \\
&= \frac{qc_{\mu/\mu,\lambda}\sigma^{\star 2}}{\sqrt{\sigma^{\star 2}(1+q^2)+q^2\frac{\sigma^{\star 4}}{2N}+\sigma_\epsilon^{\star 2}}} \\
&\quad -\frac{q^2c_{\mu/\mu,\lambda}^2\sigma^{\star 4}}{2RN\left(\sigma^{\star 2}(1+q^2)+q^2\frac{\sigma^{\star 4}}{2N}+\sigma_\epsilon^{\star 2}\right)} \\
&\quad +\frac{qc_{\mu/\mu,\lambda}\sigma^{\star 4}}{2NR\mu\sqrt{\sigma^{\star 2}(1+q^2)+q^2\frac{\sigma^{\star 4}}{2N}+\sigma_\epsilon^{\star 2}}}-\frac{\sigma^{\star 2}}{2\mu R}+\mathcal{O}(\frac{1}{\sqrt{N}})
\end{aligned}
\tag{B.64}
$$

Letting now $N \to \infty$, (B.64) changes to (B.62)

$$
\varphi_R^\star = \frac{qc_{\mu/\mu,\lambda}\sigma^{\star 2}}{\sqrt{\sigma^{\star 2}(1+q^2)+\sigma_\epsilon^{\star 2}}}-\frac{\sigma^{\star 2}}{2\mu R}
$$

Equation (B.62) will serve as an approximate formula for finite dimensional search spaces. Both progress rates, (B.48) and (B.62), were obtained for the case $\tau = 0$ and thus only applicable for small values of τ.

C The Self-Adaptation Response

This chapter presents the derivation of the self-adaptation response (SAR). The SAR is a central measure in the analysis of self-adaptive evolution strategies using the dynamic systems approach. This chapter is organized as follows: First, the general approach to determine the first-order SAR for $(\mu/\mu_I, \lambda)$-ES is introduced (C.1). During the derivations, several functions are expanded into their Taylor series'. For a first analysis, only the first derivations are needed. The second section gives the specific SARs for the sphere model and the ridge function. For a more precise approach, a general formula for determining the derivations is required. The remaining sections are devoted to this task.

C.1 A General Derivation

This section presents a general derivation of the SAR which is applicable to the sphere model as well as to the ridge functions. This derivation is only valid for small values of the learning rate τ since higher-order terms of τ which appear during the calculations are neglected.

The self-adaptation response function (SAR) denotes the expected relative change of the mean of the mutation strengths of the μ parents

$$\psi(\langle\sigma\rangle) \quad = \quad \mathrm{E}\Big[\frac{\langle\varsigma\rangle - \langle\sigma\rangle}{\langle\sigma\rangle}\Big] = \frac{1}{\mu}\sum_{m=1}^{\mu}\mathrm{E}\Big[\frac{\varsigma_{m;l} - \langle\sigma\rangle}{\langle\sigma\rangle}\Big]. \tag{C.1}$$

The random variable $\varsigma_{m;\lambda}$ denotes the mutation strength connected with the mth best quality or fitness change in λ trials. One of the main points in the derivation of the SAR is the determination of the corresponding probability density function (pdf) $p_{m;\lambda}(\varsigma)$. Note, as a rule the expectation in (C.1) depends on further variables. Since they depend in turn on the fitness model under consideration, they are not modeled at this point. They will come into play once the specific fitness models are considered. Furthermore, no normalization is introduced.

The general equation for the pdf can be given easily. Applying the concept of induced order statistics (see e.g. [3, 23, 4]) the pdf of the random variable leading to the mth highest fitness change Q in λ trials has to be derived. Putting it in another way, $m-1$ out of λ offspring must have a higher and $\lambda - m$ offspring must have a lower fitness change. Using the cdf of Q, $\mathrm{P}_Q(Q|\langle\sigma\rangle)$, the probability for the first condition is $Pr(q > Q) = 1 - Pr(q < Q) = 1 - \mathrm{P}_Q(Q|\langle\sigma\rangle)$ and $Pr(q < Q) = \mathrm{P}_Q(Q|\langle\sigma\rangle)$ in the case of the latter. It is easy to see using elementary combinatorics that there are

$$\lambda\binom{\lambda-1}{m-1} = \frac{\lambda!}{(m-1)!(\lambda-m)!} \tag{C.2}$$

different possibilities for these combinations. The resulting general equation for the pdf

$$\begin{aligned} p_{m;\lambda}(\varsigma|\langle\sigma\rangle) \quad = \quad & p_\sigma(\varsigma|\langle\sigma\rangle)\frac{\lambda!}{(m-1)!(\lambda-m)!} \\ & \times \int_{-\infty}^{\infty} p_Q(Q|\varsigma)\mathrm{P}_Q(Q|\langle\sigma\rangle)^{\lambda-m}\Big(1-\mathrm{P}_Q(Q|\langle\sigma\rangle)\Big)^{m-1}\,\mathrm{d}Q \end{aligned} \tag{C.3}$$

serves as the basis point for all further derivations. First, the cdf $\mathrm{P}_Q(Q|\langle\sigma\rangle)$ which appears in (C.3) must be determined. The cumulative density function is given as the expectation

$$\mathrm{P}_Q(Q|\langle\sigma\rangle) = \int_0^\infty P_Q(Q|\varsigma)p_\sigma(\varsigma|\langle\sigma\rangle)\,\mathrm{d}\varsigma. \tag{C.4}$$

Note, P_Q is used in (C.4) instead of P_Q to distinguish between the expectation P_Q and the cdf P_Q of Q for a ς. It was shown in [23, p.290] for a general function $f(\varsigma)$ that

$$\mathrm{E}[f(\varsigma)] = \int_0^\infty f(\varsigma)p_\sigma(\varsigma|\langle\sigma\rangle)\,\mathrm{d}\varsigma = f(\langle\sigma\rangle) + \mathcal{O}(\tau^2) \tag{C.5}$$

holds – provided that ς follows a log-normal distribution with parameter τ. A similar results holds for the symmetric two-point distribution. In the following, (C.4) is substituted by $P_Q(Q|\langle\sigma\rangle)$. The induced error vanishes for $\tau \to 0$. In the following, it is assumed that the argument of $P_Q(Q|\langle\sigma\rangle)$ is of the form

$$\frac{Q + h(\langle\sigma\rangle)}{g(\langle\sigma\rangle)} \tag{C.6}$$

with $h, g \in C^\infty(\mathbb{R})$, $g : \mathbb{R} \to \mathbb{R}^+$. This holds for example in the case of the sphere model and the ridge functions. As a next step, the standardized variable

$$\begin{aligned} z &= -\frac{Q + h(\langle\sigma\rangle)}{g(\langle\sigma\rangle)} \\ \Leftrightarrow Q &= -g(\langle\sigma\rangle)z - h(\langle\sigma\rangle) \end{aligned} \tag{C.7}$$

is introduced. Plugging (C.7) into (C.3) gives the pdf of the mutation strength

$$\begin{aligned} p_{m;\lambda}(\varsigma|\langle\sigma\rangle) &= p_\sigma(\varsigma|\langle\sigma\rangle)\frac{\lambda!}{(m-1)!(\lambda-m)!} \\ &\quad \times \int_{-\infty}^\infty g(\langle\sigma\rangle)p_z(-z|\varsigma)P_z(-z|\langle\sigma\rangle)^{\lambda-m}\Big(1 - P_z(-z|\langle\sigma\rangle)\Big)^{m-1}\,\mathrm{d}z. \end{aligned} \tag{C.8}$$

Let us now come back to the SAR (C.1) which has changed with (C.8) to

$$\begin{aligned} \psi(\langle\sigma\rangle) &= \int_0^\infty \Big(\frac{\varsigma - \langle\sigma\rangle}{\langle\sigma\rangle}\Big)p_\sigma(\varsigma|\langle\sigma\rangle)\frac{1}{\mu}\sum_{m=1}^{\mu}\frac{\lambda!}{(m-1)!(\lambda-m)!} \\ &\quad \times \int_{-\infty}^\infty g(\langle\sigma\rangle)p_z(-z|\varsigma)P_z(-z|\langle\sigma\rangle)^{\lambda-m}\Big(1 - P_z(-z|\langle\sigma\rangle)\Big)^{m-1}\,\mathrm{d}z\,\mathrm{d}\varsigma. \end{aligned} \tag{C.9}$$

In the case of the ridge functions and the sphere model, the approach can now be simplified. In both cases $P_Q(Q|\varsigma)$ is the cdf of the normal distribution with mean $h(\varsigma)$ and standard deviation $g(\varsigma)$. Thus,

$$P_Q(Q|\langle\sigma\rangle) = \Phi\Big(\frac{Q + h(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}\Big), \quad p_Q(Q|\varsigma) = \frac{1}{g(\varsigma)\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}\left(\frac{Q+h(\varsigma)}{g(\varsigma)}\right)^2}, \tag{C.10}$$

$$P_z(z) = \Phi(z), \quad \text{and} \;\; p_z(z|\varsigma) = \frac{1}{g(\varsigma)\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}\left(\frac{g(\langle\sigma\rangle)z-(h(\varsigma)-h(\langle\sigma\rangle))}{g(\varsigma)}\right)^2} \tag{C.11}$$

apply and the SAR (C.9) dissolves to

$$\begin{aligned}\psi(\langle\sigma\rangle) &= \int_0^\infty \Big(\frac{\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)p_\sigma(\varsigma|\langle\sigma\rangle)\frac{1}{\mu}\sum_{m=1}^{\mu}\frac{\lambda!}{(m-1)!(\lambda-m)!}\frac{1}{\sqrt{2\pi}} \\ &\quad\times\int_{-\infty}^\infty \frac{g(\langle\sigma\rangle)}{g(\varsigma)}\mathrm{e}^{-\frac{1}{2}\left(\frac{g(\langle\sigma\rangle)z-(h(\varsigma)-h(\langle\sigma\rangle))}{g(\varsigma)}\right)^2}\Big(1-\Phi(z)\Big)^{\lambda-m}\Phi(z)^{m-1}\,\mathrm{d}z\,\mathrm{d}\varsigma. \quad \text{(C.12)}\end{aligned}$$

In the next step, the order of the summation and the inner integration is swapped. The sum in (C.12)

$$\frac{1}{\mu}\sum_{m=1}^{\mu}\frac{\lambda!}{(m-1)!(\lambda-m)!}\Big(1-\Phi(z)\Big)^{\lambda-m}\Phi(z)^{m-1}$$

itself represents a regularized incomplete beta function [23, p. 147f] and can be substituted by an integral

$$\frac{1}{\mu}\sum_{m=1}^{\mu}\frac{\lambda!}{(m-1)!(\lambda-m)!}\Big(1-\Phi(z)\Big)^{\lambda-m}\Phi(z)^{m-1} = \frac{\lambda!}{\mu}\frac{\int_0^{1-\Phi(z)}x^{\lambda-\mu-1}(1-x)^{\mu-1}\,\mathrm{d}x}{(\lambda-\mu-1)!(\mu-1)!}. \quad \text{(C.13)}$$

Plugging the integral (C.13) into (C.9) leads to

$$\begin{aligned}\psi(\langle\sigma\rangle) &= \int_0^\infty \Big(\frac{\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)p_\sigma(\varsigma|\langle\sigma\rangle)\frac{1}{\sqrt{2\pi}}\int_{-\infty}^\infty \frac{g(\langle\sigma\rangle)}{g(\varsigma)}\mathrm{e}^{-\frac{1}{2}\left(\frac{g(\langle\sigma\rangle)z-(h(\varsigma)-h(\langle\sigma\rangle))}{g(\varsigma)}\right)^2} \\ &\quad\times\int_0^{1-\Phi(z)}x^{\lambda-\mu-1}(1-x)^{\mu-1}\frac{\lambda!}{(\lambda-\mu-1)!\mu!}\,\mathrm{d}x\,\mathrm{d}z\,\mathrm{d}\varsigma. \quad \text{(C.14)}\end{aligned}$$

Changing the integration order of the inner integrals over x and z in (C.14) gives

$$\begin{aligned}\psi(\langle\sigma\rangle) &= \int_0^\infty \frac{g(\langle\sigma\rangle)}{g(\varsigma)}\Big(\frac{\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)p_\sigma(\varsigma|\langle\sigma\rangle)\frac{1}{\sqrt{2\pi}}(\lambda-\mu)\binom{\lambda}{\mu} \\ &\quad\times\int_0^1 w^{\lambda-\mu-1}(1-w)^{\mu-1}\int_0^{\Phi^{-1}(1-w)}\mathrm{e}^{-\frac{1}{2}\left(\frac{g(\langle\sigma\rangle)z-(h(\varsigma)-h(\langle\sigma\rangle))}{g(\varsigma)}\right)^2}\mathrm{d}z\,\mathrm{d}w\,\mathrm{d}\varsigma. \quad \text{(C.15)}\end{aligned}$$

Setting finally $t=\Phi^{-1}(1-w)$,

$$\begin{aligned}\psi(\langle\sigma\rangle) &= \int_0^\infty \frac{g(\langle\sigma\rangle)}{g(\varsigma)}\Big(\frac{\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)p_\sigma(\varsigma|\langle\sigma\rangle)(\lambda-\mu)\binom{\lambda}{\mu} \\ &\quad\times\int_{-\infty}^\infty \Big(1-\Phi(t)\Big)^{\lambda-\mu-1}\Phi(t)^{\mu-1}\frac{\mathrm{e}^{-\frac{t^2}{2}}}{2\pi} \\ &\quad\times\int_{-\infty}^t \mathrm{e}^{-\frac{1}{2}\left(\frac{g(\langle\sigma\rangle)z-(h(\varsigma)-h(\langle\sigma\rangle))}{g(\varsigma)}\right)^2}\mathrm{d}z\,\mathrm{d}t\,\mathrm{d}\varsigma \quad \text{(C.16)}\end{aligned}$$

is obtained. This is the point to introduce further simplifications in order to solve the three integrals. The starting point is the innermost integral over z which leads to the cdf of the normal distribution with mean $(h(\varsigma)-h(\langle\sigma\rangle))/g(\langle\sigma\rangle)$ and standard deviation $g(\varsigma)/g(\langle\sigma\rangle)$. The SAR (C.16) changes to

$$\begin{aligned}\psi(\langle\sigma\rangle) &= \int_0^\infty \Big(\frac{\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)p_\sigma(\varsigma|\langle\sigma\rangle)(\lambda-\mu)\binom{\lambda}{\mu} \\ &\quad\times\int_{-\infty}^\infty \Big(1-\Phi(t)\Big)^{\lambda-\mu-1}\Phi(t)^{\mu-1}\frac{\mathrm{e}^{-\frac{t^2}{2}}}{\sqrt{2\pi}}\Phi\Big(\frac{g(\langle\sigma\rangle)t-(h(\varsigma)-h(\langle\sigma\rangle))}{g(\varsigma)}\Big)\,\mathrm{d}t\,\mathrm{d}\varsigma.\end{aligned}$$

The last Φ-cdf will be expanded into its Taylor series around $\langle\sigma\rangle$ since it hinders further calculations. The derivatives of Φ read

$$\Phi(f(\varsigma))' = f'(\varsigma)\frac{\mathrm{e}^{-\frac{f(\varsigma)^2}{2}}}{\sqrt{2\pi}} \tag{C.17}$$

$$\frac{\partial^{k+1}}{\partial\varsigma^{k+1}}\Phi(f(\varsigma)) = \frac{\partial^k}{\partial\varsigma^k}\left(f'(\varsigma)\frac{\mathrm{e}^{-\frac{f(\varsigma)^2}{2}}}{\sqrt{2\pi}}\right) \text{ for } k>0 \tag{C.18}$$

with $f(\varsigma):=g(\langle\sigma\rangle)/g(\varsigma)t-(h(\varsigma)-h(\langle\sigma\rangle))/g(\varsigma)$. For the time being, the exact higher order coefficients are not needed in the approach since the Taylor series will be cut off eventually after the first terms. Using (C.17) and (C.18), the Taylor series reads

$$\begin{aligned} T_\Phi(t,\varsigma) = {} & \Phi(t)+\frac{\mathrm{e}^{-\frac{t^2}{2}}}{\sqrt{2\pi}}\Big(-\frac{g'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}t-\frac{h'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}\Big)\langle\sigma\rangle\Big(\frac{\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)+ \\ & \frac{1}{\sqrt{2\pi}}\sum_{k=1}^{\infty}\frac{\langle\sigma\rangle^{k+1}}{(k+1)!}\Big(\frac{\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)^{k+1} \\ & \times\frac{\partial^k}{\partial\varsigma^k}\Bigg(\Big(-g(\langle\sigma\rangle)\frac{g'(\varsigma)}{g^2(\varsigma)}t+\frac{g'(\varsigma)(h(\varsigma)-h(\langle\sigma\rangle))}{g(\varsigma)^2}-\frac{h'(\varsigma)}{g(\varsigma)}\Big) \\ & \times\exp\Big(-\frac{1}{2}\Big(\frac{g(\langle\sigma\rangle)t-(h(\varsigma)-h(\langle\sigma\rangle))}{g(\varsigma)}\Big)^2\Big)\Bigg)|_{\varsigma=\langle\sigma\rangle}. \end{aligned} \tag{C.19}$$

Plugging (C.19) into the SAR (C.9), three integrals are obtained: one containing the normal distribution function at $\langle\sigma\rangle$, one comprising the first derivation and a quadratic $(\varsigma-\langle\sigma\rangle)$-term, and one with higher derivations and polynomials in $(\varsigma-\langle\sigma\rangle)$ with degree three or higher

$$\begin{aligned} \psi(\langle\sigma\rangle) = {} & \int_0^\infty\Big(\frac{\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)p_\sigma(\varsigma|\langle\sigma\rangle)(\lambda-\mu)\binom{\lambda}{\mu}\int_{-\infty}^{\infty}\Big(1-\Phi(t)\Big)^{\lambda-\mu-1}\Phi(t)^\mu\frac{\mathrm{e}^{-\frac{t^2}{2}}}{\sqrt{2\pi}}\,\mathrm{d}t\,\mathrm{d}\varsigma \\ & +\int_0^\infty\langle\sigma\rangle\Big(\frac{\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)^2p_\sigma(\varsigma|\langle\sigma\rangle)(\lambda-\mu)\binom{\lambda}{\mu} \\ & \times\int_{-\infty}^{\infty}\Big(1-\Phi(t)\Big)^{\lambda-\mu-1}\Phi(t)^{\mu-1}\frac{\mathrm{e}^{-t^2}}{2\pi}\Big(-\frac{g'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}t-\frac{h'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}\Big)\,\mathrm{d}t\,\mathrm{d}\varsigma \\ & +\sum_{k=1}^{\infty}\int_0^\infty\frac{\langle\sigma\rangle^{k+1}}{(k+1)!}\Big(\frac{(\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)^{k+2}p_\sigma(\varsigma|\langle\sigma\rangle) \\ & \times(\lambda-\mu)\binom{\lambda}{\mu}\int_{-\infty}^{\infty}\Big(1-\Phi(t)\Big)^{\lambda-\mu-1}\Phi(t)^{\mu-1}\frac{\mathrm{e}^{-\frac{t^2}{2}}}{2\pi} \\ & \times\frac{\partial^k}{\partial\varsigma^k}\Bigg(\Big(-g(\langle\sigma\rangle)\frac{g'(\varsigma)}{g^2(\varsigma)}t+\frac{g'(\varsigma)(h(\varsigma)-h(\langle\sigma\rangle))}{g((\varsigma)^2}-\frac{h'(\varsigma)}{g(\varsigma)}\Big) \\ & \times\exp\Big(-\frac{1}{2}\Big(\frac{g(\langle\sigma\rangle)t-(h(\varsigma)-h(\langle\sigma\rangle))}{g(\varsigma)}\Big)^2\Big)\Bigg)|_{\varsigma=\langle\sigma\rangle}\,\mathrm{d}t\,\mathrm{d}\varsigma. \end{aligned} \tag{C.20}$$

First of all, the integration over ς is addressed. The remainder of this section is restricted the log-normal distribution with learning rate τ. First of all, note that the expectation of $(\varsigma-\langle\sigma\rangle)^k$ leads to a

series in τ^{2l}. It can be shown that the expectation of $[(\varsigma - \langle\sigma\rangle)/\langle\sigma\rangle]^k$ does not include any τ^{2l}-Terms with $2l + 1 < k$. At this point, the series is expanded to the precision of τ^2, thus the expectations of terms $[(\varsigma - \langle\sigma\rangle)/\langle\sigma\rangle]^k$, $k \geq 3$, enters the error term. Section C.3 is aimed at developing a more accurate formula for the SAR. The equations obtained there are lengthy and complicated, though.

Considering (C.20) reveals that the last integral contributes only to the error term. Equation (C.20) can therefore be given by

$$\begin{aligned}
\psi(\langle\sigma\rangle) &= \overline{\Big(\frac{\varsigma - \langle\sigma\rangle}{\langle\sigma\rangle}\Big)}(\lambda-\mu)\binom{\lambda}{\mu}\int_{-\infty}^{\infty}\Big(1-\Phi(t)\Big)^{\lambda-\mu-1}\Phi(t)^{\mu}\frac{\mathrm{e}^{-\frac{t^2}{2}}}{\sqrt{2\pi}}\,\mathrm{d}t \\
&\quad +\langle\sigma\rangle\overline{\Big(\frac{\varsigma - \langle\sigma\rangle}{\langle\sigma\rangle}\Big)^2}(\lambda-\mu)\binom{\lambda}{\mu}\int_{-\infty}^{\infty}\Big(-\frac{g'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}t-\frac{h'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}\Big) \\
&\quad \times\Big(1-\Phi(t)\Big)^{\lambda-\mu-1}\Phi(t)^{\mu-1}\frac{\mathrm{e}^{-t^2}}{2\pi}\,\mathrm{d}t+\mathcal{O}\Big(\overline{(\varsigma-\langle\sigma\rangle)^3}\Big) \qquad \text{(C.21)}
\end{aligned}$$

or inserting the expectations obtained in Section C.2

$$\begin{aligned}
\psi(\langle\sigma\rangle) &= \Big(\frac{\tau^2}{2}+\mathcal{O}(\tau^4)\Big) \\
&\quad \times(\lambda-\mu)\binom{\lambda}{\mu}\int_{-\infty}^{\infty}\Big(1-\Phi(t)\Big)^{\lambda-\mu-1}\Phi(t)^{\mu}\frac{\mathrm{e}^{-\frac{t^2}{2}}}{\sqrt{2\pi}}\,\mathrm{d}t \\
&\quad +\Big(\langle\sigma\rangle\tau^2+\mathcal{O}(\tau^4)\Big)\Big(-\frac{g'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}t-\frac{h'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}\Big) \\
&\quad \times(\lambda-\mu)\binom{\lambda}{\mu}\int_{-\infty}^{\infty}\Big(1-\Phi(t)\Big)^{\lambda-\mu-1}\Phi(t)^{\mu-1}\frac{\mathrm{e}^{-t^2}}{2\pi}\,\mathrm{d}t+\mathcal{O}(\tau^4) \\
&= \tau^2\Big(\frac{1}{2}(\lambda-\mu)\binom{\lambda}{\mu}\int_{-\infty}^{\infty}\Big(1-\Phi(t)\Big)^{\lambda-\mu-1}\Phi(t)^{\mu}\frac{\mathrm{e}^{-\frac{t^2}{2}}}{\sqrt{2\pi}}\,\mathrm{d}t \\
&\quad +\langle\sigma\rangle(\lambda-\mu)\binom{\lambda}{\mu}\int_{-\infty}^{\infty}\Big(-\frac{g'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}t-\frac{h'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}\Big) \\
&\quad \times\Big(1-\Phi(t)\Big)^{\lambda-\mu-1}\Phi(t)^{\mu-1}\frac{\mathrm{e}^{-t^2}}{2\pi}\,\mathrm{d}t\Big)+\mathcal{O}(\tau^4). \qquad \text{(C.22)}
\end{aligned}$$

The value of the first integral is one. The other integral cannot be solved analytically. Instead, the generalized progress coefficients $e^{\alpha,\beta}_{\mu,\lambda}$ are used. Reconsidering the definition (A.24), p. 126

$$e^{\alpha,\beta}_{\mu,\lambda} = \frac{\lambda-\mu}{\sqrt{2\pi}^{\alpha+1}}\binom{\lambda}{\mu}\int_{-\infty}^{\infty}t^{\beta}\mathrm{e}^{-\frac{\alpha+1}{2}t^2}\Phi(t)^{\lambda-\mu-1}\Big(1-\Phi(t)\Big)^{\mu-\alpha}\,\mathrm{d}t$$

with $c_{\mu/\mu,\lambda} := e^{1,0}_{\mu,\lambda}$, the SAR is given by

$$\psi(\langle\sigma\rangle) = \tau^2\Big(\frac{1}{2}+\langle\sigma\rangle\Big(e^{1,1}_{\mu,\lambda}\frac{g'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}-c_{\mu/\mu,\lambda}\frac{h'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}\Big)\Big)+\mathcal{O}(\tau^4). \qquad \text{(C.23)}$$

The self-adaptation response (C.23) has been derived under the assumption that τ is sufficiently small. In the case of the two-point operator a similar result holds provided that the parameter β is sufficiently small. In this case, the SAR reads

$$\psi(\langle\sigma\rangle) = \beta^2(1+\beta)\left(\frac{1}{2} + \langle\sigma\rangle\left(e^{1,1}_{\mu,\lambda}\frac{g'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)} - c_{\mu/\mu,\lambda}\frac{h'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}\right)\right) + \mathcal{O}(\beta^4). \quad \text{(C.24)}$$

This can be easily verified by following the approach until Eq. (C.21), using a similar argument to drop the higher order terms of β (see Section C.2) and inserting the expectations into (C.21).

The general equations (C.23) and (C.24) can now be used to give the first-order SAR for the sphere model and the ridge functions for sufficiently small values of τ and β. In the next subsection, the SARs for the undisturbed and noisy sphere model are derived. Afterwards, the SARs for ridge functions will be given.

C.1.1 Sphere Model: The self-adaptation response function for $\tau \ll 1$

First, the SAR of the noise-free sphere model is determined. The general SAR (C.23)

$$\psi(\langle\sigma\rangle) = \tau^2\left(\frac{1}{2} + \langle\sigma\rangle\left(e^{1,1}_{\mu,\lambda}\frac{g'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)} - c_{\mu/\mu,\lambda}\frac{h'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}\right)\right) + \mathcal{O}(\tau^4)$$

requires the determination of the functions g and h and their derivatives. These functions can be obtained by considering the cdf of the fitness change $P_Q(Q|\varsigma) = P((Q + h(\varsigma))/g(\varsigma))$ (cf. Eq. (B.12)). In this section, the log-normal distribution is considered. The equation for the symmetric two-point operator (C.23) can be obtained by substituting τ^2 with $\beta^2(1+\beta)$. In this section, the fitness function is denoted by $f(\mathbf{y}) = w(\|\mathbf{y} - \hat{\mathbf{y}}\|) = w(R)$.

The Undisturbed Sphere Model In the case of the noise-free sphere, the pdf of fitness change reads

$$P_Q(Q|\varsigma) = \Phi\left(\frac{Q + \frac{Nw'(R)}{2R}\varsigma^2}{w'(R)\sqrt{\varsigma^2 + \frac{N}{2R^2}\varsigma^4}}\right). \quad \text{(C.25)}$$

The functions required, h and g are therefore

$$h(\varsigma) = \frac{Nw'(R)}{2R}\varsigma^2, \;\; h'(\varsigma) = \frac{Nw'(R)}{R}\varsigma \quad \text{(C.26)}$$

and

$$g(\varsigma) = w'(R)\sqrt{\varsigma^2 + \frac{N}{2R^2}\varsigma^4}, \;\; g'(\varsigma) = w'(R)\frac{2\varsigma + 4\frac{N}{2R^2}\varsigma^3}{2\sqrt{\varsigma^2 + \frac{N}{2R^2}\varsigma^4}}. \quad \text{(C.27)}$$

After inserting (C.26) and (C.27) into the SAR (C.23),

$$\psi(\langle\sigma\rangle) = \tau^2\left(\frac{1}{2} + \langle\sigma\rangle\left(e^{1,1}_{\mu,\lambda}\frac{2\langle\sigma\rangle + 4\frac{N}{2R^2}\langle\sigma\rangle^3}{2\left(\langle\sigma\rangle^2 + \frac{N}{2R^2}\langle\sigma\rangle^4\right)} - c_{\mu/\mu,\lambda}\frac{\frac{N}{R}\langle\sigma\rangle}{\sqrt{\langle\sigma\rangle^2 + \frac{N}{2R^2}\langle\sigma\rangle^4}}\right)\right) + \mathcal{O}(\tau^4) \quad \text{(C.28)}$$

is obtained. Introducing the usual normalization, $\sigma^* := N/R\langle\sigma\rangle$, changes (C.28) to

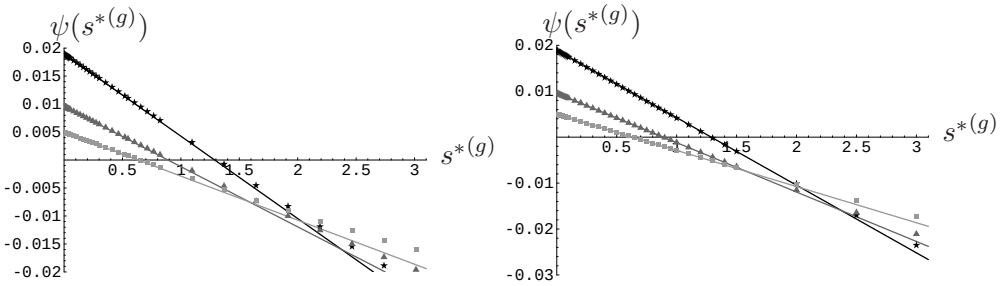
$$\begin{aligned}\psi(\sigma^*) &= \tau^2\left(\frac{1}{2} + \frac{R\sigma^*}{N}\left(e^{1,1}_{\mu,\lambda}\frac{2\frac{R\sigma^*}{N} + 2R\frac{\sigma^{*3}}{N^2}}{2\left(\frac{R^2\sigma^{*2}}{N^2} + R^2\frac{\sigma^{*4}}{2N^3}\right)} - c_{\mu/\mu,\lambda}\frac{N\frac{\sigma^*}{N}}{\sqrt{\frac{R^2\sigma^{*2}}{N^2} + R^2\frac{\sigma^{*4}}{2N^3}}}\right)\right) + \mathcal{O}(\tau^4) \\ &= \tau^2\left(\frac{1}{2} + \sigma^*\left(e^{1,1}_{\mu,\lambda}\frac{2\sigma^* + 2\frac{\sigma^{*3}}{N}}{2\left(\sigma^{*2} + \frac{\sigma^{*4}}{2N}\right)} - c_{\mu/\mu,\lambda}\frac{\sigma^*}{\sqrt{\sigma^{*2} + \frac{\sigma^{*4}}{2N}}}\right)\right) + \mathcal{O}(\tau^4). \quad \text{(C.29)}\end{aligned}$$

Letting $N \to \infty$, the limit of (C.29) is obtained as

$$\psi(\sigma^*) \quad = \quad \tau^2\Big(\frac{1}{2} + e^{1,1}_{\mu,\lambda} - c_{\mu/\mu,\lambda}\sigma^*\Big) + \mathcal{O}(\tau^4). \tag{C.30}$$

Equation (C.30) will be used in the further calculations in Chapter 4. Equations (C.29) and (C.30) were obtained under the following conditions: As small learning rate τ (or the parameter β, respectively). This is due to the derivation of the general equation of the SAR. In the case of (C.29), a high-dimensional search space is required which is due to a requirement in obtaining the cdf of the fitness change (B.12).

Equation (C.30) was compared with the results of ES-runs (Fig. C.1). In all experiments, the negative sphere function $F(\mathbf{y}) = -\|\mathbf{y}\|^2$ was used as fitness function. The N-dependency is weak in the experimental results and the agreement at least for smaller mutation strengths is good.



a) $N = 30$, $\tau = 0.1$ b) $N = 100$, $\tau = 0.1$

Figure C.1: The self-adaptation response ψ in the case of a log-normal distribution of the mutation strength. The points denote the results of one-generation experiments. Each data point was averaged over 250 000 trials. As initial vector $\mathbf{y}^{(0)} = \mathbf{10}$ was chosen. From top to bottom the results for $(10/10_I, 60)$-, $(20/20_I, 60)$-, and $(30/30_I, 60)$-ES are shown.

The SAR for the Noisy Sphere In the case of the noisy sphere, the pdf of the fitness change is given by

$$P_Q(Q|\varsigma) \quad = \quad \Phi\left(\frac{Q + \frac{Nw'(R)}{2R}\varsigma^2}{w'(R)\sqrt{\varsigma^2 + \frac{\sigma_\epsilon^2}{w'(R)} + \frac{N}{2R^2}\varsigma^4}}\right) \tag{C.31}$$

(see B.12). The functions h and g required for the general SAR (C.23) are

$$h(\varsigma) \quad = \quad \frac{Nw'(R)}{2R}\varsigma^2, \;\; h'(\varsigma) = \frac{Nw'(R)}{R}\varsigma \tag{C.32}$$

and

$$g(\varsigma) \quad = \quad w'(R)\sqrt{\varsigma^2 + \frac{\sigma_\epsilon^2}{w'(R)^2} + \frac{N}{2R^2}\varsigma^4}, \;\; g'(\varsigma) = w'(R)\frac{2\varsigma + 2\frac{N}{R^2}\varsigma^3}{2\sqrt{\varsigma^2 + \frac{\sigma_\epsilon^2}{w'(R)^2} + \frac{N}{2R^2}\varsigma^4}}. \tag{C.33}$$

Note, again w denotes the fitness function instead of the original symbol g in (B.12).

Inserting (C.32) and (C.33) into the SAR (C.23) gives

$$\begin{aligned}\psi(\langle\sigma\rangle) &= \tau^2\Big(\frac{1}{2}+\langle\langle\sigma\rangle\rangle\Big(e^{1,1}_{\mu,\lambda}\frac{\langle\sigma\rangle+\frac{N}{R^2}\langle\sigma\rangle^3}{\Big(\langle\sigma\rangle^2+\frac{\sigma_\epsilon^2}{w'(R)^2}+\frac{N}{2R^2}\langle\sigma\rangle^4\Big)}\\ &\quad -c_{\mu/\mu,\lambda}\frac{\frac{N}{R}\langle\sigma\rangle}{\sqrt{\langle\sigma\rangle^2+\frac{\sigma_\epsilon^2}{w'(R)^2}+\frac{N}{2R^2}\langle\sigma\rangle^4}}\Big)\Big)+\mathcal{O}(\tau^4).\end{aligned} \tag{C.34}$$

Now the same normalization as in the previous section is introduced – setting $\sigma^* := (N/R)\langle\sigma\rangle$ and $\sigma^*_\epsilon := [N/(w'(R)R)]\sigma_\epsilon$. The SAR changes to

$$\begin{aligned}\psi(\sigma^*) &= \tau^2\Big(\frac{1}{2}+\frac{R\sigma^*}{N}\Big(e^{1,1}_{\mu,\lambda}\frac{R\frac{\sigma^*}{N}+R\frac{\sigma^{*3}}{N^2}}{\Big(\frac{R^2\sigma^{*2}}{N^2}+\frac{R^2\sigma_\epsilon^{*2}}{2N^2}+R^2\frac{\sigma^{*4}}{2N^3}\Big)}\\ &\quad -c_{\mu/\mu,\lambda}\frac{R\sigma^*}{\sqrt{\frac{R^2\sigma^{*2}}{N^2}+\frac{R^2\sigma^{*2}}{N^2}+R^2\frac{\sigma^{*4}}{2N^3}}}\Big)\Big)+\mathcal{O}(\tau^4).\end{aligned} \tag{C.35}$$

Letting $N\to\infty$,

$$\psi(\sigma^*) = \tau^2\Big(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\frac{\sigma^{*2}}{\sigma^{*2}+\sigma_\epsilon^{*2}}-c_{\mu/\mu,\lambda}\frac{\sigma^{*2}}{\sqrt{\sigma^{*2}+\sigma_\epsilon^{*2}}}\Big)+\mathcal{O}(\tau^4) \tag{C.36}$$

is obtained. The conditions under which (C.35) and (C.36) were obtained are the same as in the case of the noise-free sphere: A small learning rate τ (or the parameter β, respectively). Again, (C.35) is obtained for large search spaces.

Both SARs (C.35) and (C.36) are compared with the results of experiments. The set-up of the experiments is similar to the noise free case. The experiments were conducted using $(\mu/\mu_I, 100)$-ES Each data point was sampled over $250,000$ runs of one-generation experiments. Two search space dimensionalities were investigated: $N = 30$ and $N = 100$. Even in the low-dimensional search space ($N = 30$), the prediction quality is reasonable good – expecially for smaller mutation strengths. Deviations occur for higher mutation strengths. This is more pronounced for smaller noise strengths than for higher. Increasing the mutation strength eventually results in a deviation from the N-dependent prediction (C.35).

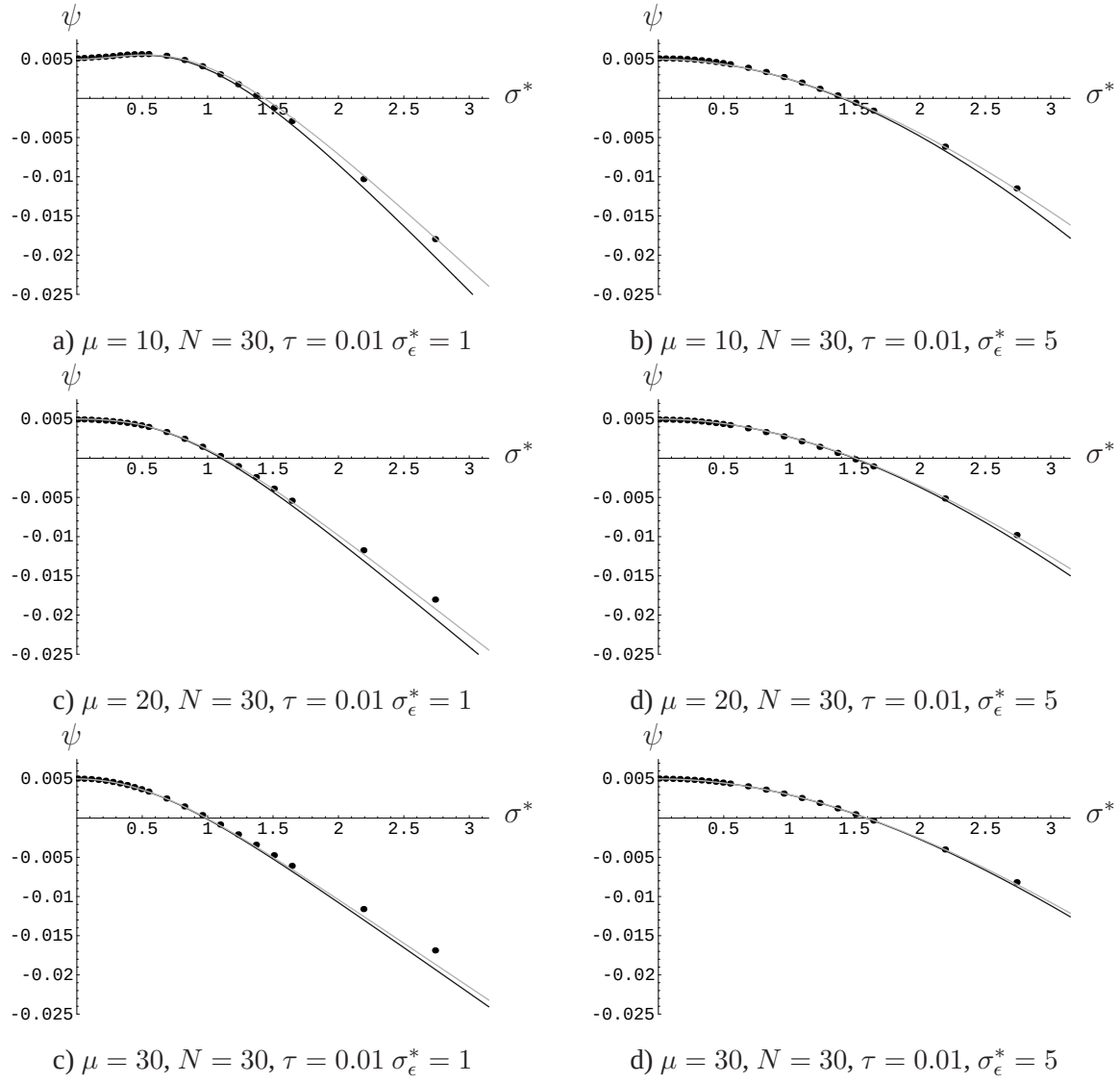
C.1.2 Ridge Functions: The Self-Adaptation Response Function for $\tau \ll 1$

This section is devoted to the task of determining the SAR for ridge functions. Recall, the SAR is given by (C.23)

$$\psi(\langle\sigma\rangle) = \tau^2\left(\frac{1}{2}+\langle\sigma\rangle\Big(e^{1,1}_{\mu,\lambda}\frac{g'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}-c_{\mu/\mu,\lambda}\frac{h'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}\Big)\right)+\mathcal{O}(\tau^4)$$

with g and h stemming from the cdf of the fitness change of the form $P_Q(Q|\varsigma) = P[(Q+h(\varsigma))/g(\varsigma)]$. In this section, the equation for the symmetric two-point operator (C.23) is not given explicitely, since it can be obtained by substituting τ^2 with $\beta^2(1+\beta)$.

As in the case of the sphere model, first undisturbed ridge functions are considered before the SAR for noisy ridge functions is derived.



a) $\mu = 10$, $N = 30$, $\tau = 0.01$ $\sigma_\epsilon^* = 1$

b) $\mu = 10$, $N = 30$, $\tau = 0.01$, $\sigma_\epsilon^* = 5$

c) $\mu = 20$, $N = 30$, $\tau = 0.01$ $\sigma_\epsilon^* = 1$

d) $\mu = 20$, $N = 30$, $\tau = 0.01$, $\sigma_\epsilon^* = 5$

c) $\mu = 30$, $N = 30$, $\tau = 0.01$ $\sigma_\epsilon^* = 1$

d) $\mu = 30$, $N = 30$, $\tau = 0.01$, $\sigma_\epsilon^* = 5$

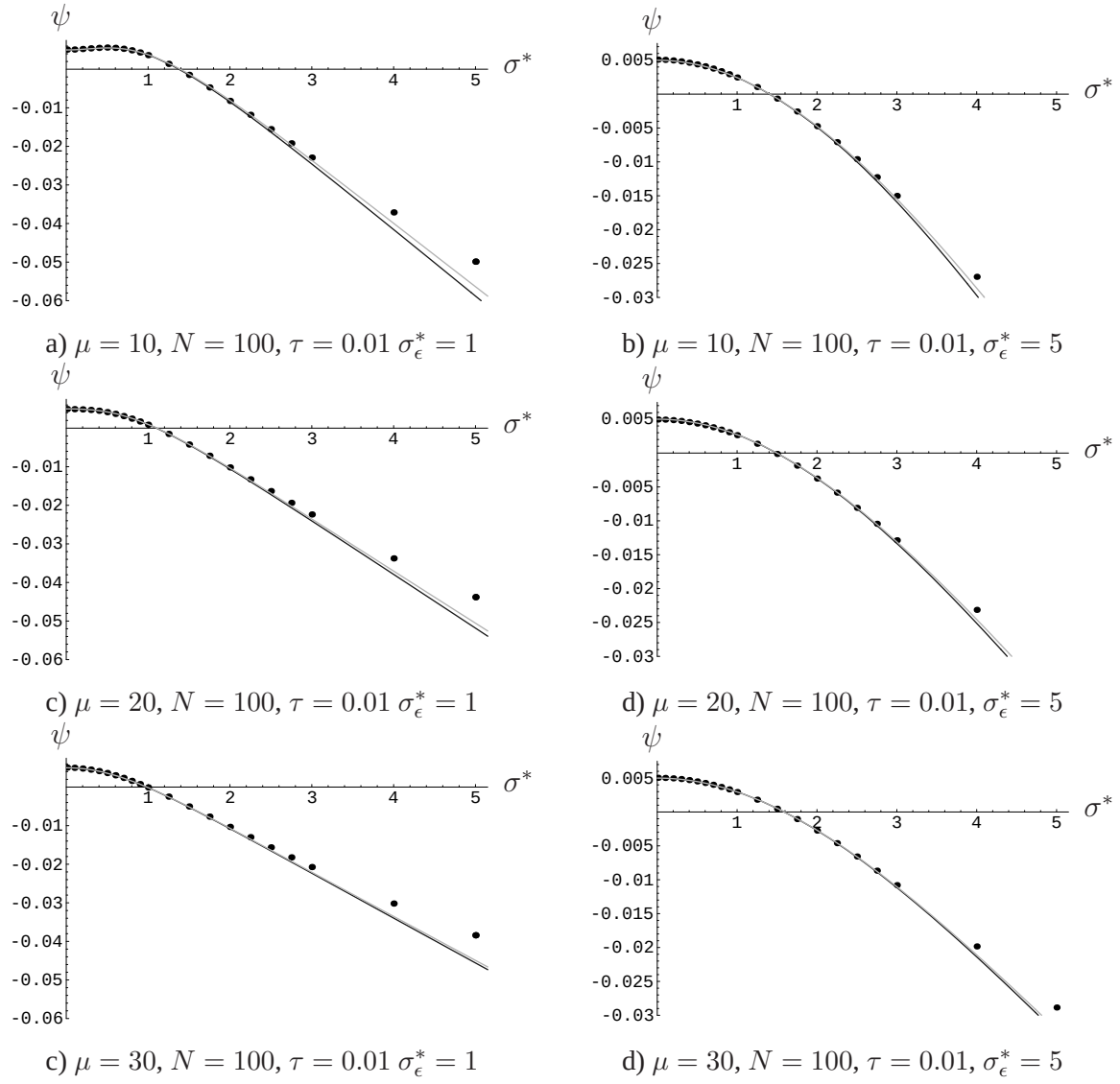
Figure C.2: The first-order self-adaptation response function ψ for some choices of τ and some $(\mu/\mu_I, 100)$-ES. Equation (C.35) is represented by the gray lines, whereas the black lines denote (C.36). The points denote the results of one-generation experiments and each was obtained by averaging over $250,000$ trials.

The Undisturbed Ridge In the case of the noise-free ridge function, the pdf of fitness change is given by (B.33), p. 133,

$$P_Q(Q|\varsigma) = \Phi\left(\frac{Q + \alpha d R^{\alpha-1}\frac{N}{2R}\varsigma^2}{\sqrt{\varsigma^2(1+\alpha^2 d^2 R^{2\alpha-2}) + \alpha^2 d^2 R^{2\alpha-2}\frac{N}{2R^2}\varsigma^4}}\right). \tag{C.37}$$

The functions required for the SAR (C.23) are therefore

$$h(\varsigma) = \alpha d R^{\alpha-1}\frac{N}{2R}\varsigma^2$$



a) $\mu = 10$, $N = 100$, $\tau = 0.01$ $\sigma_\epsilon^* = 1$ b) $\mu = 10$, $N = 100$, $\tau = 0.01$, $\sigma_\epsilon^* = 5$

c) $\mu = 20$, $N = 100$, $\tau = 0.01$ $\sigma_\epsilon^* = 1$ d) $\mu = 20$, $N = 100$, $\tau = 0.01$, $\sigma_\epsilon^* = 5$

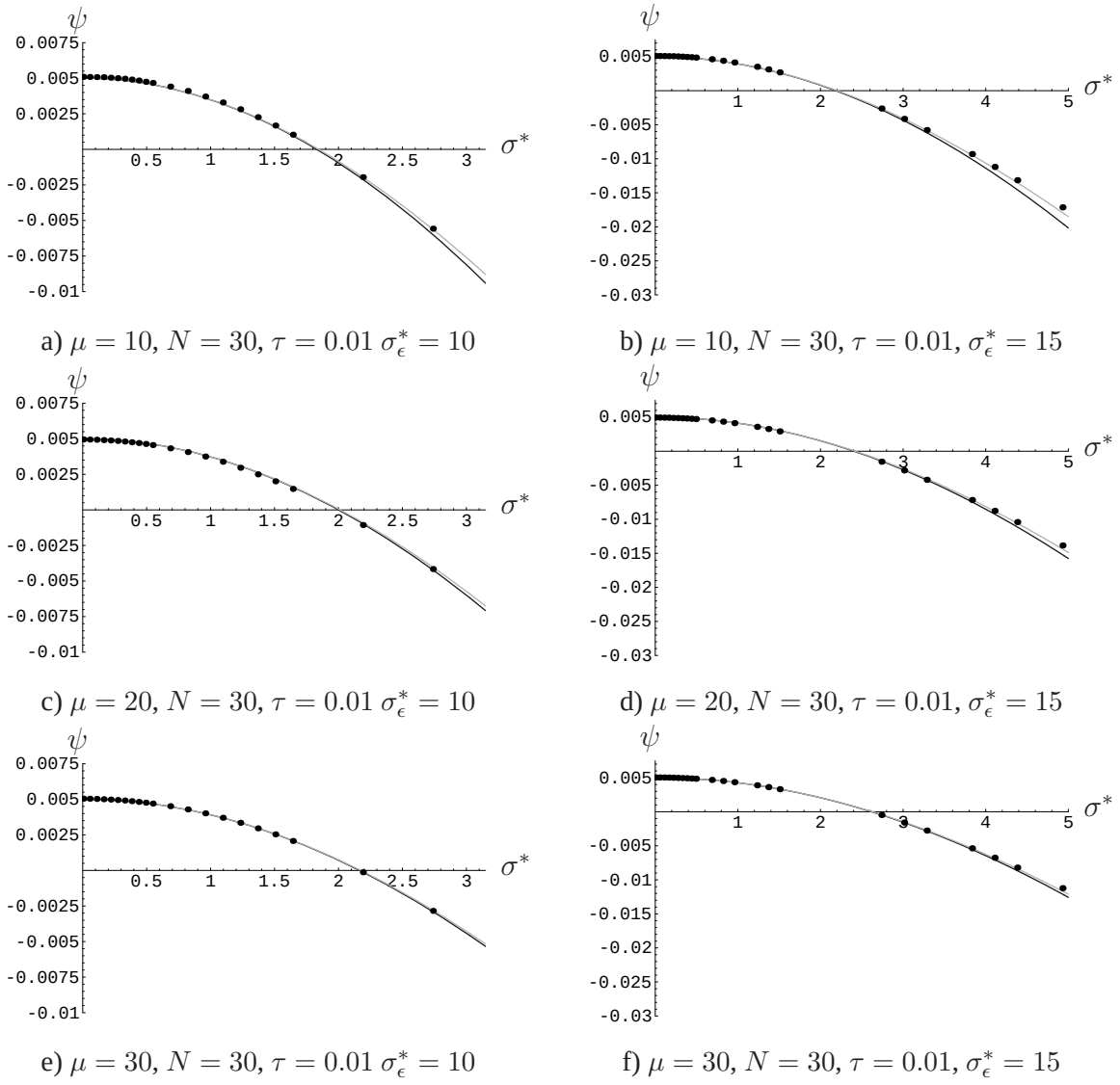
c) $\mu = 30$, $N = 100$, $\tau = 0.01$ $\sigma_\epsilon^* = 1$ d) $\mu = 30$, $N = 100$, $\tau = 0.01$, $\sigma_\epsilon^* = 5$

Figure C.3: The first-order self-adaptation response function ψ for some choices of τ and some $(\mu/\mu_I, 100)$-ES. Equation (C.35) is represented by the gray lines, whereas the black lines denote (C.36). The points denote the results of one-generation experiments and each was obtained by averaging over $250,000$ trials.

$$h'(\varsigma) \quad = \quad \alpha d R^{\alpha-1} \frac{N}{R} \varsigma \tag{C.38}$$

and

$$\begin{aligned} g(\varsigma) &= \sqrt{\varsigma^2(1+\alpha^2 d^2 R^{2\alpha-2}) + \alpha^2 d^2 R^{2\alpha-2} \frac{N}{2R^2} \varsigma^4} \\ g'(\varsigma) &= \frac{2\varsigma(1+\alpha^2 d^2 R^{2\alpha-2}) + 2\alpha^2 d^2 R^{2\alpha-2} \frac{N}{R^2} \varsigma^3}{2\sqrt{\varsigma^2(1+\alpha^2 d^2 R^{2\alpha-2}) + \alpha^2 d^2 R^{2\alpha-2} \frac{N}{2R^2} \varsigma^4}} \end{aligned} \tag{C.39}$$



a) $\mu = 10$, $N = 30$, $\tau = 0.01$ $\sigma_\epsilon^* = 10$

b) $\mu = 10$, $N = 30$, $\tau = 0.01$, $\sigma_\epsilon^* = 15$

c) $\mu = 20$, $N = 30$, $\tau = 0.01$ $\sigma_\epsilon^* = 10$

d) $\mu = 20$, $N = 30$, $\tau = 0.01$, $\sigma_\epsilon^* = 15$

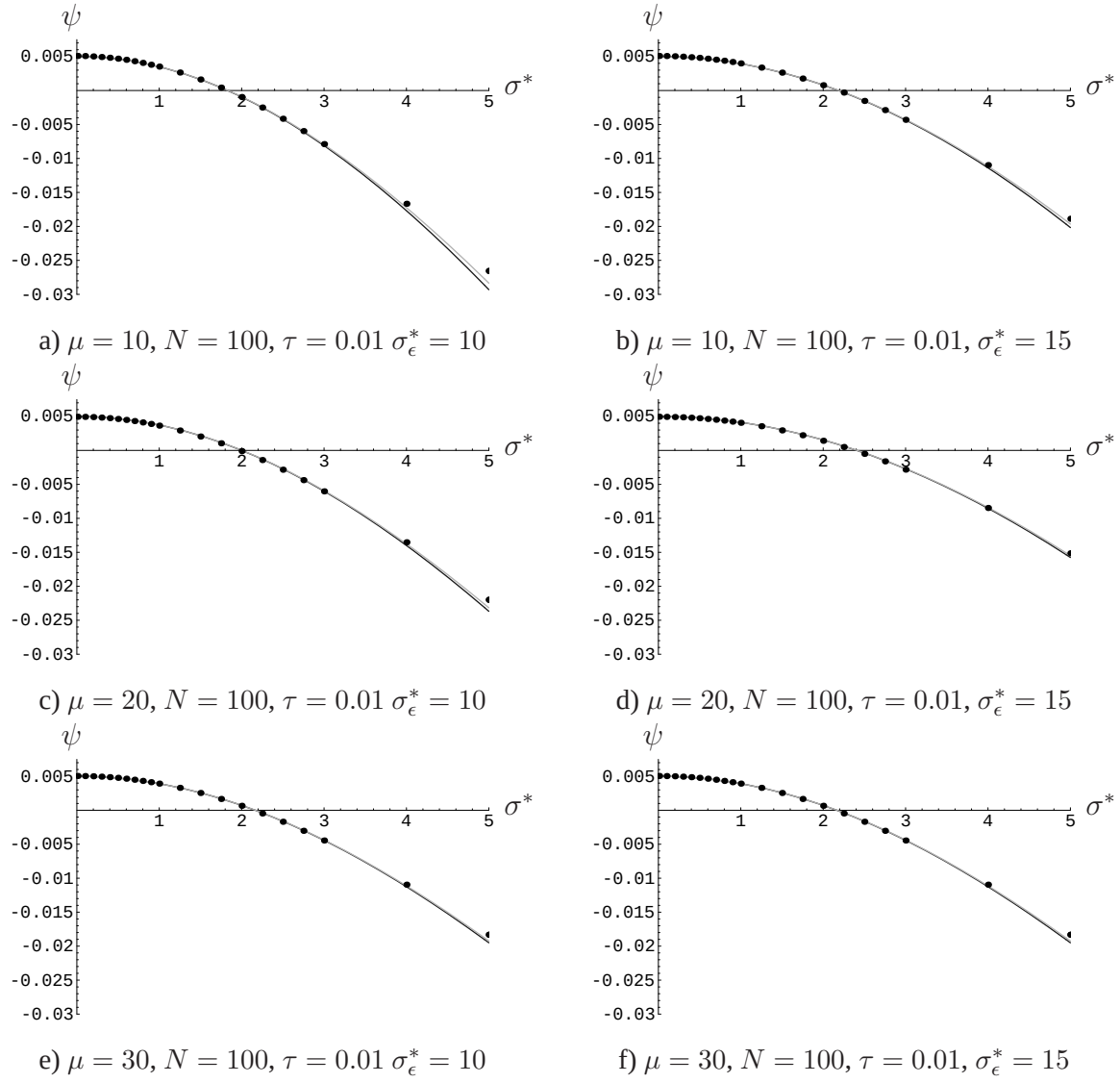
e) $\mu = 30$, $N = 30$, $\tau = 0.01$ $\sigma_\epsilon^* = 10$

f) $\mu = 30$, $N = 30$, $\tau = 0.01$, $\sigma_\epsilon^* = 15$

Figure C.4: The first-order self-adaptation response function ψ for some choices of τ and some $(\mu/\mu_I, 100)$-ES. Equation (C.35) is represented by the gray lines, whereas the black lines denote (C.36). The points denote the results of one-generation experiments and each was obtained by averaging over $250,000$ trials.

Plugging (C.38) and (C.39) into the SAR (C.22) leads to

$$\psi(\langle\sigma\rangle) = \tau^2\Big(\frac{1}{2} + \langle\langle\sigma\rangle\rangle\Big(e^{1,1}_{\mu,\lambda}\frac{\langle\sigma\rangle(1+\alpha^2d^2R^{2\alpha-2}) + \alpha^2d^2R^{2\alpha-2}\frac{N}{R^2}\langle\sigma\rangle^3}{\Big(\langle\sigma\rangle^2(1+\alpha^2d^2R^{2\alpha-2}) + \alpha^2d^2R^{2\alpha-2}\frac{N}{2R^2}\langle\sigma\rangle^4\Big)} - c_{\mu/\mu,\lambda}\frac{\alpha dR^{\alpha-1}\frac{N}{R}\langle\sigma\rangle}{\sqrt{\langle\sigma\rangle^2(1+\alpha^2d^2R^{2\alpha-2}) + \alpha^2d^2R^{2\alpha-2}\frac{N}{2R^2}\langle\sigma\rangle^4}}\Big)\Big) + \mathcal{O}(\tau^4). \quad \text{(C.40)}$$



a) $\mu = 10$, $N = 100$, $\tau = 0.01$ $\sigma^*_\epsilon = 10$

b) $\mu = 10$, $N = 100$, $\tau = 0.01$, $\sigma^*_\epsilon = 15$

c) $\mu = 20$, $N = 100$, $\tau = 0.01$ $\sigma^*_\epsilon = 10$

d) $\mu = 20$, $N = 100$, $\tau = 0.01$, $\sigma^*_\epsilon = 15$

e) $\mu = 30$, $N = 100$, $\tau = 0.01$ $\sigma^*_\epsilon = 10$

f) $\mu = 30$, $N = 100$, $\tau = 0.01$, $\sigma^*_\epsilon = 15$

Figure C.5: The first-order self-adaptation response function ψ for some choices of τ and some $(\mu/\mu_I, 100)$-ES. Equation (C.35) is represented by the gray lines, whereas (C.36) is represented by the black lines. The points denote the results of one-generation experiments and each was obtained by averaging over $250,000$ trials.

Using the normalization $\sigma^\star := N\langle\sigma\rangle$, the SAR changes to

$$\psi(\sigma^\star) = \tau^2\Big(\frac{1}{2} + \frac{\sigma^\star}{N}\Big(e^{1,1}_{\mu,\lambda}\frac{\frac{\sigma^\star}{N}(1+\alpha^2d^2R^{2\alpha-2}) + \alpha^2d^2R^{2\alpha-2}\frac{N}{R^2}\frac{\sigma^{\star 3}}{N^3}}{\Big(\frac{\sigma^{\star 2}}{N^2}(1+\alpha^2d^2R^{2\alpha-2}) + \alpha^2d^2R^{2\alpha-2}\frac{N}{2R^2}\frac{\sigma^{\star 4}}{N^4}\Big)} - c_{\mu/\mu,\lambda}\frac{\alpha dR^{\alpha-1}\frac{N}{R}\frac{\sigma^\star}{N}}{\sqrt{\frac{\sigma^{\star 2}}{N^2}(1+\alpha^2d^2R^{2\alpha-2}) + \alpha^2d^2R^{2\alpha-2}\frac{N}{2R^2}\frac{\sigma^{\star 4}}{N^4}}}\Big)\Big) + \mathcal{O}(\tau^4) \quad \text{(C.41)}$$

$$= \tau^2\Bigg(\frac{1}{2} + e^{1,1}_{\mu,\lambda}\frac{\sigma^{\star 2}(1+\alpha^2 d^2 R^{2\alpha-2}) + \alpha^2 d^2 R^{2\alpha-2}\frac{\sigma^{\star 4}}{NR^2}}{\left(\sigma^{\star 2}(1+\alpha^2 d^2 R^{2\alpha-2}) + \alpha^2 d^2 R^{2\alpha-2}\frac{\sigma^{\star 4}}{2R^2 N}\right)} - c_{\mu/\mu,\lambda}\frac{\alpha d R^{\alpha-1}\frac{\sigma^{\star 2}}{R}}{\sqrt{\sigma^{\star 2}(1+\alpha^2 d^2 R^{2\alpha-2}) + \alpha^2 d^2 R^{2\alpha-2}\frac{\sigma^{\star 4}}{2R^2 N}}}\Bigg) + \mathcal{O}(\tau^4). \quad \text{(C.42)}$$

Letting $N \to \infty$,

$$\begin{aligned}\psi(\sigma^\star) &= \tau^2\Big(\frac{1}{2} + e^{1,1}_{\mu,\lambda}\frac{\sigma^{\star 2}(1+\alpha^2 d^2 R^{2\alpha-2})}{\sigma^{\star 2}(1+\alpha^2 d^2 R^{2\alpha-2})} - c_{\mu/\mu,\lambda}\frac{\alpha d R^{\alpha-1}\sigma^{\star 2}}{R\sqrt{\sigma^{\star 2}(1+\alpha^2 d^2 R^{2\alpha-2})}}\Big) + \mathcal{O}(\tau^4)\\ &= \tau^2\Big(\frac{1}{2} + e^{1,1}_{\mu,\lambda} - c_{\mu/\mu,\lambda}\frac{\alpha d R^{\alpha-1}\sigma^\star}{R\sqrt{1+\alpha^2 d^2 R^{2\alpha-2}}}\Big) + \mathcal{O}(\tau^4) \quad \text{(C.43)}\end{aligned}$$

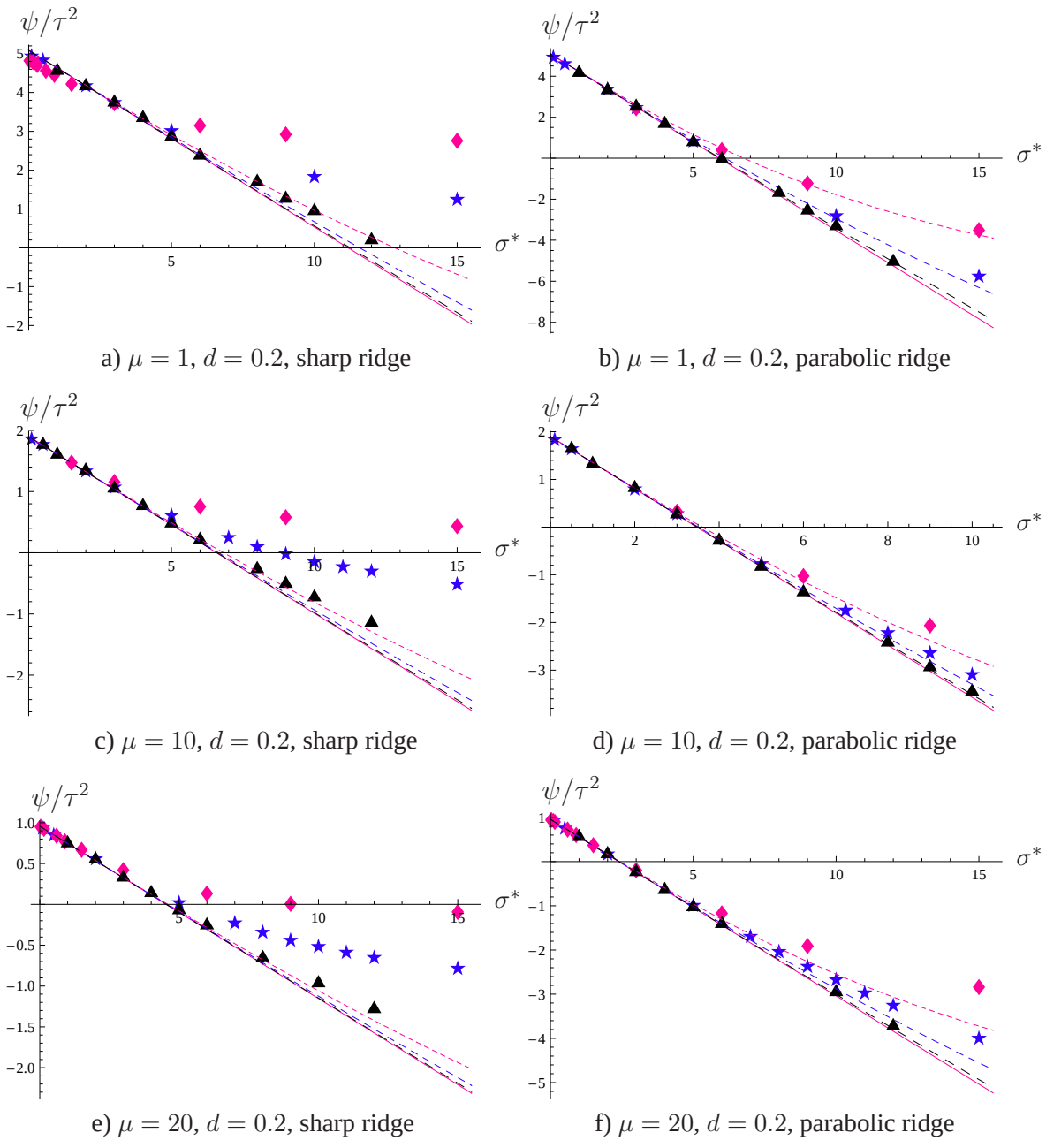
is obtained. Let us summarize the conditions under which (C.42) and (C.43) were derived: The general SAR requires the learning rate τ (or the parameter β, respectively) to be small. The cdf of the fitness change was obtained for large value N. Therefore, (C.42) only holds in large dimensional search spaces. Finally, (C.43) is obtained for $N \to \infty$.

It remains to compare both SARs (C.42) and (C.43) with the results of experiments (see Fig. C.6). For the experiments, a $(1, 60)$-ES, a $(10/10_I, 60)$-ES, and a $(30/30_I, 60)$-ES were chosen and run on the sharp ($\alpha = 1$) and parabolic ridge ($\alpha = 2$). The learning rate was set to $\tau = 1/\sqrt{N}$ and the d-constant was set to $d = 0.2$. The starting vectors $\mathbf{y}^{\mathbf{m}}_{\mathbf{0}}$ were randomly chosen and normalized to $\|\mathbf{y}^{\mathbf{m}}_{\mathbf{0}}\| = 1$.

In the case of the parabolic ridge, the prediction quality is good. This even holds for the low dimensional search space ($N = 30$). In the case of the sharp ridge, considerable deviations can be found for $N = 30$. This is especially true for the $(1, 60)$-ES which deviates very soon from the values predicted by (C.43). Smaller deviations can be observed for the experiments in the higher dimensional search space ($N = 500$). It should be noted that the N-dependent (C.43) also fails to capture the exact behavior of the measured SAR. The prediction quality is far better in the case of the parabolic ridge. Several causes may contribute to the behavior of the SAR. First, higher-order terms of τ were neglected during the derivation of the SAR. Second, the derivation of the fitness gain relied on the assumption that the changes of the components of mutation vector are small w.r.t. the distance. This allowed the Taylor expansion of the square root in Eq. (B.31), p. 133 and the subsequent cutting off of the series after the linear term. The learning rate for $N = 30$ and $N = 100$ may thus contribute together with the N-dependent terms to the deviation. It should be noted that in the case of the parabolic ridge this smallness assumption is not required. This is also stressed by considering the prediction quality of the SAR in [28] which used an alternative fitness change for $(1, \lambda)$-ES. This alternative fitness change resulted in a better prediction quality for smaller N. Therefore in the next section an alternative derivation for the SAR is given.

The SAR for Noisy Ridge Functions In the case of noisy ridge functions, the pdf of fitness change is given by

$$P_Q(Q|\varsigma) = \Phi\Bigg(\frac{Q + \alpha d R^{\alpha-1}\frac{N}{2R}\varsigma^2}{\sqrt{\varsigma^2(1+\alpha^2 d^2 R^{2\alpha-2}) + \sigma_\epsilon^2 + \alpha^2 d^2 R^{2\alpha-2}\frac{N}{2R^2}\varsigma^4}}\Bigg). \quad \text{(C.44)}$$



a) $\mu = 1$, $d = 0.2$, sharp ridge

b) $\mu = 1$, $d = 0.2$, parabolic ridge

c) $\mu = 10$, $d = 0.2$, sharp ridge

d) $\mu = 10$, $d = 0.2$, parabolic ridge

e) $\mu = 20$, $d = 0.2$, sharp ridge

f) $\mu = 20$, $d = 0.2$, parabolic ridge

Figure C.6: The first-order SAR Eq. (C.42) (dashed lines) and Eq. (C.43) (solid lines) on the sharp and parabolic ridge for some $(\mu/\mu_I, 60)$-ES. Shown are the results for $\mu = 1$, $\mu = 10$, and $\mu = 20$. The distance to the ridge was set to $R = 1$. Each data point was obtained by sampling over $100,000$ one-generation experiments for $N = 30$, $200,000$ for $N = 100$, and $250,000$ for $N = 500$. The results for $N = 30$ are denoted by diamond shaped symbols (red), whereas stars (blue) stand for $N = 100$, and triangles (black) for $N = 500$.

Considering the general form of the SAR (C.23), the functions h and g and their derivatives are

$$h(\varsigma) = \alpha d R^{\alpha-1} \frac{N}{2R} \varsigma^2, \quad h'(\varsigma) = \alpha d R^{\alpha-1} \frac{N}{R} \varsigma \tag{C.45}$$

and

$$\begin{aligned} g(\varsigma) &= \sqrt{\varsigma^2(1+\alpha^2d^2R^{2\alpha-2})+\sigma_\epsilon^2+\alpha^2d^2R^{2\alpha-2}\frac{N}{2R^2}\varsigma^4}, \\ g'(\varsigma) &= \frac{2\varsigma(1+\alpha^2d^2R^{2\alpha-2})+2\alpha^2d^2R^{2\alpha-2}\frac{N}{R^2}\varsigma^3}{2\sqrt{\varsigma^2(1+\alpha^2d^2R^{2\alpha-2})+\sigma_\epsilon^2+\alpha^2d^2R^{2\alpha-2}\frac{N}{2R^2}\varsigma^4}}. \end{aligned} \tag{C.46}$$

As one can easily see, the rest of the steps in obtaining the SAR are entirely analogous to the noise-free case. Inserting (C.45) and (C.46) into the SAR (C.23) leads to

$$\begin{aligned} \psi(\langle\sigma\rangle) &= \tau^2\Big(\frac{1}{2}+\langle\langle\sigma\rangle\rangle\Big(e_{\mu,\lambda}^{1,1}\frac{\langle\sigma\rangle(1+\alpha^2d^2R^{2\alpha-2})+\alpha^2d^2R^{2\alpha-2}\frac{N}{R^2}\langle\sigma\rangle^3}{\Big(\langle\sigma\rangle^2(1+\alpha^2d^2R^{2\alpha-2})+\sigma_\epsilon^2+\alpha^2d^2R^{2\alpha-2}\frac{N}{2R^2}\langle\sigma\rangle^4\Big)} \\ &\quad -\frac{c_{\mu/\mu,\lambda}\alpha dR^{\alpha-1}\frac{N}{R}\langle\sigma\rangle}{\sqrt{\langle\sigma\rangle^2(1+\alpha^2d^2R^{2\alpha-2})+\sigma_\epsilon^2+\alpha^2d^2R^{2\alpha-2}\frac{N}{2R^2}\langle\sigma\rangle^4}}\Big)\Big)+\mathcal{O}(\tau^4). \end{aligned} \tag{C.47}$$

Now the same normalization as before is introduced – setting $\sigma^\star := N\langle\sigma\rangle$ and $\sigma_\epsilon^\star := N\sigma_\epsilon$. The SAR (C.47) changes to

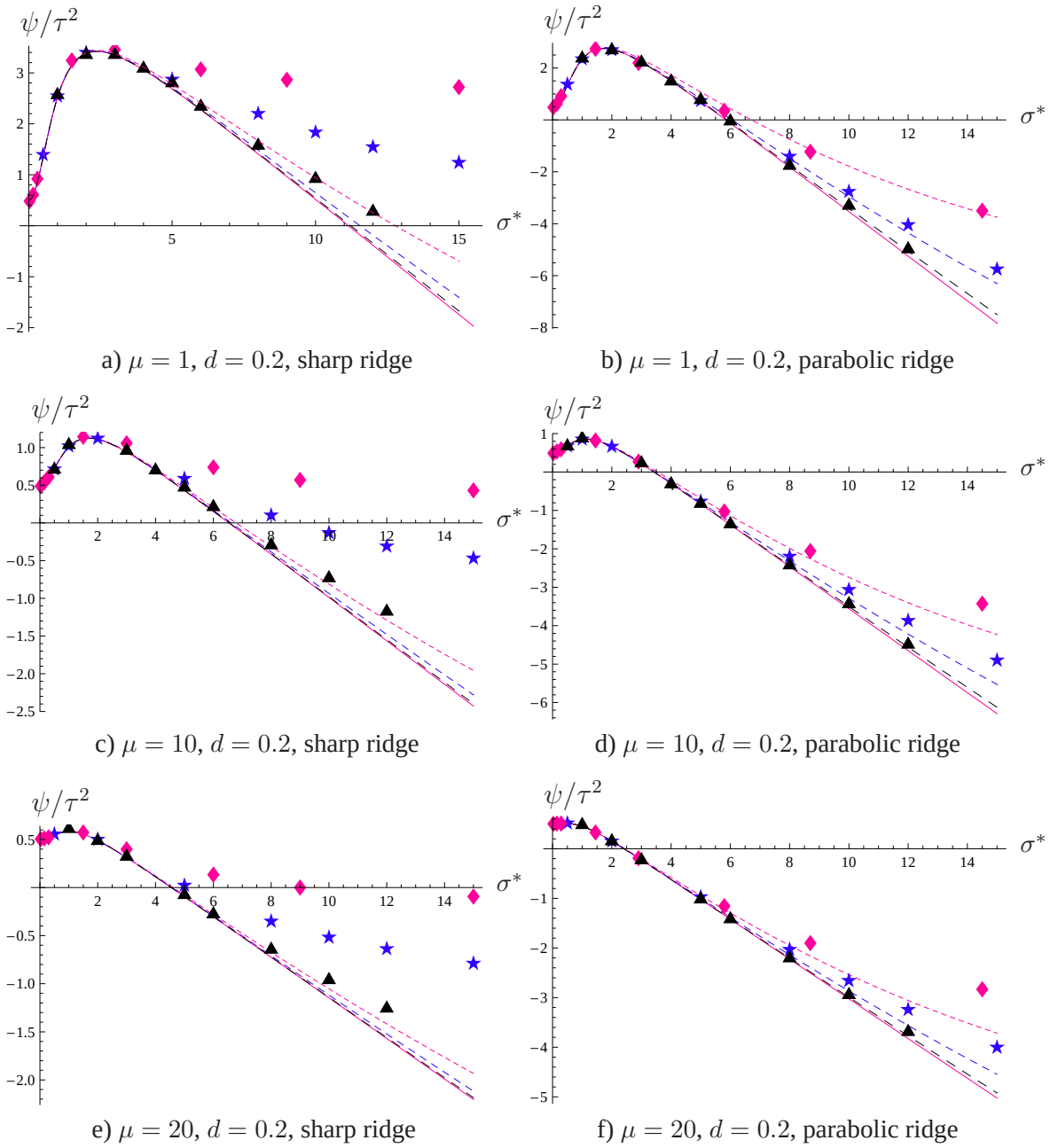
$$\begin{aligned} \psi(\sigma^\star) &= \tau^2\Big(\frac{1}{2}+\frac{\sigma^\star}{N}\Big(e_{\mu,\lambda}^{1,1}\frac{\frac{\sigma^\star}{N}(1+\alpha^2d^2R^{2\alpha-2})+\alpha^2d^2R^{2\alpha-2}\frac{N}{R^2}\frac{\sigma^{\star3}}{N^3}}{\Big(\frac{\sigma^{\star2}}{N^2}(1+\alpha^2d^2R^{2\alpha-2})+\frac{\sigma_\epsilon^{\star2}}{N^2}+\alpha^2d^2R^{2\alpha-2}\frac{N}{2R^2}\frac{\sigma^{\star4}}{N^4}\Big)} \\ &\quad -c_{\mu/\mu,\lambda}\frac{\alpha dR^{\alpha-1}\frac{N}{R}\frac{\sigma^\star}{N}}{\sqrt{\frac{\sigma^{\star2}}{N^2}(1+\alpha^2d^2R^{2\alpha-2})+\frac{{\sigma_\epsilon}^2}{N^2}+\alpha^2d^2R^{2\alpha-2}\frac{N}{2R^2}\frac{\sigma^{\star4}}{N^4}}}\Big)\Big)+\mathcal{O}(\tau^4) \\ &= \tau^2\Big(\frac{1}{2}+e_{\mu,\lambda}^{1,1}\frac{\sigma^{\star2}(1+\alpha^2d^2R^{2\alpha-2})+\alpha^2d^2R^{2\alpha-2}\frac{\sigma^{\star4}}{R^2N}}{\Big(\sigma^{\star2}(1+\alpha^2d^2R^{2\alpha-2})+\sigma_\epsilon^{\star2}+\alpha^2d^2R^{2\alpha-2}\frac{\sigma^{\star4}}{2R^2N}\Big)} \\ &\quad -c_{\mu/\mu,\lambda}\frac{\alpha dR^{\alpha-1}\frac{\sigma^{\star2}}{R}}{\sqrt{\sigma^{\star2}(1+\alpha^2d^2R^{2\alpha-2})+{\sigma_\epsilon}^2+\alpha^2d^2R^{2\alpha-2}\frac{\sigma^{\star4}}{2R^2N}}}\Big)+\mathcal{O}(\tau^4). \end{aligned} \tag{C.48}$$

Computing the limes $N\to\infty$ of (C.48) gives

$$\begin{aligned} \psi(\sigma^\star) &= \tau^2\Big(\frac{1}{2}+e_{\mu,\lambda}^{1,1}\frac{\sigma^{\star2}(1+\alpha^2d^2R^{2\alpha-2})}{\sigma^{\star2}(1+\alpha^2d^2R^{2\alpha-2})+\sigma_\epsilon^{\star2}} \\ &\quad -c_{\mu/\mu,\lambda}\frac{\alpha dR^{\alpha-1}\sigma^{\star2}}{R\sqrt{\sigma^{\star2}(1+\alpha^2d^2R^{2\alpha-2})+\sigma_\epsilon^{\star2}}}\Big)+\mathcal{O}(\tau^4). \end{aligned} \tag{C.49}$$

The conditions under which (C.48) and (C.49) were derived are the same as in the case of the noise-free ridge: The learning rate τ (or the parameter β, respectively) has to be small and the search space must be high-dimensional.

Both SARs (C.48) and (C.49) are compared with the results of experiments in Figure C.7. The set-up of the experiments is nearly the same as in the noise-free case. The noise strengths was set to $\sigma_\epsilon = 1$ for $N = 100$ and to $\sigma_\epsilon = 0.33$ for $N = 30$. The learning rate was set to $\tau = 1/\sqrt{N}$. In the case of the parabolic ridge, the prediction quality is reasonably good – even in the low dimensional search space ($N = 30$). In the case of the sharp ridge, the prediction quality is only good for higher dimensional search spaces.



a) $\mu = 1$, $d = 0.2$, sharp ridge

b) $\mu = 1$, $d = 0.2$, parabolic ridge

c) $\mu = 10$, $d = 0.2$, sharp ridge

d) $\mu = 10$, $d = 0.2$, parabolic ridge

e) $\mu = 20$, $d = 0.2$, sharp ridge

f) $\mu = 20$, $d = 0.2$, parabolic ridge

Figure C.7: The first-order SAR (C.48) (dashed lines) and (C.49) (solid lines) on the sharp and parabolic ridge for some $(\mu/\mu_I, 60)$-ES. Shown are the results for $\mu = 1$, $\mu = 10$, and $\mu = 20$. The distance to the ridge was set to $R = 1$. Each data point was obtained by sampling over $100,000$ one-generation experiments for $N = 30$, $200,000$ for $N = 100$, and $250,000$ for $N = 500$. The results for $N = 30$ are denoted by diamond shaped symbols (red), whereas stars (blue) stand for $N = 100$, and triangles (black) for $N = 500$.

C.1.3 Ridge Functions: An Alternative Derivation of the Self-Adaptation Response Function for the Sharp Ridge

In this section, an alternative fitness change is used to determined the SAR for the sharp ridge. This fitness change uses a different approach to give the density of r [23, p.111]. The starting point is the Taylor series expansion of $f(r) = r^\alpha$ in B.2.1. Again, the Taylor series around R, $T_f(r) = R^\alpha + \alpha R^{\alpha-1}(r-R) + \mathcal{O}[(r-R)^2]$, is cut off after the linear term. This leads to the fitness change

$$\begin{aligned} Q &= z_x + dR^\alpha - dR\alpha - d\alpha R^{\alpha-1}(r-R) - \mathcal{O}[(r-R)^2] + \epsilon \\ &= z_x - d\alpha R^{\alpha-1}(r-R) - \mathcal{O}[(r-R)^2] + \epsilon. \end{aligned} \tag{C.50}$$

Similar to (B.29), p. 132, z_x denotes the change in the first component of the vector, whereas ϵ stands for the noise term. Following [22], a normal approximation for the pdf of r

$$p(r|\varsigma) = \frac{\exp\left(-\frac{1}{2}\left(\frac{r-\sqrt{R^2+\varsigma^2 N}}{\varsigma\sqrt{\frac{R^2+\frac{\varsigma^2 N}{2}}{R^2+\varsigma^2 N}}}\right)\right)}{\sqrt{2\pi}\varsigma\sqrt{\frac{R^2+\frac{\varsigma^2 N}{2}}{R^2+\varsigma^2 N}}} \tag{C.51}$$

is used [23, p.111]. This results in the following cdf of the fitness change

$$P_Q(Q|\varsigma) = \Phi\left(\frac{Q + \alpha d R^{\alpha-1}(\sqrt{R^2+\varsigma^2 N} - R)}{\sqrt{\varsigma^2 + \sigma_\epsilon^2 + \alpha^2 d^2 R^{2\alpha-2}\varsigma^2\left(\frac{R^2+\varsigma^2 N/2}{R^2+\varsigma^2 N}\right)}}\right). \tag{C.52}$$

Recall, the SAR is given by (C.23)

$$\psi(\langle\sigma\rangle) = \tau^2\left(\frac{1}{2} + \langle\sigma\rangle\left(e^{1,1}_{\mu,\lambda}\frac{g'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)} - c_{\mu/\mu,\lambda}\frac{h'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}\right)\right) + \mathcal{O}(\tau^4)$$

with g and h stemming from the cdf of the fitness change of the form $P_Q(Q|\varsigma) = P((Q+h(\varsigma))/g(\varsigma))$.

Again, first undisturbed ridge functions are considered before the SAR for noisy ridge functions is derived.

The Undisturbed Ridge In the case of the undisturbed ridge, the variance of the noise term in (C.52) is zero yielding

$$P_Q(Q|\varsigma) = \Phi\left(\frac{Q + \alpha d R^{\alpha-1}(\sqrt{R^2+\varsigma^2 N} - R)}{\sqrt{\varsigma^2 + \alpha^2 d^2 R^{2\alpha-2}\varsigma^2\left(\frac{R^2+\varsigma^2 N/2}{R^2+\varsigma^2 N}\right)}}\right). \tag{C.53}$$

The functions needed for the SAR (C.23) are

$$\begin{aligned} h(\varsigma) &= \alpha d R^{\alpha-1}(\sqrt{R^2+\varsigma^2 N} - R) \\ h'(\varsigma) &= \alpha d R^{\alpha-1}\frac{N\varsigma}{\sqrt{R^2+\varsigma^2 N}} \end{aligned} \tag{C.54}$$

and

$$
\begin{aligned}
g(\varsigma) &= \sqrt{\varsigma^2 + \alpha^2 d^2 R^{2\alpha-2}\varsigma^2\Big(\frac{R^2+\varsigma^2 N/2}{R^2+\varsigma^2 N}\Big)} \\
g'(\varsigma) &= \frac{2\varsigma + 2\alpha^2 d^2 R^{2\alpha-2}\varsigma\Big(\frac{R^2+\varsigma^2 N/2}{R^2+\varsigma^2 N}\Big) + \alpha^2 d^2 R^{2\alpha-2}\varsigma^2\Big(\frac{N\varsigma(R^2+\varsigma^2 N)-2N\varsigma(R^2+\varsigma^2 N/2)}{\big(R^2+\varsigma^2 N\big)^2}\Big)}{2\sqrt{\varsigma^2 + \alpha^2 d^2 R^{2\alpha-2}\varsigma^2\Big(\frac{R^2+\varsigma^2 N/2}{R^2+\varsigma^2 N}\Big)}} \\
&= \frac{2\varsigma\Big[1+\alpha^2 d^2 R^{2\alpha-2}\Big(\frac{R^2+\varsigma^2 N/2}{R^2+\varsigma^2 N}\Big)\Big] + \alpha^2 d^2 R^{2\alpha-2} N\varsigma^3\Big(\frac{R^2+\varsigma^2 N-2R^2-\varsigma^2 N}{\big(R^2+\varsigma^2 N\big)^2}\Big)}{2\sqrt{\varsigma^2 + \alpha^2 d^2 R^{2\alpha-2}\varsigma^2\Big(\frac{R^2+\varsigma^2 N/2}{R^2+\varsigma^2 N}\Big)}} \\
&= \frac{2\varsigma\Big[1+\alpha^2 d^2 R^{2\alpha-2}\Big(\frac{R^2+\varsigma^2 N/2}{R^2+\varsigma^2 N}\Big)\Big] - \alpha^2 d^2 R^{2\alpha-2} N\varsigma^3\Big(\frac{R^2}{\big(R^2+\varsigma^2 N\big)^2}\Big)}{2\sqrt{\varsigma^2 + \alpha^2 d^2 R^{2\alpha-2}\varsigma^2\Big(\frac{R^2+\varsigma^2 N/2}{R^2+\varsigma^2 N}\Big)}}.
\end{aligned}
\tag{C.55}
$$

Plugging (C.54) and (C.55) into the SAR (C.23) gives

$$
\begin{aligned}
\psi(\langle\sigma\rangle) &= \tau^2\Bigg(\frac{1}{2} + \langle\sigma\rangle\Bigg(- c_{\mu/\mu,\lambda}\frac{\alpha d R^{\alpha-1} N\langle\sigma\rangle}{\sqrt{R^2+N\langle\sigma\rangle}\sqrt{\langle\sigma\rangle^2 + \alpha^2 d^2 R^{2\alpha-2}\langle\sigma\rangle^2\Big(\frac{R^2+\varsigma^2 N/2}{R^2+\langle\sigma\rangle^2 N}\Big)}} \\
&\quad + e^{1,1}_{\mu,\lambda} 2\langle\sigma\rangle\frac{1+\alpha^2 d^2 R^{2\alpha-2}\Big(\frac{R^2+\langle\sigma\rangle^2 N/2}{R^2+\langle\sigma\rangle^2 N}\Big)}{2\Bigg(\langle\sigma\rangle^2 + \alpha^2 d^2 R^{2\alpha-2}\langle\sigma\rangle^2\Big(\frac{R^2+\langle\sigma\rangle^2 N/2}{R^2+\langle\sigma\rangle^2 N}\Big)\Bigg)} \\
&\quad - e^{1,1}_{\mu,\lambda}\frac{\alpha^2 d^2 R^{2\alpha-2} N\langle\sigma\rangle^3\Big(\frac{R^2}{\big(R^2+\langle\sigma\rangle^2 N\big)^2}\Big)}{2\Bigg(\langle\sigma\rangle^2 + \alpha^2 d^2 R^{2\alpha-2}\langle\sigma\rangle^2\Big(\frac{R^2+\langle\sigma\rangle^2 N/2}{R^2+\langle\sigma\rangle^2 N}\Big)\Bigg)}\Bigg)\Bigg) + \mathcal{O}(\tau^4) \\
&= \tau^2\Bigg(\frac{1}{2} + e^{1,1}_{\mu,\lambda} - c_{\mu/\mu,\lambda}\frac{\alpha d R^{\alpha-1} N\langle\sigma\rangle^2}{\sqrt{R^2+N\langle\sigma\rangle}\sqrt{\langle\sigma\rangle^2 + \alpha^2 d^2 R^{2\alpha-2}\langle\sigma\rangle^2\Big(\frac{R^2+\varsigma^2 N/2}{R^2+\langle\sigma\rangle^2 N}\Big)}} \\
&\quad - e^{1,1}_{\mu,\lambda}\frac{\alpha^2 d^2 R^{2\alpha-2} N\langle\sigma\rangle^4\Big(\frac{R^2}{\big(R^2+\langle\sigma\rangle^2 N\big)^2}\Big)}{2\Bigg(\langle\sigma\rangle^2 + \alpha^2 d^2 R^{2\alpha-2}\langle\sigma\rangle^2\Big(\frac{R^2+\langle\sigma\rangle^2 N/2}{R^2+\langle\sigma\rangle^2 N}\Big)\Bigg)}\Bigg) + \mathcal{O}(\tau^4).
\end{aligned}
\tag{C.56}
$$

Using the same normalization as before, i.e., $\sigma^\star := N\langle\sigma\rangle$, the SAR changes to

$$
\begin{aligned}
\psi(\sigma^\star) &= \tau^2\Bigg(\frac{1}{2} + e^{1,1}_{\mu,\lambda} - c_{\mu/\mu,\lambda}\frac{\alpha d R^{\alpha-1}\sigma^{\star 2}}{\sqrt{R^2+\frac{\sigma^{\star 2}}{N}}\sqrt{\sigma^{\star 2} + \alpha^2 d^2 R^{2\alpha-2}\sigma^{\star 2}\Big(\frac{R^2+\frac{\sigma^{\star 2}}{2N}}{R^2+\frac{\sigma^{\star 2}}{N}}\Big)}} \\
&\quad - e^{1,1}_{\mu,\lambda}\frac{\alpha^2 d^2 R^{2\alpha-2}\frac{\sigma^{\star 4}}{N}\Big(\frac{R^2}{\big(R^2+\frac{\sigma^{\star 2}}{2N}\big)^2}\Big)}{2\Bigg(\sigma^{\star 2} + \alpha^2 d^2 R^{2\alpha-2}\sigma^{\star 2}\Big(\frac{R^2+\frac{\varsigma^{\star 2}}{2N}}{R^2+\frac{\sigma^{\star 2}}{N}}\Big)\Bigg)}\Bigg) + \mathcal{O}(\tau^4).
\end{aligned}
\tag{C.57}
$$

Letting $N \to \infty$,

$$\psi(\sigma^\star) \;=\; \tau^2\Big(\frac{1}{2} + e^{1,1}_{\mu,\lambda} - c_{\mu/\mu,\lambda}\frac{\alpha d R^{\alpha-1}\sigma^\star}{R\sqrt{1+\alpha^2 d^2 R^{2\alpha-2}}}\Big) + \mathcal{O}(\tau^4) \tag{C.58}$$

is obtained which equals (C.43). The conditions for the derivation of (C.57) and (C.58) are the same as for (C.42) and (C.43): A small learning rate τ and a large value of N in the case of (C.57).

In the following, the SARs (C.57) and (C.58) are compared with the results of experiments. The experiments were already described in the previous section. As expected, the prediction quality of the SAR (C.57) is not good in the case of the parabolic ridge (see Fig. C.8). This is due to the derivation of the fitness change (C.52). Equation (C.57) agrees very well with the experiments in the case of the sharp ridge.

The SAR for Noisy Ridge Functions In the case of noisy ridge functions, the pdf of fitness change is given by

$$P_Q(Q|\varsigma) \;=\; \Phi\left(\frac{Q + \alpha d R^{\alpha-1}(\sqrt{R^2+\varsigma^2 N} - R)}{\sqrt{\varsigma^2 + \sigma_\epsilon^2 + \alpha^2 d^2 R^{2\alpha-2}\varsigma^2\Big(\frac{R^2+\varsigma^2 N/2}{R^2+\varsigma^2 N}\Big)}}\right). \tag{C.59}$$

Considering the general form of the SAR (C.23), the functions h and g and their derivatives are

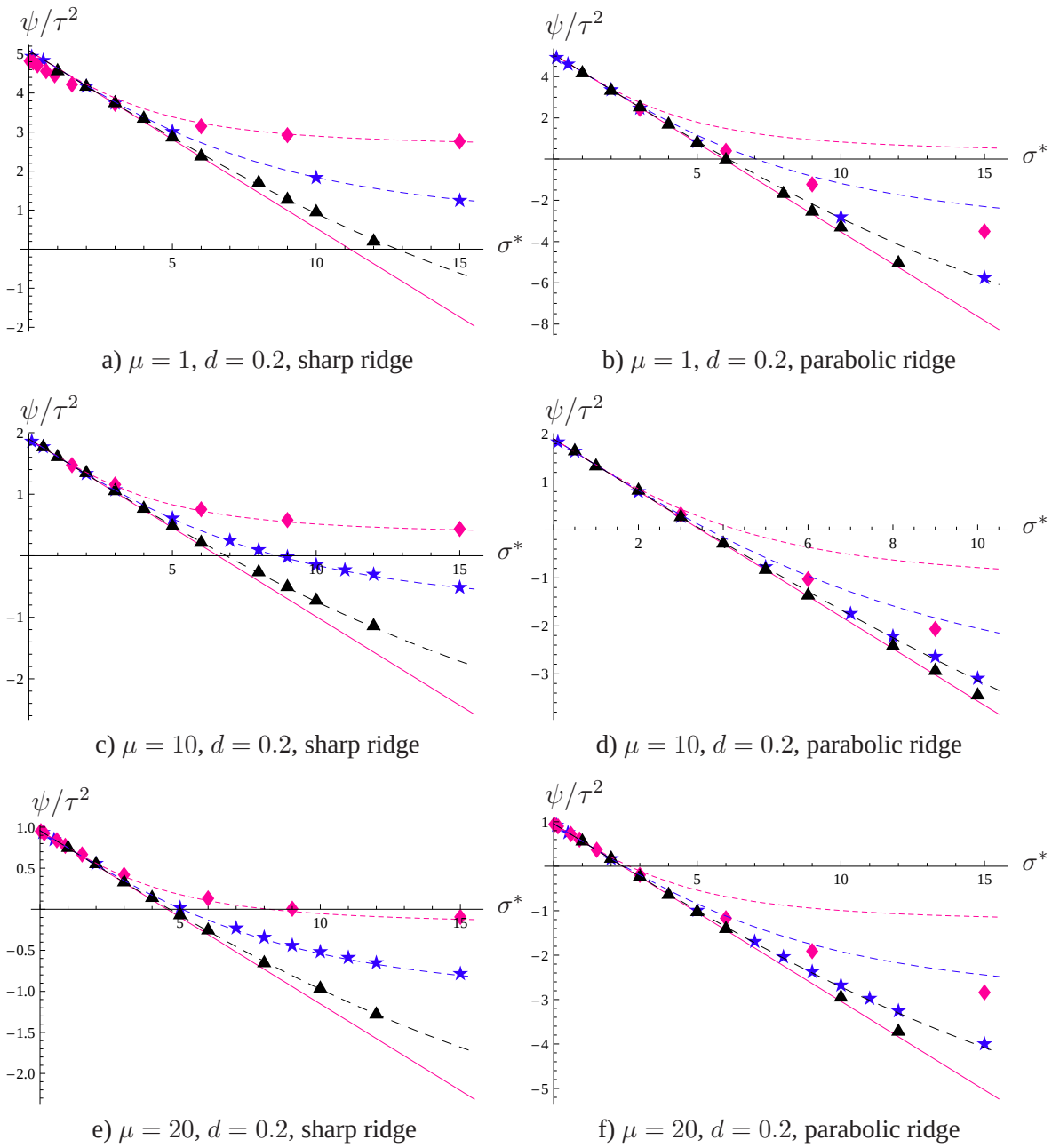
$$\begin{aligned} h(\varsigma) &= \alpha d R^{\alpha-1}(\sqrt{R^2+\varsigma^2 N} - R) \\ h'(\varsigma) &= \alpha d R^{\alpha-1}\frac{\varsigma N}{\sqrt{R^2+\varsigma^2 N}} \end{aligned} \tag{C.60}$$

and

$$\begin{aligned} g(\varsigma) &= \sqrt{\varsigma^2 + \sigma_\epsilon^2 + \alpha^2 d^2 R^{2\alpha-2}\varsigma^2\Big(\frac{R^2+\varsigma^2 N/2}{R^2+\varsigma^2 N}\Big)} \\ g'(\varsigma) &= \frac{2\varsigma\Big[1 + \alpha^2 d^2 R^{2\alpha-2}\Big(\frac{R^2+\varsigma^2 N/2}{R^2+\varsigma^2 N}\Big)\Big] - \alpha^2 d^2 R^{2\alpha-2} N\varsigma^3\Big(\frac{R^2}{\left(R^2+\varsigma^2 N\right)^2}\Big)}{2\sqrt{\sigma_\epsilon^2 + \varsigma^2 + \alpha^2 d^2 R^{2\alpha-2}\varsigma^2\Big(\frac{R^2+\varsigma^2 N/2}{R^2+\varsigma^2 N}\Big)}}. \end{aligned} \tag{C.61}$$

As one can easily see, the rest of the steps in obtaining the SAR are entirely analogous to the noise-free case. Inserting (C.61) into the SAR (C.23) leads to

$$\begin{aligned} \psi(\langle\sigma\rangle) \;=\; \tau^2\Bigg(&\frac{1}{2} - c_{\mu/\mu,\lambda}\alpha d R^{\alpha-1}\frac{\langle\sigma\rangle^2 N}{\sqrt{R^2+\langle\sigma\rangle^2 N}\sqrt{\langle\sigma\rangle^2 + \sigma_\epsilon^2 + \alpha^2 d^2 R^{2\alpha-2}\langle\sigma\rangle^2\Big(\frac{R^2+\langle\sigma\rangle^2 N/2}{R^2+\langle\sigma\rangle^2 N}\Big)}} \\ &+ 2e^{1,1}_{\mu,\lambda}\langle\sigma\rangle^2\frac{1 + \alpha^2 d^2 R^{2\alpha-2}\Big(\frac{R^2+\langle\sigma\rangle^2 N/2}{R^2+\langle\sigma\rangle^2 N}\Big)}{2\Big(\sigma_\epsilon^2 + \langle\sigma\rangle^2 + \alpha^2 d^2 R^{2\alpha-2}\langle\sigma\rangle^2\Big(\frac{R^2+\langle\sigma\rangle^2 N/2}{R^2+\langle\sigma\rangle^2 N}\Big)\Big)} \\ &- e^{1,1}_{\mu,\lambda}\frac{\alpha^2 d^2 R^{2\alpha-2} N\langle\sigma\rangle^4\Big(\frac{R^2}{\left(R^2+\langle\sigma\rangle^2 N\right)^2}\Big)}{2\Big(\sigma_\epsilon^2 + \langle\sigma\rangle^2 + \alpha^2 d^2 R^{2\alpha-2}\langle\sigma\rangle^2\Big(\frac{R^2+\langle\sigma\rangle^2 N/2}{R^2+\langle\sigma\rangle^2 N}\Big)\Big)}\Bigg) + \mathcal{O}(\tau^4). \end{aligned} \tag{C.62}$$



a) $\mu = 1$, $d = 0.2$, sharp ridge

b) $\mu = 1$, $d = 0.2$, parabolic ridge

c) $\mu = 10$, $d = 0.2$, sharp ridge

d) $\mu = 10$, $d = 0.2$, parabolic ridge

e) $\mu = 20$, $d = 0.2$, sharp ridge

f) $\mu = 20$, $d = 0.2$, parabolic ridge

Figure C.8: The first-order SAR Eq. (C.57) (dashed lines) and Eq. (C.58) (solid lines) for some $(\mu/\mu_I, 60)$-ES. Shown are the results for $\mu = 1$, $\mu = 10$, and $\mu = 20$. The distance to the ridge was set to $R = 1$. Each data point was obtained by sampling over $100,000$ one-generation experiments for $N = 30$, $200,000$ for $N = 100$, and $250,000$ for $N = 500$. The results for $N = 30$ are denoted by diamond shaped symbols (red), whereas stars (blue) stand for $N = 100$, and triangles (black) for $N = 500$.

Now the same normalizations as before are introduced – setting $\sigma^\star := N\langle\sigma\rangle$ and $\sigma_\epsilon^\star := N\sigma_\epsilon$. The SAR (C.62) changes to

$$\psi(\sigma^\star) \quad = \quad \tau^2\left(\frac{1}{2} - c_{\mu/\mu,\lambda}\alpha d R^{\alpha-1}\frac{\sigma^{\star 2}}{\sqrt{R^2+\frac{\sigma^{\star 2}}{N}}\sqrt{\sigma^{\star 2}+\sigma_\epsilon^{\star 2}+\alpha^2 d^2 R^{2\alpha-2}\sigma^{\star 2}\left(\frac{R^2+\frac{\sigma^{\star 2}}{2N}}{R^2+\frac{\sigma^{\star 2}}{N}}\right)}}\right)$$

$$+e^{1,1}_{\mu,\lambda}\sigma^{\star 2}\frac{1+\alpha^2 d^2 R^{2\alpha-2}\left(\frac{R^2+\frac{\sigma^{\star 2}}{2N}}{R^2+\frac{\sigma^{\star 2}}{N}}\right)}{\left(\sigma_\epsilon^{\star 2}+\sigma^{\star 2}+\alpha^2 d^2 R^{2\alpha-2}\sigma^{\star 2}\left(\frac{R^2+\frac{\sigma^{\star 2}}{2N}}{R^2+\frac{\sigma^{\star 2}}{N}}\right)\right)}$$
$$-e^{1,1}_{\mu,\lambda}\frac{\alpha^2 d^2 R^{2\alpha-2}\frac{\sigma^{\star 4}}{N}\left(\frac{R^2}{\left(R^2+\frac{\sigma^{\star 2}}{N}\right)^2}\right)}{2\left(\sigma_\epsilon^{\star 2}+\sigma^{\star 2}+\alpha^2 d^2 R^{2\alpha-2}\sigma^{\star 2}\left(\frac{R^2+\frac{\sigma^{\star 2}}{2N}}{R^2+\frac{\sigma^{\star 2}}{N}}\right)\right)}\Bigg) + \mathcal{O}(\tau^4). \tag{C.63}$$

Letting $N \to \infty$ gives

$$\begin{aligned}\psi(\sigma^\star) &= \tau^2\Big(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\frac{\sigma^{\star 2}(1+\alpha^2 d^2 R^{2\alpha-2})}{\sigma^{\star 2}(1+\alpha^2 d^2 R^{2\alpha-2})+\sigma_\epsilon^{\star 2}} \\ &\quad -c_{\mu/\mu,\lambda}\frac{\alpha d R^{\alpha-1}\sigma^{\star 2}}{R\sqrt{\sigma^{\star 2}(1+\alpha^2 d^2 R^{2\alpha-2})+\sigma_\epsilon^{\star 2}}}\Big)+\mathcal{O}(\tau^4).\end{aligned} \tag{C.64}$$

Again, the conditions under which (C.63) and (C.64) were derived are the same as in the case of (C.48) and (C.49)

Figure C.9 shows a comparison of (C.63) and (C.64) with the results of experiments. Again, there is a good agreement of (C.63) with the experiments in the case of the sharp ridge. In the case of the parabolic ridge, the same observations can be made as in the case of the undisturbed ridge: Due to the derivation of (C.64), (C.49) serves better to predict the experiments.

C.2 Calculating the Expectation

In this section, the expectations of

$$\mathrm{E}\Big[\Big(\frac{\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)^k\Big] = \int_0^\infty \Big(\frac{\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)^k p_\sigma(\varsigma|\langle\sigma\rangle)\,\mathrm{d}\varsigma \tag{C.65}$$

are determined for the log-normal and the symmetric two-point operator. In Section C.1 it was claimed that if the Taylor series in τ^{2k} is cut off after the $(n+1)$th summand, the expectation of (C.65) with degree $\geq 2n+1$ is zero. In this section, this claim is verified. First, the log-normal distribution is considered, before the case of the two-point distribution is discussed for the sake of completeness.

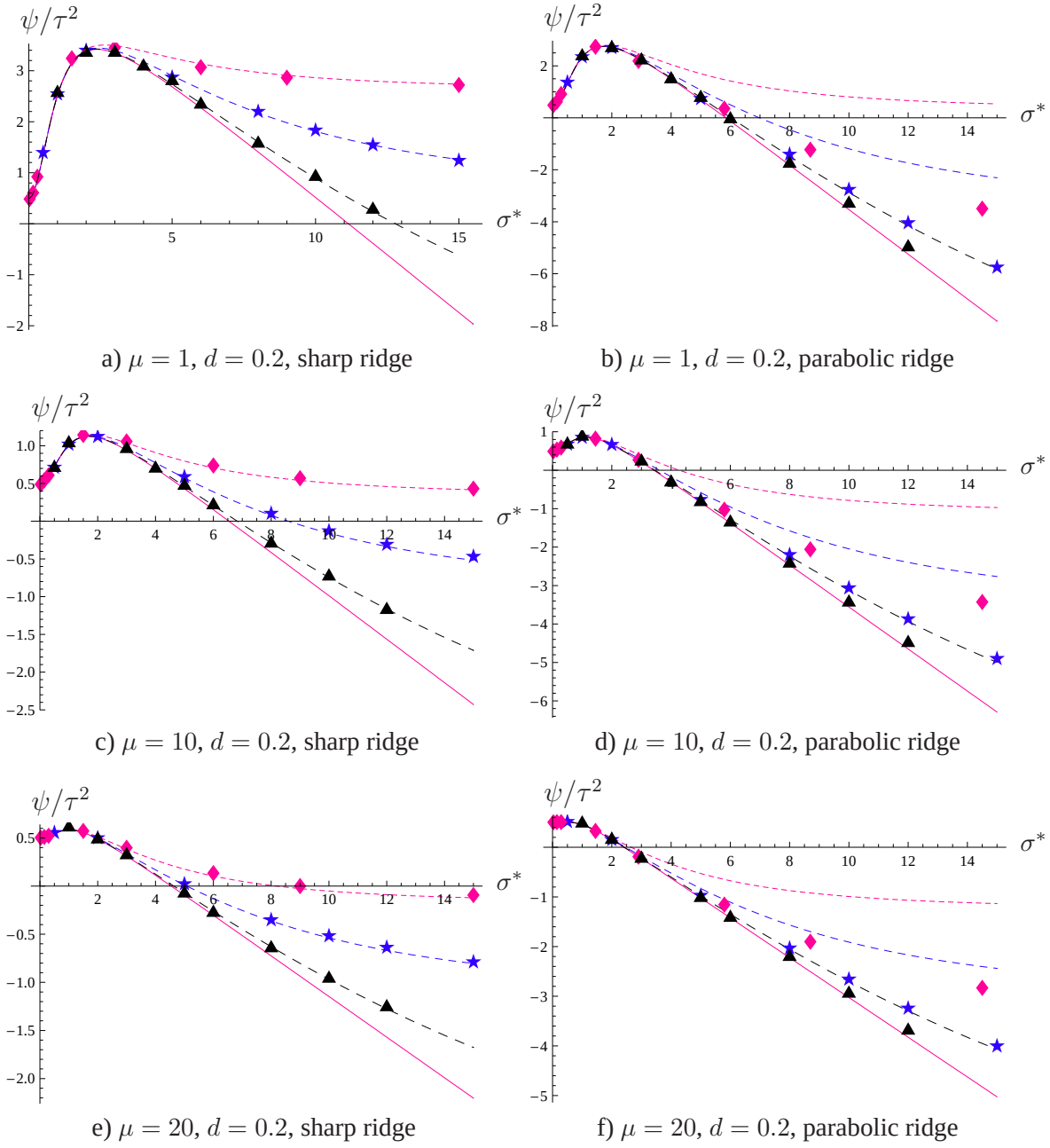
C.2.1 The log-normal operator

The moments of a log-normal distribution are given by $\overline{(\varsigma)^k}=(\langle\sigma\rangle)^k \mathrm{e}^{\frac{k^2\tau^2}{2}}$. The aim of the section is to derive expressions for $\overline{\left(\frac{\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\right)^k}$. Since

$$\Big(\frac{\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)^k = \sum_{l=0}^{k}\binom{k}{l}(\varsigma)^l(-1)^{k-l}(\langle\sigma\rangle)^{-l}, \tag{C.66}$$

the expectation is given by

$$\overline{\Big(\frac{\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)^k} = (-1)^k\sum_{l=0}^{k}\binom{k}{l}(-1)^l \mathrm{e}^{\frac{l^2\tau^2}{2}}. \tag{C.67}$$



a) $\mu = 1$, $d = 0.2$, sharp ridge

b) $\mu = 1$, $d = 0.2$, parabolic ridge

c) $\mu = 10$, $d = 0.2$, sharp ridge

d) $\mu = 10$, $d = 0.2$, parabolic ridge

e) $\mu = 20$, $d = 0.2$, sharp ridge

f) $\mu = 20$, $d = 0.2$, parabolic ridge

Figure C.9: The first-order SARs (C.64) (solid lines) and (C.63) (dashed lines) on the sharp and parabolic ridge for some $(\mu/\mu_I, 60)$-ES. The distance to the ridge was set to $R = 1$ and the noise strength to $\sigma_\epsilon^\star = 1$. Each data point was obtained by sampling over $100,000$ one-generation experiments for $N = 30$, $200,000$ for $N = 100$, and $250,000$ for $N = 500$. The results for $N = 30$ are denoted by diamond shaped symbols (red), whereas stars (blue) stand for $N = 100$, and triangles (black) for $N = 500$.

Since $\mathrm{e}^{\frac{l^2\tau^2}{2}} = \sum_{n=0}^{\infty} \frac{l^{2n}}{n!2^n}\tau^{2n}$,

$$\begin{aligned}\overline{\left(\frac{\varsigma - \langle\sigma\rangle}{\langle\sigma\rangle}\right)^k} &= (-1)^k \sum_{n=0}^{\infty} \frac{\tau^{2n}}{n!2^n} \sum_{l=0}^{k} \binom{k}{l} (-1)^l l^{2n} \\ &= (-1)^k \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} \sum_{l=0}^{k} \binom{k}{l} (-1)^l l^{2n} \end{aligned} \tag{C.68}$$

has to be considered. As it is shown later, $\sum_{l=0}^{k} \binom{k}{l}(-1)^l l^{2n} = 0$ holds for $k \geq 2n+1$. The expected values are therefore given by

$$\begin{aligned}\overline{\left(\frac{\varsigma - \langle\sigma\rangle}{\langle\sigma\rangle}\right)^k} &= \left\{ \begin{array}{ll} (-1)^k \sum_{n=k/2}^{\infty} \frac{\tau^{2n}}{n!2^n} & \\ \times \sum_{l=0}^{k} \binom{k}{l}(-1)^l l^{2n} & \text{if } k = 2j \\ (-1)^k \sum_{n=(k+1)/2}^{\infty} \frac{\tau^{2n}}{n!2^n} & \\ \times \sum_{l=0}^{k} \binom{k}{l}(-1)^l l^{2n} & \text{if } k = 2j+1 \end{array} \right\} \\ &= (-1)^k \sum_{n=\lceil k/2 \rceil}^{\infty} \frac{\tau^{2n}}{n!2^n} \sum_{l=0}^{k} \binom{k}{l} (-1)^l l^{2n}. \end{aligned} \tag{C.69}$$

As a result, if $\tau \ll 1$ is assumed and the Taylor series is cut off after $n = n_0$, accordingly, the expected values for $\left(\frac{\varsigma - \langle\sigma\rangle}{\varsigma^{(g)}}\right)^k$ with $k \geq 2n_0 + 1$ do not have to be taken into account. In the following, the expectation of $\left(\frac{\varsigma - \langle\sigma\rangle}{\varsigma^{(g)}}\right)^k$ is given for some choices of k, i.e., for

$$\begin{aligned} k = 1 \Rightarrow \overline{\left(\frac{\varsigma - \langle\sigma\rangle}{\langle\sigma\rangle}\right)^1} &= -\sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} \sum_{l=1}^{1} \binom{1}{l} (-1)^l l^{2n} = \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} \\ &= \frac{\tau^2}{2} + \frac{\tau^4}{8} + \ldots \\ k = 2 \Rightarrow \overline{\left(\frac{\varsigma - \langle\sigma\rangle}{\langle\sigma\rangle}\right)^2} &= \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} \sum_{l=1}^{2} \binom{2}{l} (-1)^l l^{2n} \\ &= \sum_{n=1}^{\infty} [2^n - 2^{1-n}] \frac{\tau^{2n}}{n!} = \tau^2 + \frac{7\tau^4}{4} + \ldots \\ k = 3 \Rightarrow \overline{\left(\frac{\varsigma - \langle\sigma\rangle}{\langle\sigma\rangle}\right)^3} &= -\sum_{n=2}^{\infty} \frac{\tau^{2n}}{n!2^n} \sum_{l=1}^{3} \binom{3}{l} (-1)^l l^{2n} \\ &= \sum_{n=2}^{\infty} 3[3^{2n-1} + 1 - 2^{2n}] \frac{\tau^{2n}}{2^n n!} = \frac{9\tau^4}{2} + \ldots \\ k = 4 \Rightarrow \overline{\left(\frac{\varsigma - \langle\sigma\rangle}{\langle\sigma\rangle}\right)^4} &= \sum_{n=2}^{\infty} \frac{\tau^{2n}}{n!2^n} \sum_{l=1}^{4} \binom{4}{l} (-1)^l l^{2n} \\ &= \sum_{n=2}^{\infty} [(6)2^{2n} + 2^{4n} - 4(1 + 3^{2n})] \frac{\tau^{2n}}{2^n n!} \\ &= 3\tau^4 + \ldots \end{aligned}$$

$$
\begin{aligned}
k=5 \Rightarrow \overline{\left(\frac{\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\right)^5} &= -\sum_{n=3}^{\infty}\frac{\tau^{2n}}{n!2^n}\sum_{l=1}^{5}\binom{5}{l}(-1)^l l^{2n} \\
&= \sum_{n=3}^{\infty}[1+2(3^{2n})+5^{2n-1}-2^{2n+1}-2^{4n}] \\
&\quad \times\frac{5\tau^{2n}}{2^n n!} = \frac{15\tau^6}{2}+\ldots. \qquad \text{(C.70)}
\end{aligned}
$$

The remainder of the section is devoted to show that $\sum_{l=0}^{k}\binom{k}{l}(-1)^l l^{2n}=0$ holds if $k \geq 2n+1$. This is done using induction. Let $m=2n$ and start with $m=0$. Splitting the sum into even and uneven terms and considering Pascal's triangle

$$
\begin{aligned}
\sum_{l=0}^{k}\binom{k}{l}(-1)^l &= \begin{cases} \sum_{l=0}^{(k)/2}\binom{k}{2l}-\sum_{l=0}^{k/2-1}\binom{k}{2l+1} & \text{if } k=2j \\ \sum_{l=0}^{(k-1)/2}\binom{k}{2l}-\sum_{l=0}^{(k-1)/2-1}\binom{k}{2l+1} & \text{if } k=2j+1 \end{cases} \\
&= 2^{k-1}-2^{k-1}=0. \qquad \text{(C.71)}
\end{aligned}
$$

Let now $m=1$. In this case

$$
\begin{aligned}
\sum_{l=0}^{k}\binom{k}{l}(-1)^l l &= k\sum_{l=1}^{k}\binom{k-1}{l-1}(-1)^l \\
&= -k\sum_{l=0}^{k-1}\binom{k-1}{l}(-1)^l = 0 \qquad \text{(C.72)}
\end{aligned}
$$

holds. Finally for $m \to m+1$, remember that l^m can be written as $l^m=\sum_{j=0}^{m}c_{m,j}\prod_{i=0}^{j-1}(l-i)$ with constants $c_{m,j}$. This leads to

$$
\begin{aligned}
\sum_{l=0}^{k}\binom{k}{l}(-1)^l l^{m+1} &= \sum_{l=1}^{k}\binom{k}{l}(-1)^l(l-m)l^m + m\sum_{l=1}^{k}\binom{k}{l}(-1)^l l^m \\
&= \sum_{l=1}^{k}\binom{k}{l}(-1)^l(l-m)\sum_{j=0}^{m}c_{m,j}\prod_{i=0}^{j-1}(l-i) \\
&= \sum_{l=1}^{k}\binom{k}{l}(-1)^l\sum_{j=0}^{m}c_{m,j}(l-j)\prod_{i=0}^{j-1}(l-i) \\
&\quad -\sum_{l=1}^{k}\binom{k}{l}(-1)^l\sum_{j=0}^{m}c_{m,j}(m-j)\prod_{i=0}^{j-1}(l-i) \\
&= \sum_{j=0}^{m}c_{m,j}\prod_{i=0}^{j}(k-i)\sum_{l=j+1}^{k}(-1)^l\binom{k-j-1}{l-j-1} \\
&\quad -\sum_{j=0}^{m}c_{m,j}(m-j)\prod_{i=0}^{j-1}(k-i) \\
&\quad \times\sum_{l=j}^{k}(-1)^l\binom{k-j}{l-j}=0. \qquad \text{(C.73)}
\end{aligned}
$$

Thus, the expectation of higher order terms vanishes. An analogous result holds for the two-point operator.

C.2.2 The two-point operator

The moments of the random variable ς^k are given by $\overline{\varsigma^k} = \langle\sigma\rangle^k/2\,(\alpha^k + \alpha^{-k})$. The analysis will be restricted to the case of $\alpha \approx 1$. Setting thus $\alpha := 1 + \beta$, $\beta \ll 1$ follows. The function $f(\beta) = (1+\beta)^{-k}$ will be developed into its Taylor series around zero. The Taylor series $T_f(\beta)$ is given by

$$\begin{aligned} T_f(\beta) &= \sum_{i=0}^{\infty} \frac{(k+j-1)!}{(k-1)!j!}(-1)^j\beta^j \\ &= 1 - k\beta + \frac{k(k+1)}{2}\beta^2 - \frac{k(k+1)(k+2)}{6}\beta^3 + \mathcal{O}(\beta^4). \end{aligned} \tag{C.74}$$

The term $(1+\beta)^k$ is given by the binomial formula

$$\begin{aligned} (1+\beta)^k &= \sum_{i=0}^{k} \binom{k}{i}\beta^i \\ &= 1 + k\beta + \frac{k(k-1)}{2}\beta^2 + \frac{k(k-1)(k-2)}{6}\beta^3 + \mathcal{O}(\beta^4). \end{aligned} \tag{C.75}$$

Thus, the sum of (C.74) and (C.75) reads

$$\begin{aligned} (1+\beta)^k + (1+\beta)^{-k} &= 1 - k\beta + \frac{k(k+1)}{2}\beta^2 - \frac{k(k+1)(k+2)}{6}\beta^3 \\ &\quad +1 + k\beta + k\frac{k(k-1)}{2}\beta^2 + \frac{k(k-1)(k-2)}{6}\beta^3 + \mathcal{O}(\beta^4) \\ &= 2 + \frac{k+1+k-1}{2}k\beta^2 + \frac{(k-1)(k-2)-(k+1)(k+2)}{6}k\beta^3 + \mathcal{O}(\beta^4) \\ &= 2 + k^2\beta^2 + k^2\beta^3 + \mathcal{O}(\beta^4) = 2 + k^2\beta^2(1+\beta) + \mathcal{O}(\beta^4). \end{aligned} \tag{C.76}$$

Addressing the expectation of

$$\left(\frac{\varsigma - \langle\sigma\rangle}{\langle\sigma\rangle}\right)^k = (-1)^k \sum_{l=0}^{k} \binom{k}{l}\varsigma^l(-1)^l\langle\sigma\rangle^{-l} \tag{C.77}$$

gives

$$\begin{aligned} \overline{\left(\frac{\varsigma - \langle\sigma\rangle}{\langle\sigma\rangle}\right)^k} &= (-1)^k \sum_{l=0}^{k} \binom{k}{l}(-1)^l\langle\sigma\rangle^{-l}\overline{\varsigma^l} \\ &= (-1)^k\frac{1}{2}\sum_{l=0}^{k} \binom{k}{l}(-1)^l\Big(2 + l^2\beta^2(1+\beta)\Big) + \mathcal{O}(\beta^4) \\ &= (-1)^k \sum_{l=0}^{k} \binom{k}{l}(-1)^l + \beta^2(1+\beta)(-1)^k\frac{1}{2}\sum_{l=0}^{k} \binom{k}{l}(-1)^l l^2 + \mathcal{O}(\beta^4) \\ &= (-1)^k \sum_{l=0}^{k} \binom{k}{l}(-1)^l + \frac{\beta^2}{2}(1+\beta)(-1)^k \sum_{l=0}^{k} \binom{k}{l}(-1)^l l^2 + \mathcal{O}(\beta^4). \end{aligned} \tag{C.78}$$

As it was shown in the previous section, the value of the first addend of (C.78) is zero. Therefore, no power of β below two appears in the approximation.

Considering the results for the log-normal distribution, we see that the expectation (C.78) contains only terms of order $\mathcal{O}(\beta^4)$ if $k \geq 3$. For the SAR, the values of (C.78) for $k = 1$ and $k = 2$, i.e.,

$$\begin{aligned}
\overline{\left(\frac{\varsigma - \langle\sigma\rangle}{\langle\sigma\rangle}\right)^1} &= -\beta^2(1+\beta)\frac{1}{2}\sum_{l=1}^{1}\binom{1}{l}(-1)^l l^2 + \mathcal{O}(\beta^4) \\
&= \frac{\beta^2}{2}(1+\beta) + \mathcal{O}(\beta^4) \qquad \text{(C.79)} \\
\overline{\left(\frac{\varsigma - \langle\sigma\rangle}{\langle\sigma\rangle}\right)^2} &= \beta^2(1+\beta)\frac{1}{2}\sum_{l=1}^{2}\binom{2}{l}(-1)^l l^2 + \mathcal{O}(\beta^4) \\
&= \frac{\beta^2}{2}(1+\beta)\left[-\binom{2}{1} + 4\binom{2}{2}\right] + \mathcal{O}(\beta^4) = \beta^2(1+\beta) + \mathcal{O}(\beta^4) \qquad \text{(C.80)}
\end{aligned}$$

need to be determined.

C.3 A General Formula

The section is devoted to the task of determining a recursive equation. The ultimate aim is to gain an equation or a MATHEMATICA-program which can be used to give the SAR in an (arbitrary) precision of τ. This section still does not include the τ-dependent terms of P_Q (C.4) in the derivation. The main point of this section is to illustrate some points of the derivation which are also relevant for the next section which presents an approach which accounts for all τ-dependent terms.

C.3.1 The Derivation

The starting point is (C.20) in Appendix C.1, p. 142,

$$\begin{aligned}
\psi(\langle\sigma\rangle) = &\int_0^\infty \Big(\frac{\varsigma - \langle\sigma\rangle}{\langle\sigma\rangle}\Big) p_\sigma(\varsigma|\langle\sigma\rangle)(\lambda-\mu)\binom{\lambda}{\mu}\int_{-\infty}^{\infty}\Big(1-\Phi(t)\Big)^{\lambda-\mu-1}\Phi(t)^\mu \frac{\mathrm{e}^{-\frac{t^2}{2}}}{\sqrt{2\pi}}\,\mathrm{d}t\,\mathrm{d}\varsigma \\
&+\int_0^\infty \langle\sigma\rangle\Big(\frac{\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)^2 p_\sigma(\varsigma|\langle\sigma\rangle)(\lambda-\mu)\binom{\lambda}{\mu} \\
&\times\int_{-\infty}^{\infty}\Big(1-\Phi(t)\Big)^{\lambda-\mu-1}\Phi(t)^{\mu-1}\frac{\mathrm{e}^{-t^2}}{2\pi}\Big(-\frac{g'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}t - \frac{h'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}\Big)\,\mathrm{d}t\,\mathrm{d}\varsigma \\
&+\sum_{k=1}^{\infty}\int_0^\infty \frac{\langle\sigma\rangle^{k+1}}{(k+1)!}\Big(\frac{(\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)^{k+2} p_\sigma(\varsigma|\langle\sigma\rangle) \\
&\times(\lambda-\mu)\binom{\lambda}{\mu}\int_{-\infty}^{\infty}\Big(1-\Phi(t)\Big)^{\lambda-\mu-1}\Phi(t)^{\mu-1}\frac{\mathrm{e}^{-\frac{t^2}{2}}}{2\pi} \\
&\times\frac{\partial^k}{\partial\varsigma^k}\Bigg(\Big(-g(\langle\sigma\rangle)\frac{g'(\varsigma)}{g^2(\varsigma)}t + \frac{g'(\varsigma)(h(\varsigma)-h(\langle\sigma\rangle))}{g((\varsigma)^2} - \frac{h'(\varsigma)}{g(\varsigma)}\Big) \\
&\times\exp\Big(-\frac{1}{2}\Big(\frac{g(\langle\sigma\rangle)t-(h(\varsigma)-h(\langle\sigma\rangle))}{g(\varsigma)}\Big)^2\Big)\Bigg)|_{\varsigma=\langle\sigma\rangle}\,\mathrm{d}t\,\mathrm{d}\varsigma.
\end{aligned}$$

Again, the three integrals have to be considered. The first two can be easily developed into a general formula of τ. Recall from Appendix C.2 that

$$\begin{aligned}\overline{\Big(\frac{\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)} &= -\sum_{n=1}^{\infty}\frac{\tau^{2n}}{2^n n!}\sum_{l=0}^{1}\binom{1}{l}(-1)^l l^{2n} = \sum_{n=1}^{\infty}\frac{\tau^{2n}}{2^n n!}\\ \overline{\Big(\frac{\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)^2} &= \sum_{n=1}^{\infty}\frac{\tau^{2n}}{2^n n!}\sum_{l=0}^{2}\binom{2}{l}(-1)^l l^{2n} = \sum_{n=1}^{\infty}\frac{\tau^{2n}}{2^n n!}\Big(-2+2^{2n}\Big)\end{aligned} \tag{C.81}$$

holds. Considering the results obtained so far in Appendix C.1 for the first two integrals in (C.20), it is easy to see that

$$\begin{aligned}I_1+I_2 &= \sum_{n=1}^{\infty}\frac{\tau^{2n}}{2^n n!}\\ &\quad+\langle\sigma\rangle\Big(\frac{g'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}e^{1,1}_{\mu,\lambda}-c_{\mu/\mu,\lambda}\frac{h'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}\Big)\sum_{n=1}^{\infty}\frac{\tau^{2n}}{2^n n!}\Big(-2+2^{2n}\Big)\\ &= \sum_{n=1}^{\infty}\frac{\tau^{2n}}{2^n n!}\Big(1+\Big(-2+2^{-2n}\Big)\langle\sigma\rangle\Big)\Big(\frac{g'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}e^{1,1}_{\mu,\lambda}-c_{\mu/\mu,\lambda}\frac{h'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}\Big)\Big)\end{aligned} \tag{C.82}$$

(cf. (C.22)) holds. As already mentioned in Appendix C.1, the third integral poses more difficulties. This concerns the appearance of higher derivatives and of course the integration over t. Let us now focus on

$$\begin{aligned}I_3 &= \sum_{k=1}^{\infty}\int_0^{\infty}\frac{\langle\sigma\rangle^{k+1}}{(k+1)!}\Big(\frac{\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)^{k+2}p_\sigma(\varsigma|\langle\sigma\rangle)\\ &\quad\times(\lambda-\mu)\binom{\lambda}{\mu}\int_{-\infty}^{\infty}\Big(1-\Phi(t)\Big)^{\lambda-\mu-1}\Phi(t)^{\mu-1}\frac{\mathrm{e}^{-\frac{t^2}{2}}}{2\pi}\\ &\quad\times\frac{\partial^k}{\partial\varsigma^k}\Bigg(\Big(-g(\langle\sigma\rangle)\frac{g'(\varsigma)}{g^2(\varsigma)}t+\frac{g'(\varsigma)(h(\varsigma)-h(\langle\sigma\rangle))}{g((\varsigma)^2}-\frac{h'(\varsigma)}{g(\varsigma)}\Big)\\ &\quad\times\exp\Big(-\frac{1}{2}\Big(\frac{g(\langle\sigma\rangle)t-(h(\varsigma)-h(\langle\sigma\rangle))}{g(\varsigma)}\Big)^2\Big)\Bigg)|_{\varsigma=\langle\sigma\rangle}\,\mathrm{d}t\,\mathrm{d}\varsigma\end{aligned} \tag{C.83}$$

and start with the derivatives. In the following let

$$u(\varsigma) := g(\langle\sigma\rangle)\frac{g'(\varsigma)}{g^2(\varsigma)}t+\frac{g'(\varsigma)(h(\varsigma)-h(\langle\sigma\rangle))}{g((\varsigma)^2}-\frac{h'(\varsigma)}{g(\varsigma)} \tag{C.84}$$

and

$$v(\varsigma) := \exp\Big(-\frac{1}{2}\Big(\frac{g(\langle\sigma\rangle)t-(h(\varsigma)-h(\langle\sigma\rangle))}{g(\varsigma)}\Big)^2\Big). \tag{C.85}$$

The kth derivative of a product of two functions simply reads

$$(u(\varsigma)v(\varsigma))^{(k)} = \sum_{l=0}^{k}\binom{k}{l}u^{(k-l)}(\varsigma)v^{(l)}(\varsigma) \tag{C.86}$$

with $u^{(k)}(\varsigma) := \frac{\partial^k}{\partial \varsigma^k} u(\varsigma)$. The kth derivative of a composite function is not so easily obtained. Following [32] it reads

$$\frac{\mathrm{d}^n}{\mathrm{d}\varsigma^n} u(v(\varsigma)) = n! \sum_{\{k_m\}} \frac{\mathrm{d}^r}{\mathrm{d}y^r} u(y)|_{y=v(\varsigma)} \prod_{m=1}^{n} \frac{1}{k_m!} \Big(\frac{1}{m!} v^{(m)}(\varsigma)\Big)^{k_m} \tag{C.87}$$

with $r = k_1 + \ldots + k_n$ and $\{k_m\}$ the set of all non-negative integer solutions of the so-called Diophantine equation (see, e.g., [32])

$$k_1 + 2k_1 + \ldots + nk_n = n. \tag{C.88}$$

The lth derivative of u is of the form

$$u^{(l)} = \Big(\frac{g'(\varsigma)}{g(\varsigma)^2}\Big)^{(l)} g(\langle\sigma\rangle)t + \Big(\frac{g'(\varsigma)}{g(\varsigma)^2}(h(\varsigma) - h(\langle\sigma\rangle))\Big)^{(l)} - \Big(\frac{h'(\varsigma)}{g(\varsigma)}\Big)^{(l)} \tag{C.89}$$

with

$$\Big(\frac{g'(\varsigma)}{g(\varsigma)^2}\Big)^{(l)} = \sum_{j=0}^{l} \binom{l}{j} g^{(l+1-j)}(\varsigma) \Big(g(\varsigma)^{-2}\Big)^{(j)}. \tag{C.90}$$

The jth derivative of the composite function is given by

$$\Big(g(\varsigma)^{-2}\Big)^{(j)} = j! \sum_{\{k_m\}} (2+r)!(-1)^r g(\varsigma)^{-2-r} \prod_{m=1}^{j} \frac{1}{k_m!} \Big(\frac{1}{m!} g^{(m)}(\varsigma)\Big)^{k_m} \tag{C.91}$$

since $(y^{-2})^{(j)} = (-1)^j (2+j)! y^{-2-j}$. The derivation of the third term in (C.89) can be obtained by

$$\Big(\frac{h'(\varsigma)}{g(\varsigma)}\Big)^{(l)} = \sum_{j=0}^{l} \binom{l}{j} h^{(l+1-j)}(\varsigma) \Big(g(\varsigma)^{-1}\Big)^{(j)} \tag{C.92}$$

with

$$\Big(g(\varsigma)^{-1}\Big)^{(j)} = j! \sum_{\{k_m\}} (1+r)!(-1)^r g(\varsigma)^{-1-r} \prod_{m=1}^{j} \frac{1}{k_m!} \Big(\frac{1}{m!} g^{(m)}(\varsigma)\Big)^{k_m}. \tag{C.93}$$

The remaining derivation of the last composite term of (C.89) can be determined using

$$\begin{aligned} \Big(\frac{g'(\varsigma)}{g(\varsigma)^2}(h(\varsigma) - h(\langle\sigma\rangle))\Big)^{(l)} &= \sum_{j=0}^{l} \binom{l}{j} \Big(g'(\varsigma)(h(\varsigma) - h(\langle\sigma\rangle))\Big)^{(l-j)} \Big(g(\varsigma)^{-2}\Big)^{(j)} \\ &= \sum_{j=0}^{l} \sum_{k=0}^{l-j} \binom{l}{j} \binom{lj}{k} g^{(l-j-k+1)}(\varsigma)(h(\varsigma) - h(\langle\sigma\rangle))\Big)^{(k)} \\ &\quad \times \Big(g(\varsigma)^{-2}\Big)^{(j)}. \end{aligned} \tag{C.94}$$

Concerning t the lth derivative of u stays linear. This is not the case if v is considered. The function itself is a composite function of the form $v(\varsigma) = \exp(w)$ and therefore the derivative is

$$v^{(l)}(\varsigma) = l! \sum_{\{k_m\}} \mathrm{e}^{w(\varsigma)} \prod_{m=1}^{l} \frac{1}{k_m!} \Big(\frac{1}{m!} w^{(m)}(\varsigma)\Big)^{k_m}. \tag{C.95}$$

The function w is again a composite function with $w(\varsigma) = -1/2z(\varsigma)^2$ leading to

$$w^{(l)}(\varsigma) = -\frac{1}{2}l! \sum_{\{k_m\}} 2 \times \ldots \times (2-r) z(\varsigma)^{2-r} \prod_{m=1}^{l} \frac{1}{k_m!} \Big(\frac{1}{m!} w^{(m)}(y)|_{y=z(\varsigma)} \Big)^{k_m}. \quad \text{(C.96)}$$

Finally, the last remaining derivatives remain those of the arguments of z leading to

$$z^{(l)}(\varsigma) = g(\langle\sigma\rangle) t \Big(g(\varsigma)^{-1} \Big)^{(l)} + \sum_{m=0}^{l} \binom{l}{m} \Big(h(\varsigma) - h(\sigma) \Big)^{(l-m)} \Big(g(\varsigma)^{-1} \Big)^{(m)}. \quad \text{(C.97)}$$

Some simplifications can be made:

1. The lth summand in (C.97) vanishes completely

$$\begin{aligned} z^{(l)}(\varsigma)|_{\varsigma=\langle\sigma\rangle} &= g(\langle\sigma\rangle) t \Big(g(\varsigma)^{-1} \Big)^{(l)}|_{\varsigma=\langle\sigma\rangle} \\ &\quad + \sum_{m=0}^{l-1} \binom{l}{m} h(\varsigma)^{(l-m)}|_{\varsigma=\langle\sigma\rangle} \Big(g(\varsigma)^{-1} \Big)^{(m)}|_{\varsigma=\langle\sigma\rangle}. \end{aligned} \quad \text{(C.98)}$$

2. Since $z(\langle\sigma\rangle) = t$, only sets with at most two elements have to be taken into account

$$w^{(l)}(\varsigma) = -\frac{1}{2}l! \sum_{\{k_m\}} 2 \times \ldots \times (2-r) t^{2-r} \prod_{m=1}^{l} \frac{1}{k_m!} \Big(\frac{1}{m!} z^{(m)}(\varsigma) \Big)^{k_m}. \quad \text{(C.99)}$$

3. Concerning t the lth derivative of $v\mathrm{e}^{t^2/2}$ is a polynomial in t

$$v^{(l)}(\varsigma) = l! \mathrm{e}^{-\frac{t^2}{2}} \sum_{k_m} \prod_{m=1}^{l} \frac{1}{k_m!} \Big(\frac{1}{m!} w^{(m)}(\varsigma) \Big)^{k_m}. \quad \text{(C.100)}$$

In principle, (C.86) to (C.100) can be used to determine the SAR. However, performing the calculations is lengthy and the results are not easily usable. Therefore, the remainder of the section is aimed at providing a MATHEMATICA-program for determining the SAR. To this end, reconsider (C.82) and (C.83). Equation (C.82) can be directly transferred. In the case of Eq. (C.83), the first step is to swap the integration order – computing first the integral over ς

$$\begin{aligned} I_3 = \sum_{k=1}^{\infty} \int_0^{\infty} & \frac{\langle\sigma\rangle^{k+1}}{(k+1)!} \Big(\frac{(\varsigma - \langle\sigma\rangle}{\langle\sigma\rangle} \Big)^{k+2} p_\sigma(\varsigma|\langle\sigma\rangle) \, \mathrm{d}\varsigma \\ & \times (\lambda - \mu) \binom{\lambda}{\mu} \int_{-\infty}^{\infty} \Big(1 - \Phi(t) \Big)^{\lambda-\mu-1} \Phi(t)^{\mu-1} \frac{\mathrm{e}^{-\frac{t^2}{2}}}{2\pi} \\ & \times \frac{\partial^k}{\partial \varsigma^k} \Bigg(\Big(g(\langle\sigma\rangle) \frac{g'(\varsigma)}{g^2(\varsigma)} t + \frac{g'(\varsigma)(h(\varsigma) - h(\langle\sigma\rangle))}{g((\varsigma)^2} - \frac{h'(\varsigma)}{g(\varsigma)} \Big) \\ & \times \exp\Big(-\frac{1}{2} \Big(\frac{g(\langle\sigma\rangle)t - (h(\varsigma) - h(\langle\sigma\rangle))}{g(\varsigma)} \Big)^2 \Big) \Bigg)|_{\varsigma=\langle\sigma\rangle} \, \mathrm{d}t. \end{aligned} \quad \text{(C.101)}$$

The integration result for every term of the series in n gives a series in τ^{2l}. To obtain the general series in powers of τ^2, the summation order must be swapped. For notation convenience let $C_k^t(\langle\sigma\rangle)$ denote the integral over t of $k$the derivative in (C.101). After integrating over ς,

$$\begin{aligned} I_3 &= \sum_{k=1}^{\infty} C_k^t \frac{\langle\sigma\rangle^{k+1}}{(k+1)!}(-1)^{k+2} \sum_{j=\lceil k/2+1\rceil}^{\infty} \frac{\tau^{2j}}{j!2^j} \sum_{h=0}^{k+2} \binom{k+2}{h}(-1)^h h^{2j} \\ &= \sum_{j=2}^{\infty} \frac{\tau^{2j}}{j!2^j} \sum_{k=1}^{2j-2} C_k^t \frac{\langle\sigma\rangle^{k+1}}{(k+1)!}(-1)^k \sum_{h=0}^{k+2} \binom{k+2}{h}(-1)^h h^{2j} \end{aligned} \tag{C.102}$$

is obtained. As shown, the coefficient C_k^t leads to expressions of the form $\mathrm{e}^{-t^2/2}\sum_{i=0}^{2k+1} a_i(\langle\sigma\rangle)t^i$. The integration over t in (C.101) leads therefore to special cases of the progress coefficients (A.24)

$$\begin{aligned} C_k^t &= \sum_{i=0}^{2k+1} a_i(\langle\sigma\rangle)(\lambda-\mu)\binom{\lambda}{\mu} \int_{-\infty}^{\infty} \Big(1-\Phi(t)\Big)^{\lambda-\mu-1} \Phi(t)^{\mu-1} t^i \frac{\mathrm{e}^{-\frac{t^2}{2}}}{2\pi}\,\mathrm{d}t \\ C_k^t &= \sum_{i=0}^{2k+1} (-1)^i a_i(\langle\sigma\rangle)(\lambda-\mu)\binom{\lambda}{\mu} \int_{-\infty}^{\infty} \Phi(t)^{\lambda-\mu-1} \Big(1-\Phi(t)\Big)^{\mu-1} t^i \frac{\mathrm{e}^{-\frac{t^2}{2}}}{2\pi}\,\mathrm{d}t \\ &= \sum_{i=0}^{2k+1} (-1)^i a_i(\langle\sigma\rangle) e_{\mu,\lambda}^{1,i}. \end{aligned} \tag{C.103}$$

The task remains to determine the coefficients in (C.103) which can be done using MATHEMATICA. The SAR can then be obtained by combining (C.82) and (C.102) as

$$\begin{aligned} \psi(\langle\sigma\rangle) &= \sum_{n=1}^{\infty} \frac{\tau^{2n}}{2^n n!} \left(1 + \Big(-2+2^{-2n}\Big)\langle\sigma\rangle\right) \left(\frac{g'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)} e_{\mu,\lambda}^{1,1} - c_{\mu/\mu,\lambda} \frac{h'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}\right) \\ &\quad + \sum_{j=2}^{\infty} \frac{\tau^{2j}}{j!2^j} \sum_{k=1}^{2j-2} C_k^t \frac{\langle\sigma\rangle^{k+1}}{(k+1)!}(-1)^k \sum_{h=0}^{k+2} \binom{k+2}{h}(-1)^h h^{2j} \end{aligned} \tag{C.104}$$

In the following section, the effects of including higher-order terms of τ in the SAR (C.104) are discussed. The parabolic ridge is used to as a test function for the SAR (C.104).

C.3.2 Comparison with the Parabolic Ridge

Let us compare the obtained SAR (C.104) with the results of experiments for the parabolic ridge. Three evolution strategies were examined: a $(1,60)$-ES, a $(10/10_I,60)$-ES, and a $(20/20_I,60)$-ES. The SAR was expanded up to τ^6. The ridge constant d was set to $d=0.2$ and the distance to the ridge to $R=1$. In the following, the SARs are numbered in accordance to the expansion, i.e., ψ_i denotes the result up to the power of τ^{2i}. Figure C.11 compares the prediction with the results of experiments for $N=100$. In the derivation of the SAR, the N-dependent version was used. The influence of the higher order τ-terms is relatively minor. Although, ψ_k with $k>1$ deviate from the results obtained for $k=1$, the effect wished for cannot be obtained in general. In the case of $\mu=1$, Fig. C.11 a), the deviations from the result for ψ_1 do not lead to a better prediction quality. In the case of $\mu=10$, Fig. C.11 b), ψ_2 and ψ_3 move closer to the experimental results for higher mutation strengths, but ψ_2 and ψ_3 do not overlap with the measured data. Furthermore, ψ_3 does not deviate far from ψ_2. In

```
Get["eabml.mat"]
Clear[fh,fg,awts,cwkt,bw,wkt,psi1,psi3,psi]
n=100
d=0.2
a=2
fh[s_]:=a*d*s^2/2
dfh[s_]:=D[fh[x],{x,1}]/.x->s
fg[s_]:=s*Sqrt[1+a^2*d^2+a^2*d^2*s^2/(2*n)]
dfg[s_]:=D[fg[x],{x,1}]/.x->s
awts[k_,t_,s_]:=
Module[{x,y,l,erg},
expo[x_, y_, l_] :=
 D[expo[x, y, l - 1], x] -
  expo[x, y, 1]*expo[x, y, l - 1]*expo[x, y, 0];
expo[x_, y_, 0] := fg[s]/fg[x]*y - (fh[x] - fh[s])/fg[x];
expo[x_, y_, 1] := D[fg[s]/fg[x]*y - (fh[x] - fh[s])/fg[x],x];
erg=expo[x,y,k+1]/.x->s/.y->t
]
cwkt[k_,s_,m_,l_]:=Module[{alist,erg,coefs,y,as},
        coefs=CoefficientList[awts[k,y,as],y];
        erg=Sum[(-1)^(j-1)*coefs[[j]]*eabml[1,j-1,m,l],
                                    {j,1,Length[coefs]}];
        erg/.as->s]
wkt[k_,j_]:=Sum[Binomial[k,h]*(-1)^h*h^(2*j),{h,0,k}]
bw[w_,s_,m_,l_]:=Module[{as,k,erg},
        erg=If[2*w-2<=0,
        0,
        Sum[cwkt[k,as,m,l]*as^(k+1)/((k+1)!)*(-1)^k*wkt[(k+2),w],
                                                {k,1,w*2-2}]
        ];
        erg/.as->s]
psi3[tau_,i_,s_,m_,l_]:=Module[{as,t,erg},
        erg=If[i<2,0,Sum[tau^(2*w)/((w!)*2^w)*bw[w,s,m,l],{w,2,i}]]
]
psi1[tau_,i_,s_,m_,l_]:=
        Sum[tau^(2*w)/((w!)*2^w)
                *(1+(-2+2^(2*w))*s
                *(dfg[s]/fg[s]*eabml[1,1,m,l]-cmmkl[m,l]*dfh[s]/fg[s])),
                {w,1,i}]
psi[tau_,i_,s_,m_,l_]:=psi1[tau,i,s,m,l]+psi3[tau,i,s,m,l]
```

Figure C.10: The MATHEMATICA source code for the SAR

the case of $\mu = 20$, finally, ψ_2 and ψ_3 are very close to the experimental data. Note, ψ_2 (indicated by the dashed line with the shorter dots in Fig. C.11 c)) gives better results. The behavior may have several causes: First of all, there are all still neglected τ^{2k} terms which may cause deviations. Second, it should be noted that taking more terms of the τ^2 series does not necessarily improve the prediction quality for any fixed τ.

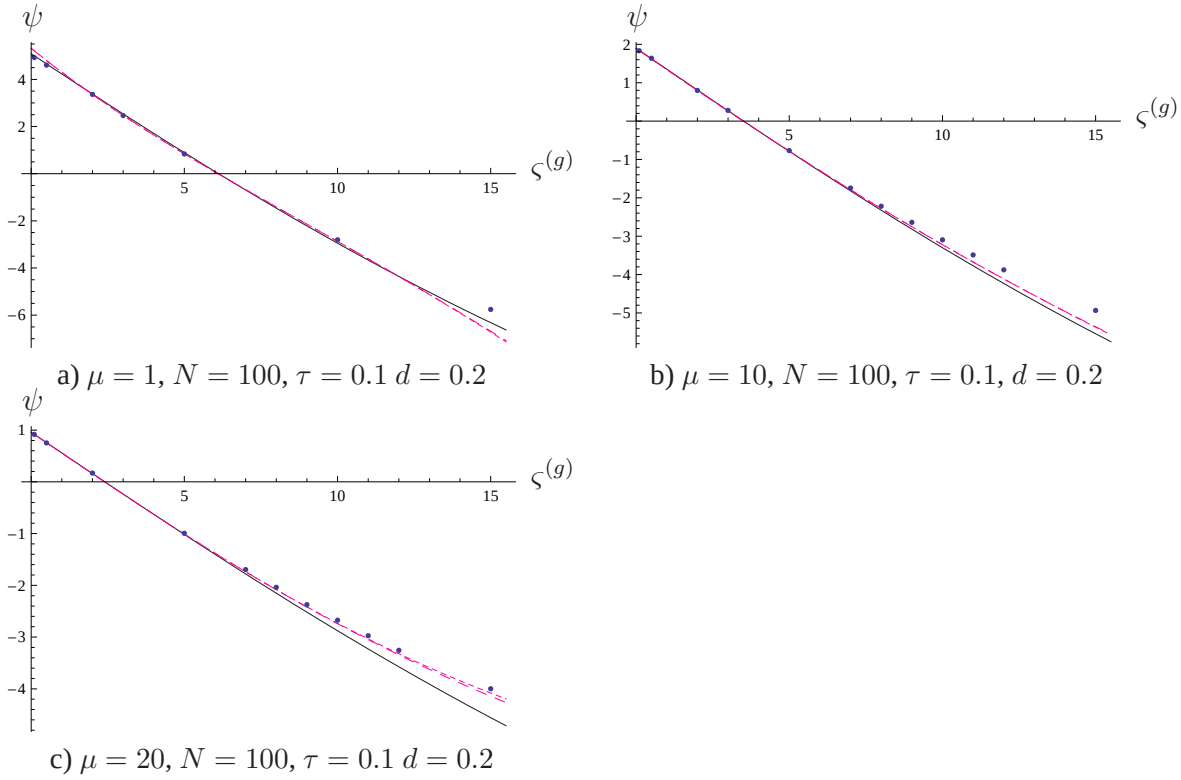


Figure C.11: Comparison of the SAR (C.104) with the results of experiments. Three SARs, ψ_1, ψ_2, and ψ_3 are shown. The solid line represents ψ_1, the dotted ψ_2 (dashed, short dots) and ψ_3 (dashed, longer dots). The results for ψ_2 and ψ_3 cannot be distinguished, since the lines nearly overlap.

C.4 A General Formula: A Second Approach

First of all, a minimization problem will be considered. In other words, the mth best fitness change is not the mth highest fitness change but the mth smallest. It is easy to see that the fitnes change of an offspring retains the general form of the previous sections. In other words, it is assumed that first,

$$P_Q(Q) = P_Q\Big(\frac{Q - h(\varsigma)}{g(\varsigma)}\Big) \tag{C.105}$$

and second

$$P_Q(Q) = \Phi\Big(\frac{Q - h(\varsigma)}{g(\varsigma)}\Big) \tag{C.106}$$

holds in accordance with (C.4). Let us reconsider the SAR (C.1) which is now given by

$$\psi(\sigma) = \frac{\lambda}{\mu}\sum_{m=1}^{\mu}\binom{\lambda-1}{m-1}\int_0^\infty \Big(\frac{\varsigma-\sigma}{\sigma}\Big)p_\sigma(\varsigma|\sigma)$$

$$\times \int_{-\infty}^{\infty} p_Q(Q|\varsigma)\mathrm{P}_Q(Q|\sigma)^{m-1}\Big(1-\mathrm{P}_Q(Q|\sigma)\Big)^{\lambda-m}\,\mathrm{d}Q\,\mathrm{d}\varsigma \tag{C.107}$$

with

$$\mathrm{P}_Q(Q|\sigma) \;=\; \int_0^{\infty} P_Q(Q|\varsigma)p_\sigma(\varsigma|\varsigma)\,\mathrm{d}\varsigma. \tag{C.108}$$

At this point the deviation is changed. Instead of switching to standardized integration variables, the steps for (C.13) - (C.16) are performed first. First, the order of summation and integration is swapped and the sum is substituted by an integral

$$\frac{\lambda}{\mu}\sum_{m=1}^{\mu}\binom{\lambda-1}{m-1}\Big(1-\mathrm{P}_Q(Q)\Big)^{\lambda-m}\mathrm{P}_Q(Q)^{m-1} = \frac{\lambda!}{\mu}\frac{\int_0^{1-\mathrm{P}_Q(Q)} x^{\lambda-\mu-1}(1-x)^{\mu-1}\,\mathrm{d}x}{(\lambda-\mu-1)!(\mu-1)!}. \tag{C.109}$$

Again, the integral is reinserted into the SAR (C.107) giving

$$\begin{aligned}\psi(\sigma) \;&=\; (\lambda-\mu)\binom{\lambda}{\mu}\int_0^{\infty}\Big(\frac{\varsigma-\sigma}{\sigma}\Big)p_\sigma(\varsigma|\sigma)\int_{-\infty}^{\infty}p_Q(Q|\varsigma)\\ &\quad\times\int_0^{1-\mathrm{P}_Q(Q)} x^{\lambda-\mu-1}(1-x)^{\mu-1}\,\mathrm{d}x\,\mathrm{d}Q\,\mathrm{d}\varsigma.\end{aligned} \tag{C.110}$$

After some calculations and subsequent reordering, the SAR

$$\begin{aligned}\psi(\sigma) \;&=\; (\lambda-\mu)\binom{\lambda}{\mu}\int_0^{\infty}\Big(\frac{\varsigma-\sigma}{\sigma}\Big)p_\sigma(\varsigma|\sigma)\\ &\quad\times\int_{-\infty}^{\infty}(1-\mathrm{P}_Q(Q))^{\lambda-\mu-1}\mathrm{P}_Q(Q)^{\mu-1}\mathrm{p}_Q(Q)P_Q(Q|\varsigma)\,\mathrm{d}Q\,\mathrm{d}\varsigma\end{aligned} \tag{C.111}$$

is obtained with $\mathrm{p}_Q := \partial/(\partial Q)\mathrm{P}_Q$. At this point the integral

$$\mathrm{P}_Q(Q|\sigma) \;=\; \int_0^{\infty} P_Q(Q|\varsigma)p_\sigma(\varsigma|\sigma)\,\mathrm{d}\varsigma \tag{C.112}$$

has to be reconsidered. Expanding P_Q into its Taylor series around σ gives

$$\begin{aligned}\mathrm{P}_Q(Q|\sigma) \;&=\; \int_0^{\infty}\sum_{k=0}^{\infty}\frac{\partial^k}{\partial\varsigma^k}P_Q(Q|\varsigma)|_{\varsigma=\sigma}\Big(\frac{\varsigma-\sigma}{\sigma}\Big)^k\frac{\sigma^k}{k!}p_\sigma(\varsigma|\sigma)\,\mathrm{d}\varsigma\\ &=\; \sum_{k=0}^{\infty}\frac{\partial^k}{\partial\varsigma^k}P_Q(Q|\varsigma)|_{\varsigma=\sigma}\overline{\Big(\frac{\varsigma-\sigma}{\sigma}\Big)^k}\frac{\sigma^k}{k!}\\ &=\; \Phi\Big(\frac{Q-h(\sigma)}{g(\sigma)}\Big)+\sum_{k=1}^{\infty}\frac{\partial^k}{\partial\varsigma^k}P_Q(Q|\varsigma)|_{\varsigma=\sigma}\overline{\Big(\frac{\varsigma-\sigma}{\sigma}\Big)^k}\frac{\sigma^k}{k!}.\end{aligned} \tag{C.113}$$

Again, it is refrained from computing the derivatives $\partial^k/(\partial\varsigma^k)P_Q(Q|\varsigma)|_{\varsigma=\sigma}$. This will be done eventually using MATHEMATICA. Note the following, though: The kth derivative of P_Q can be given as a product of the pdf p_Q and a polynomial in Q. This will finally lead to coefficients similar to the progress coefficients $e_{\mu,\lambda}^{\alpha,\beta}$ (A.24). The second step consists of taking the expectation $\mathrm{E}[((\varsigma-\sigma)/\sigma)^k]$ and developing it into a series in τ^2 similar to Eqs. (C.66)f. in Appendix C.2.1. Accordingly,

$$\mathrm{P}_Q(Q|\sigma) \;=\; \Phi\Big(\frac{Q-h(\sigma)}{g(\sigma)}\Big)+\sum_{k=1}^{\infty}a_k(Q,\sigma)\sum_{n=\lceil k/2\rceil}^{\infty}\frac{\tau^{2n}}{n!2^n}\sum_{l=0}^{k}\binom{k}{l}(-1)^l l^{2n} \tag{C.114}$$

is obtained (cf. (C.69)). The coefficient a_k denotes

$$a_k(Q,\sigma) \;:=\; (-1)^k \frac{\partial^k}{\partial \varsigma^k} P_Q(Q|\varsigma)|_{\varsigma=\sigma} \frac{\sigma^k}{k!}. \tag{C.115}$$

The last calculation concerning $\mathrm{P}_Q(Q|\sigma)$ at this moment is to change the order of summation leading to

$$\begin{aligned} \mathrm{P}_Q(Q|\sigma) &= \Phi\Big(\frac{Q-h(\sigma)}{g(\sigma)}\Big) + \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} \sum_{k=1}^{2n} a_k(Q,\sigma) \sum_{l=0}^{k} \binom{k}{l} (-1)^l l^{2n} \\ &=: \Phi\Big(\frac{Q-h(\sigma)}{g(\sigma)}\Big) + \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} c_n(Q,\sigma). \end{aligned} \tag{C.116}$$

Accordingly, the product $(1-\mathrm{P}_Q)^{\lambda-\mu-1}$ in (C.111) reads

$$\begin{aligned} \Big(1-\mathrm{P}_Q(Q|\sigma)\Big)^{\lambda-\mu-1} &= \sum_{l=0}^{\lambda-\mu-1} \binom{\lambda-\mu-1}{l} \Big(1-\Phi\Big(\frac{Q-h(\sigma)}{g(\sigma)}\Big)\Big)^{\lambda-\mu-1-l} \\ &\quad \times (-1)^l \Big(\sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} c_n(Q,\sigma)\Big)^l \end{aligned} \tag{C.117}$$

whereas $\mathrm{P}_Q^{\mu-1}$ in (C.111) is given by

$$\mathrm{P}_Q(Q|\sigma)^{\mu-1} = \sum_{m=0}^{\mu-1} \binom{\mu-1}{m} \Phi\Big(\frac{Q-h(\sigma)}{g(\sigma)}\Big)^{\mu-1-m} \Big(\sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} c_n(Q,\sigma)\Big)^m. \tag{C.118}$$

The product of (C.116) and (C.117) reads in turn

$$\begin{aligned} \Big(1-\mathrm{P}_Q(Q|\sigma)\Big)^{\lambda-\mu-1} \mathrm{P}_Q(Q|\sigma)^{\mu-1} &= \sum_{l=0}^{\lambda-\mu-1} \sum_{m=0}^{\mu-1} \binom{\lambda-\mu-1}{l} \binom{\mu-1}{m} \\ &\quad \times \Big(1-\Phi\Big(\frac{Q-h(\sigma)}{g(\sigma)}\Big)\Big)^{\lambda-\mu-1-l} \Phi\Big(\frac{Q-h(\sigma)}{g(\sigma)}\Big)^{\mu-1-m} \\ &\quad \times (-1)^l \Big(\sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} c_n(Q,\sigma)\Big)^{m+l}. \end{aligned} \tag{C.119}$$

A similar series is obtained for the integral

$$I_Q(Q,\sigma) = \int_0^{\infty} \Big(\frac{\varsigma-\sigma}{\sigma}\Big) P_Q(Q|\varsigma) p_\sigma(\varsigma|\sigma)\,\mathrm{d}\varsigma. \tag{C.120}$$

Taylor series expansion of (C.120) around σ leads to

$$\begin{aligned} I_Q &= \sum_{k=1}^{\infty} \frac{\partial^{k-1}}{\partial \varsigma^{k-1}} P_Q(Q|\varsigma)|_{\varsigma=\sigma} \frac{\sigma^{k-1}}{(k-1)!} \overline{\Big(\frac{\varsigma-\sigma}{\sigma}\Big)^k} \\ &= \Phi\Big(\frac{Q-h(\sigma)}{g(\sigma)}\Big)\Big(\frac{\varsigma-\sigma}{\sigma}\Big) + \sum_{k=2}^{\infty} \frac{\partial^{k-1}}{\partial \varsigma^{k-1}} P_Q(Q|\varsigma)|_{\varsigma=\sigma} \frac{\sigma^{k-1}}{(k-1)!} \overline{\Big(\frac{\varsigma-\sigma}{\sigma}\Big)^k}. \end{aligned} \tag{C.121}$$

Computing the expectation and reordering according to powers of τ^2 gives

$$\begin{aligned} I_Q(Q|\sigma) &= \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n}\Big(\Phi\Big(\frac{Q-h(\sigma)}{g(\sigma)}\Big) - \sum_{k=1}^{2n-1} a_k(Q,\sigma) \sum_{l=0}^{k+1} \binom{k+1}{l} (-1)^l l^{2n}\Big) \\ &= \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n}\Big(w_n(Q|\sigma) + \Phi\Big(\frac{Q-h(\sigma)}{g(\sigma)}\Big)\Big). \end{aligned} \tag{C.122}$$

The function p_Q remains to be considered. Since $\mathrm{p}_Q = \partial/(\partial Q)\mathrm{P}_Q$, it is obtained using (C.116) as

$$\begin{aligned} \mathrm{p}_Q(Q|\sigma)\,\mathrm{d}Q &= \phi\Big(\frac{Q-h(\sigma)}{g(\sigma)}\Big) + \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} \sum_{k=1}^{2n} \frac{\partial}{\partial Q} a_k(Q,\sigma) \sum_{l=0}^{k} \binom{k}{l} (-1)^l l^{2n}\,\mathrm{d}Q \\ &= \phi\Big(\frac{Q-h(\sigma)}{g(\sigma)}\Big) + \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} v_n(Q|\sigma)\,\mathrm{d}Q. \end{aligned} \tag{C.123}$$

The product of (C.122) and (C.123) reads

$$\begin{aligned} \mathrm{p}_Q(Q|\sigma) I_Q(Q|\sigma)\,\mathrm{d}Q &= \phi\Big(\frac{Q-h(\sigma)}{g(\sigma)}\Big)\Phi\Big(\frac{Q-h(\sigma)}{g(\sigma)}\Big) \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n}\,\mathrm{d}Q \\ &\quad + \phi\Big(\frac{Q-h(\sigma)}{g(\sigma)}\Big) \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} w_n(Q|\sigma)\,\mathrm{d}Q \\ &\quad + \Phi\Big(\frac{Q-h(\sigma)}{g(\sigma)}\Big) \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} v_n(Q|\sigma)\,\mathrm{d}Q \\ &\quad + \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} w_n(Q|\sigma) \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} v_n(Q|\sigma)\,\mathrm{d}Q. \end{aligned} \tag{C.124}$$

Now, the integration variable Q is transformed to $t = (Q - h(\sigma))/g(\sigma)$. We arrive at

$$\begin{aligned} \mathrm{p}_t(t|\sigma) I_t(t|\sigma)\,\mathrm{d}t &= \phi(t)\Phi(t) \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n}\,\mathrm{d}t \\ &\quad + \phi(t) \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} w_n(t|\sigma)\,\mathrm{d}t \\ &\quad + \Phi(t) g(\sigma) \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} v_n(t|\sigma)\,\mathrm{d}t \\ &\quad + \sum_{n=1}^{\infty} g(\sigma) \frac{\tau^{2n}}{n!2^n} w_n(t|\sigma) \sum_{n=1}^{\infty} \frac{\tau^{2n}}{n!2^n} v_n(t|\sigma)\,\mathrm{d}t. \end{aligned} \tag{C.125}$$

Since

$$\phi\Big(\frac{Q-h(\sigma)}{g(\sigma)}\Big) = \frac{1}{g(\sigma)\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{Q-h(\sigma)}{g(\sigma)}\right)^2}, \tag{C.126}$$

it follows that

$$\phi(t)\,\mathrm{d}t = \frac{\mathrm{e}^{-\frac{t^2}{2}}}{\sqrt{2\pi}} = g(\sigma)\phi\Big(\frac{Q-h(\sigma)}{g(\sigma)}\Big). \tag{C.127}$$

Furthermore, $\mathrm{d}Q = g(\sigma)\,\mathrm{d}t$ holds and leads together with (C.127) to (C.125). It remains to compute the product of (C.119) and (C.124) $\Big(1 - \mathrm{P}_t(t|\sigma)\Big)^{\lambda-\mu-1}\mathrm{P}_t(t|\sigma)^{\mu-1}\mathrm{p}_t(t|\sigma)I_t(t|\sigma)\,\mathrm{d}t$. Again, four terms can be distinguished

$$\begin{aligned} I_1(t|\sigma)\,\mathrm{d}t &= \phi(t)\Phi(t)\sum_{n=1}^{\infty}\frac{\tau^{2n}}{n!2^n}\sum_{l=0}^{\lambda-\mu-1}\sum_{m=0}^{\mu-1}\binom{\lambda-\mu-1}{l}\binom{\mu-1}{m} \\ &\quad\times\Big(1-\Phi(t)\Big)^{\lambda-\mu-1-l}\Phi(t)^{\mu-1-m}(-1)^l\Big(\sum_{n=1}^{\infty}\frac{\tau^{2n}}{n!2^n}c_n(t,\sigma)\Big)^{m+l}\,\mathrm{d}t \\ &= \phi(t)\sum_{n=1}^{\infty}\frac{\tau^{2n}}{n!2^n}\sum_{l=0}^{\lambda-\mu-1}\sum_{m=0}^{\mu-1}\binom{\lambda-\mu-1}{l}\binom{\mu-1}{m} \\ &\quad\times\Big(1-\Phi(t)\Big)^{\lambda-\mu-1-l}\Phi(t)^{\mu-m}(-1)^l\Big(\sum_{n=1}^{\infty}\frac{\tau^{2n}}{n!2^n}c_n(t,\sigma)\Big)^{m+l}\,\mathrm{d}t, \quad \text{(C.128)} \end{aligned}$$

$$\begin{aligned} I_2(t|\sigma)\,\mathrm{d}t &= \phi(t)\sum_{n=1}^{\infty}\frac{\tau^{2n}}{n!2^n}w_n(t|\sigma)\sum_{l=0}^{\lambda-\mu-1}\sum_{m=0}^{\mu-1}\binom{\lambda-\mu-1}{l}\binom{\mu-1}{m} \\ &\quad\times\Big(1-\Phi(t)\Big)^{\lambda-\mu-1-l}\Phi(t)^{\mu-1-m}(-1)^l\Big(\sum_{n=1}^{\infty}\frac{\tau^{2n}}{n!2^n}c_n(t,\sigma)\Big)^{m+l}\,\mathrm{d}t, \quad \text{(C.129)} \end{aligned}$$

$$\begin{aligned} I_3(t|\sigma)\,\mathrm{d}t &= \Phi(t)g(\sigma)\sum_{n=1}^{\infty}\frac{\tau^{2n}}{n!2^n}\sum_{n=1}^{\infty}\frac{\tau^{2n}}{n!2^n}v_n(t|\sigma)\sum_{l=0}^{\lambda-\mu-1}\sum_{m=0}^{\mu-1}\binom{\lambda-\mu-1}{l}\binom{\mu-1}{m} \\ &\quad\times\Big(1-\Phi(t)\Big)^{\lambda-\mu-1-l}\Phi(t)^{\mu-1-m}(-1)^l\Big(\sum_{n=1}^{\infty}\frac{\tau^{2n}}{n!2^n}c_n(t,\sigma)\Big)^{m+l}\,\mathrm{d}t \\ &= g(\sigma)\sum_{n=1}^{\infty}\frac{\tau^{2n}}{n!2^n}\sum_{n=1}^{\infty}\frac{\tau^{2n}}{n!2^n}v_n(t|\sigma)\sum_{l=0}^{\lambda-\mu-1}\sum_{m=0}^{\mu-1}\binom{\lambda-\mu-1}{l}\binom{\mu-1}{m} \\ &\quad\times\Big(1-\Phi(t)\Big)^{\lambda-\mu-1-l}\Phi(t)^{\mu-m}(-1)^l\Big(\sum_{n=1}^{\infty}\frac{\tau^{2n}}{n!2^n}c_n(t,\sigma)\Big)^{m+l}\,\mathrm{d}t, \quad \text{(C.130)} \end{aligned}$$

$$\begin{aligned} I_4(t|\sigma)\,\mathrm{d}t &= \sum_{n=1}^{\infty}g(\sigma)\frac{\tau^{2n}}{n!2^n}w_n(t|\sigma)\sum_{n=1}^{\infty}\frac{\tau^{2n}}{n!2^n}v_n(t|\sigma)\sum_{l=0}^{\lambda-\mu-1}\sum_{m=0}^{\mu-1}\binom{\lambda-\mu-1}{l}\binom{\mu-1}{m} \\ &\quad\times\Big(1-\Phi(t)\Big)^{\lambda-\mu-1-l}\Phi(t)^{\mu-1-m}(-1)^l\Big(\sum_{n=1}^{\infty}\frac{\tau^{2n}}{n!2^n}c_n(t,\sigma)\Big)^{m+l}\,\mathrm{d}t \quad \text{(C.131)} \end{aligned}$$

with

$$\psi(\sigma) = (\lambda-\mu)\binom{\lambda}{\mu}\int_{-\infty}^{\infty} I_1(t) + I_2(t) + I_3(t) + I_4(t)\,\mathrm{d}t. \quad \text{(C.132)}$$

The aim is now to give ψ (C.132) up to a precision of τ^{2K}. As it can be seen easily, the summation over l and m can be cut off after $\min\{K, \lambda-\mu-1\}$ in the case of l and $\min\{K, \mu-1\}$ in m. This

means that (C.130) and (C.131) only contribute to the SAR if $K > 2$. As stated before, the coefficients c_n, w_n, and v_n contain products of a polynomial in t and $\exp(-t^2/2)$. Therefore, expressions similar to the definition of the progress coefficients (A.24) can be obtained. Also note that the free $g(\sigma)$-term in (C.130) and (C.131) averages out eventually. In the following the MATHEMATICA-code is described. Let us start with Fig. C.12 which defines some progress coefficients. These stem from considering the sums over m and l in Eqs. (C.128)-(C.131) which contain products of the form

$$\binom{\lambda-\mu-1}{l}\binom{\mu-1}{m}\Big(1-\Phi(t)\Big)^{\lambda-\mu-1-l}\Phi(t)^{\mu-1-m}. \tag{C.133}$$

Additionally, they are multiplicated with one or two pdfs of the standard normal distribution and with polynomials in t. In other words (C.132) contains terms of the following general form

$$\begin{aligned} e_{\mu,\lambda}^{m,l,h,k,j} &= (\lambda-\mu)\binom{\lambda}{\mu}\binom{\lambda-\mu-1}{l}\binom{\mu-1}{m} \\ &\quad \int_{-\infty}^{\infty}\Big(1-\Phi(t)\Big)^{\lambda-\mu-1-l}\Phi(t)^{\mu-1-m+j}\phi(t)^{l+m+h}t^k\,\mathrm{d}t. \end{aligned} \tag{C.134}$$

The t-dependent terms can now be expressed in terms of (C.134). It remains to determine

1. which powers of t actually appear for τ^{2k}, k fixed
2. the coefficients which are connected to t^i in τ^{2k}.

So first of all, the coefficients v_n, w_n, and c_n have to be obtained. This requires some further calculations. These coefficients are given as follows

$$\begin{aligned} v_n &:= \sum_{k=1}^{2n}\frac{\partial}{\partial t}a_k(t,\sigma)\sum_{l=0}^{k}\binom{k}{l}(-1)^l l^{2n} \\ &= \sum_{k=1}^{2n}\frac{\partial}{\partial t}a_k(t,\sigma)w_{k,n} \end{aligned} \tag{C.135}$$

$$\begin{aligned} w_n &:= -\sum_{k=1}^{2n-1}a_k(t,\sigma)\sum_{l=0}^{k+1}\binom{k+1}{l}(-1)^l l^{2n} \\ &= -\sum_{k=1}^{2n-1}a_k(t,\sigma)w_{k+1,n} \end{aligned} \tag{C.136}$$

$$\begin{aligned} c_n &:= \sum_{k=1}^{2n}a_k(t,\sigma)\sum_{l=0}^{k}\binom{k}{l}(-1)^l l^{2n} \\ &= \sum_{k=1}^{2n}a_k(t,\sigma)w_{k,n} \end{aligned} \tag{C.137}$$

```
fi[x_] := (1 + Erf[x/2^(1/2)])/2

lfi[x_] := Module[ {aaa}, aaa = N[fi[x]];
        If[ aaa =!= 0.0,  Log[aaa],
        -Log[2*Pi]/2 - x^2/2 - Log[-x]]
        ]

l1fi[x_] := Module[ {aaa}, aaa = 1-N[fi[x]];
        If[ aaa =!= 0.0,  Log[aaa],
        -Log[2*Pi]/2 - x^2/2 - Log[x]]
        ]

eijkml[i_,j_,h_,k_,w_,  mu_, lambda_] :=
  eijkml[i,j,h,k,w, mu, lambda] =
    Module[ {aa,m,l},m=mu; l=lambda; If[ l-m-i-1 == 0,
        (aa=
         Log[Binomial[l-m-1,i]*If[m-1==0,1,Binomial[m-1,j]]*Binomial[l,m]];
        (l-m)*(2*Pi)^(-(i+j+h)/2) * NIntegrate[
        If[k == 0, 1, t^k] * Exp[ -(i+j+h)/2 * t*t + aa +
        If[(l-m-i-1) == 0, 0, (l-m-i-1)*l1fi[t]] +
        If[(m-j-1-w) <= 0, 0, (m-j-1-w)*lfi[t]]],
        {t, -8, -2, 2, 8}, MaxRecursion -> 45] ),
        (aa =
         (Log[Binomial[l-m-1,i]*If[m-1==0,1,
                 Binomial[m-1,j]]*Binomial[l,m]])/(l-j-i-w-2);
        (l-m)*(2*Pi)^(-(i+j+h)/2) * NIntegrate[
        If[k == 0, 1, t^k] * Exp[ -(i+j+h)/2 * t*t +
        If[(l-m-i-1) == 0, 0, (l-m-i-1)*(aa + l1fi[t])]
        + If[(m-j-1-w) <= 0, 0, (m-j-1-w)*(aa + lfi[t])]],
        {t, -8, -2, 2, 8}, MaxRecursion -> 45] ) ]
           ]
```

Figure C.12: The MATHEMATICA source code for the coefficients $e_{\mu,\lambda}^{i,j,h,k,w}$ (C.134). The code is oriented after the MATHEMATICA code for the $e_{\mu,\lambda}^{i,j}$ coefficients of Beyer.

with $a_k(t,\sigma)$ given by (C.115), i.e., by

$$a_k(t,\sigma) \;=\; (-1)^k \frac{\partial^k}{\partial \varsigma^k} P_Q(t|\varsigma)|_{\varsigma=\sigma} \frac{\sigma^k}{k!}$$

and $w_{k,n}$ by

$$w_{k,n} \;:=\; \sum_{l=0}^{k} \binom{k}{l} (-1)^l l^{2n}. \tag{C.138}$$

Figure C.13 shows how these coefficients are obtained. Note, the coefficient `akts` gives $\partial^k/(\partial s^k)P_Q$ whereas `bkts` computes $\partial^{k+1}/(\partial t \partial s^k)P_Q$. It remains to treat the remaining sums in (C.128)-(C.131). First of all, let us consider $\sum_{i=1}^{\infty} \tau^{2i}/(i!2^i)c_i$. Of course, the series can be cut off after the wished precision is reached. In the following, let K denote the maximal power of τ^2. The sum `Pt` in Figure C.13 computes then

$$P_t \;=\; \sum_{n=1}^{K} c_n(t,\sigma) \frac{\tau^{2n}}{n!2^n}. \tag{C.139}$$

whereas `pt` gives

$$p_t = \sum_{n=1}^{K} v_n(t,\sigma)\frac{\tau^{2n}}{n!2^n}. \tag{C.140}$$

Let us now reconsider (C.121) where two sums are given – one containing $\Phi(t)$ and the other derivatives of Φ. The sum `theintPhi` determines the first, where `theint` stands for the latter.

Now the single factors can be combined (see Fig. C.14). Let us first consider a single m and l addend in (C.128)-(C.131). First of all, the product of the series in τ^2 has to be determined. Afterwards, only the terms up to the power of τ^{2K} need to be retained. The addend in (C.128)

$$\phi(t)\Phi(t)\frac{\tau^{2n}}{n!2^n}\Big(\sum_{n=1}^{K} c_n\frac{\tau^{2n}}{n!2^n}\Big)^{l+m} \tag{C.141}$$

is given by `prodphiPhi` whereas `prodphi` computes the addend in (C.129)

$$\phi(t)\sum_{n=1}^{K} w_n\frac{\tau^{2n}}{n!2^n}\Big(\sum_{n=1}^{K} c_n\frac{\tau^{2n}}{n!2^n}\Big)^{l+m}. \tag{C.142}$$

The remaining addends in (C.130)

$$\Phi(t)g(\sigma)\frac{\tau^{2n}}{n!2^n}\sum_{n=1}^{K} v_n\frac{\tau^{2n}}{n!2^n}\Big(\sum_{n=1}^{K} c_n\frac{\tau^{2n}}{n!2^n}\Big)^{l+m} \tag{C.143}$$

and (C.130)

$$\sum_{n=1}^{K} v_n\frac{\tau^{2n}}{n!2^n}\sum_{n=1}^{K} w_n\frac{\tau^{2n}}{n!2^n}\Big(\sum_{n=1}^{K} c_n\frac{\tau^{2n}}{n!2^n}\Big)^{l+m} \tag{C.144}$$

are then given by `prodPhi` and `prod`, respectively.

It remains to combine the obtained addends with the corresponding $e_{\mu,\lambda}^{i,j,h,k,w}$-coefficients (C.134). Therefore, the addends for each equation (C.128)-(C.131) are reconsidered in Figures C.15 and C.16. First, the coefficient for each τ^{2k} is obtained. Afterwards, the results are used to determine the coefficients of t^i. These are combined with the appropriate $e_{\mu,\lambda}^{i,j,h,k,w}$-coefficients before, finally, the results are gathered up again in a polynomial in τ^{2k}. The SAR can then be obtained by summing up the single addends in (C.128)-(C.131) and computing the remaining sums over m and l.

```
dim=100
d=0.2
a=2
h[s_]:=a*d*s^2/2
g[s_]:=s*Sqrt[1+d^2*a^2+a^2*d^2*s^2/(2*dim)]
akts[k_,t_,s_]:=Module[{x,y,l,w},
expo[x_, y_, l_] := D[expo[x, y, l - 1], x] -
                        expo[x, y, 0]*expo[x, y, l - 1]*expo[x, y, 1];
expo[x_, y_, 0] := g[s]/g[x]*y - (h[x] - h[s])/g[x];
expo[x_, y_, 1] := D[expo[x, y, 0],x];
w=expo[x,y,k]/.x->s/.y->t]
bkts[k_,t_,s_]:=Module[{x,w,y,l},
expo[x_, y_, l_] :=
 D[expo[x, y, l - 1], x] -
  expo[x, y, 0]*expo[x, y, l - 1]*expo[x, y, 1];
expo[x_, y_, 0] := g[s]/g[x]*y - (h[x] - h[s])/g[x];
expo[x_, y_, 1] := D[expo[x, y, 0],x];
w=D[expo[x,y,k],y]/.x->s/.y->t]
wkn[k_,n_]:=Sum[Binomial[k,l]*(-1)^l*l^(2*n),{l,0,k}]
[...]
result[n_,t_,s_,mu_,la_]:=Module[{tau,erg,as,y},
akList=Table[akts[k,y,as]*as^k*(-1)^k/k!,{k,1,2*n}];
bkList=Table[akts[k,y,as]*as^k*(-1)^k/k!,{k,1,2*n}];
ckList=Table[If[k==1,0,akts[k-1,y,as]*as^(k-1)*(-1)^k/(k-1)!],{k,1,2*n}];
cn[i_,y_,as_]:=Sum[akList[[k]]*wkn[k,i],{k,1,2*i}];
bn[i_,y_,as_]:=Sum[bkList[[k]]*wkn[k,i],{k,1,2*i}];
dn[i_,y_,as_]:=Sum[ckList[[k]]*wkn[k,i],{k,2,2*i}];
Pt[i_,tau_,as_,y_]:=If[i>0,Sum[tau^(2*j)/(j!*2^j)*cn[j,y,as],{j,1,i}],0];
pt[i_,tau_,as_,y_]:=If[i>0,Sum[tau^(2*j)/(j!*2^j)*bn[j,y,as],{j,1,i}],0];
theint[i_,tau_,as_,y_]:=If[i>0,Sum[tau^(2*(j))/((j)!*2^(j))*dn[j,y,as],
                                                          {j,1,i}],0];
theintPhi[i_,tau_,as_,y_]:=If[i>0,Sum[tau^(2*(j))/((j)!*2^(j)),{j,1,i}],0];
intErg=theint[n,tau,as,y];
intPhiErg=theintPhi[n,tau,as,y];
ptErg=pt[n,tau,as,y];
PtErg=Pt[n,tau,as,y];
erg=Sum[Sum[(-1)^l*(getPower4[n,tau,s,m,l,mu,la,PtErg,intPhiErg,y]+
        getPower3[n,tau,s,m,l,mu,la,PtErg,ptErg,intPhiErg,y]+
        getPower2[n,tau,s,m,l,mu,la,PtErg,intErg,y]+
        getPower1[n,tau,s,m,l,mu,la,PtErg,ptErg,intErg,y]),
              {l,0,Min[la-mu-1,n]}],{m,0,Min[mu-1,n]}];
erg=erg/.as->s/.tau->t
]
```

Figure C.13: The MATHEMATICA source code for obtaining the coefficients. Some lines are missing (indicated by [...]) which will be explained later.

C.4.1 Comparison with the Parabolic Ridge

Again, the parabolic ridge is taken as a test function for the SAR (C.132). Let us now compare the SAR (C.132) with the results of experiments for the parabolic ridge. Three evolution strategies were examined: a $(1, 60)$-ES, a $(10/10_I, 60)$-ES, and a $(20/20_I, 60)$-ES. The SAR was expanded up to τ^6. The ridge constant d was set to $d = 0.2$ and the distance to the ridge R was $R = 1$. In the following, the SARs are again numbered in accordance to the highest power of τ^2 in the expansion,

```
prod[n_,tau_,s_,y_,m_,l_,PtIn_,ptIn_,theIntIn_]:=Module[
                        {res,end,erg,t,ay,as},
                        If[m+l+2>n,erg=0,
                        res=PtIn^(m+l)*ptIn*theIntIn;
                        end=Min[Exponent[res,tau],2*n];
                        erg=If[end<0,0,
                            Sum[CoefficientList[res,tau][[i]]*tau^(i-1),
                                                {i,1,end+1}]]
                        ]]
prodPhi[n_,tau_,s_,y_,m_,l_,PtIn_,ptIn_,theintPhiIn_]:=Module[
                        {res,end,erg,t,ay,as},
                        If[m+l+2>n,erg=0,
                        res=PtIn^(m+l)*ptIn*theintPhiIn;
                        end=Min[Exponent[res,tau],2*n];
                        erg=If[end<0,0,
                            Sum[CoefficientList[res,tau][[i]]*tau^(i-1),
                                                {i,1,end+1}]]
                        ]]
prodphi[n_,tau_,s_,y_,m_,l_,PtIn_,theintIn_]:=Module[
                        {res,end,erg,t,ay,as},
                        If[m+l+1>n,erg=0,
                        res=PtIn^(m+l)*theintIn;
                        end=Min[Exponent[res,tau],2*n];
                        erg=If[end<0,0,
                            Sum[CoefficientList[res,tau][[i]]*tau^(i-1),
                                                {i,1,end+1}]]
                        ]]
prodphiPhi[n_,tau_,s_,y_,m_,l_,PtIn_,theintPhiIn_]:=Module[
                        {res,end,erg,t,ay,as},
                        If[m+l+1>n,erg=0,
                        res=PtIn^(m+l)*theintPhiIn;
                        end=Min[Exponent[res,tau],2*n];
                        erg=If[end<0,0,
                            Sum[CoefficientList[res,tau][[i]]*tau^(i-1),
                                                {i,1,end+1}]]
                        ]]
```

Figure C.14: The single m and l addends in (C.128)-(C.131).

i.e., ψ_i denotes the result up to the power of τ^{2i}. Figure C.17 compares the prediction with the results of experiments for $N = 100$. In the derivation of the SAR, the N-dependent version was used. Additionally, Figure C.18 shows a comparison of the two approaches. It can be seen easily that using (C.132) has no significant advantage over using (C.104) – at least up to the power of τ^6. In the case of $\mu = 10$, apparently (C.104) leads to better results. It should be noted, though, that ψ_3 obtained by (C.132) is closer to the experimental data that ψ_2. This might indicate that higher-order expansions lead to better results. Unfortunately, the MATHEMATICA-programm takes far too long to determine ψ_4.

```
getPower1[n_,tau_,s_,m_,l_,mu_,la_,PtIn_,ptIn_,theIntIn_,y_]:=Module[
                {t,tList,yList,coefList,prod,res,max,amax,end,end2,res2},
                If[n<2,erg=0,If[m+l+2>n,erg=0,
                res=prod[n,tau,s,y,m,l,PtIn,ptIn,theIntIn];
                tList=CoefficientList[res,tau];
                yList=CoefficientList[tList,y];
                end=0;
                end=Min[2*n+1,Exponent[res,tau]];
                end2=Max[end,0];
                If[end2==0,0,
                max=0;
                For[i=1;amax=0,i<Length[tList],i++;
                amax=Exponent[tList[[i]],y];If[max<amax,max=amax,max];];
                coefList=Table[eijkml[m,l,2,i-1,0,mu,la],{i,1,max+1}];
                res2=If[Length[yList]>0,Table[If[yList[[i]]!={},
                If[Length[yList[[i]]]==0,yList[[i]]*coefList[[1]],
                        Sum[yList[[i]][[j]]*coefList[[j]],
                        {j,1,
                        Min[Length[yList[[i]]],Length[coefList]]}]],0],
                                        {i,1,Length[yList]}],0];
                erg=Sum[res2[[i]]*tau^(i-1),{i,1,Length[res2]}]
                ]]]]
getPower2[n_,tau_,s_,m_,l_,mu_,la_,PtIn_,theintIn_,y_]:=Module[
                {t,tList,yList,coefList,res2,res,erg,end2,as},
                If[m+l+1>n,erg=0,
                res=prodphi[n,tau,s,y,m,l,PtIn,theintIn];
                tList=CoefficientList[res,tau];
                end=Min[2*n+1,Exponent[res,tau]];
                end2=Max[end,0];
                If[end2==0,0,
                yList=CoefficientList[tList,y];
                max=0;
                For[i=1;amax=0,i<Length[tList],i++;
                amax=Exponent[tList[[i]],y];If[max<amax,max=amax,max];];
                coefList=Table[eijkml[m,l,2,i-1,0,mu,la],{i,1,max+1}];
                res2=If[Length[yList]>0,Table[If[yList[[i]]!={},
                If[Length[yList[[i]]]==0,yList[[i]]*coefList[[1]],
                         Sum[yList[[i]][[j]]*coefList[[j]],
                        {j,1,
                        Min[Length[yList[[i]]],Length[coefList]]}]],0],
                                        {i,1,Length[yList]}],0];
                erg=Sum[res2[[i]]*tau^(i-1),{i,1,Length[res2]}]
                ]] ]
```

Figure C.15: Computing the sums I.

C.5 The Second-Order SAR for $\tau \ll 1$

In this section, the second-order SAR for $\tau \ll 1$ is derived. The second-order SAR is needed in the second-order approximation of the dynamics of self-adaptive ES. The approach followed is similar to the one used in Section C.1 for the determination of a more general first-order self-adaptation

```
getPower3[n_,tau_,s_,m_,l_,mu_,la_,PtIn_,ptIn_,theintPhiIn_,y_]:=Module[
        {t,tList,yList,coefList,res2,res,erg,end2,as},
        If[n<2,erg=0,If[m+l+2>n,erg=0,
        res=prodPhi[n,tau,s,y,m,l,PtIn,ptIn,theintPhiIn];
        tList=CoefficientList[res,tau];
        end=Min[2*n+1,Exponent[res,tau]];
        end2=Max[end,0];
        If[end2==0,0,
        yList=CoefficientList[tList,y];
        max=0;
        For[i=1;amax=0,i<Length[tList],i++;
        amax=Exponent[tList[[i]],y];If[max<amax,max=amax,max];];
        coefList=Table[eijkml[m,l,1,i-1,(-1),mu,la],{i,1,max+1}];
        res2=If[Length[yList]>0,Table[If[yList[[i]]!={},
                If[Length[yList[[i]]]==0,yList[[i]]*coefList[[1]],
                   Sum[yList[[i]][[j]]*coefList[[j]],
                   {j,1,
                    Min[Length[yList[[i]]],Length[coefList]]}]],0],
                                        {i,1,Length[yList]}],0];
        erg=Sum[res2[[i]]*tau^(i-1),{i,1,Length[res2]}];
        erg/.as->s/.t->tau]]]
]
getPower4[n_,tau_,s_,m_,l_,mu_,la_,PtIn_,theintPhiIn_,y_]:=Module[
        {t,tList,yList,coefList,res2,res,erg,end2,as},
        If[m+l+1>n,erg=0,
        res=prodphiPhi[n,tau,s,y,m,l,PtIn,theintPhiIn];
        tList=CoefficientList[res,tau];
        end=Min[2*n+1,Exponent[res,tau]];
        end2=Max[end,0];
        If[end2==0,0,
        yList=CoefficientList[tList,y];
        max=0;
        For[i=1;amax=0,i<Length[tList],i++;
        amax=Exponent[tList[[i]],y];If[max<amax,max=amax,max];];
        coefList=Table[eijkml[m,l,1,i-1,(-1),mu,la],{i,1,max+1}];
        res2=If[Length[yList]>0,Table[If[yList[[i]]!={},
        If[Length[yList[[i]]]==0,yList[[i]]*coefList[[1]],
                   Sum[yList[[i]][[j]]*coefList[[j]],
                  {j,1,
                   Min[Length[yList[[i]]],Length[coefList]]}]],0],
                                       {i,1,Length[yList]}],0];
        erg=Sum[res2[[i]]*tau^(i-1),{i,1,Length[res2]}]
]]]
```

Figure C.16: Computing the sums II.

response. The distributions considered are again the log-normal and the symmetric two-point distribution.

The second-order SAR is defined as

$$\psi^{(2)}(\langle\sigma\rangle) \quad = \quad \mathrm{E}\left[\left(\frac{\langle\varsigma\rangle - \langle\sigma\rangle}{\langle\sigma\rangle}\right)^2 |\langle\sigma\rangle\right]. \tag{C.145}$$

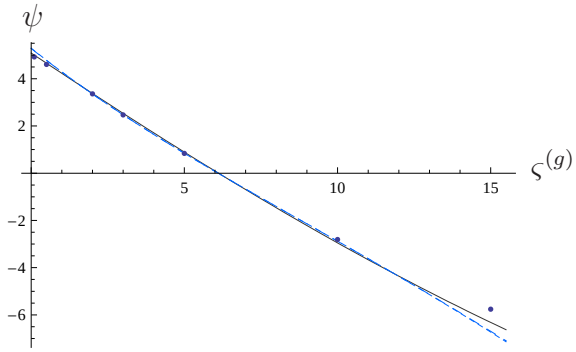
Again, the further dependencies of $\psi^{(2)}$ which may include the distance to the optimizer or the noise strength are not denoted at this point. Considering (C.145) and performing the multiplication,

$$\psi^{(2)}(\langle\sigma\rangle, R, \langle\sigma\rangle_\epsilon) \quad = \quad \frac{1}{\langle\sigma\rangle^2}\mathrm{E}\left[\langle\varsigma\rangle^2 - 2\langle\varsigma\rangle + \langle\sigma\rangle^2\right] \tag{C.146}$$

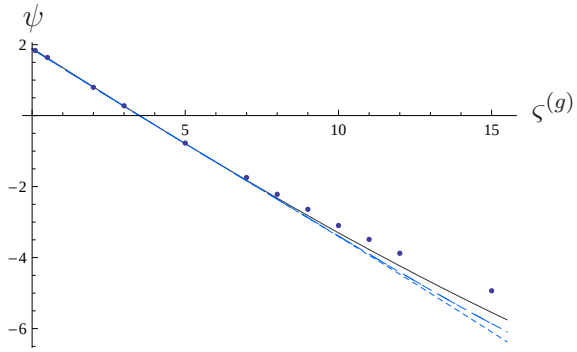
needs to be computed. Since $\langle\varsigma\rangle = 1/\mu \sum_{m=1}^{\mu} \varsigma_{m;\lambda}$, the terms inside the expectation can be split into

$$\begin{aligned}\langle\varsigma\rangle^2 - 2\langle\varsigma\rangle + \langle\sigma\rangle^2 \quad &= \quad \frac{1}{\mu^2}\sum_{m=1}^{\mu}\varsigma_{m;\lambda}^2 + \frac{2}{\mu^2}\sum_{k=2}^{\mu}\sum_{m=1}^{k-1}\varsigma_{k;\lambda}\varsigma_{m;\lambda} \\ &\quad -\frac{2\langle\sigma\rangle}{\mu}\sum_{m=1}^{\mu}\varsigma_{m;\lambda} + \langle\sigma\rangle^2. \end{aligned} \tag{C.147}$$

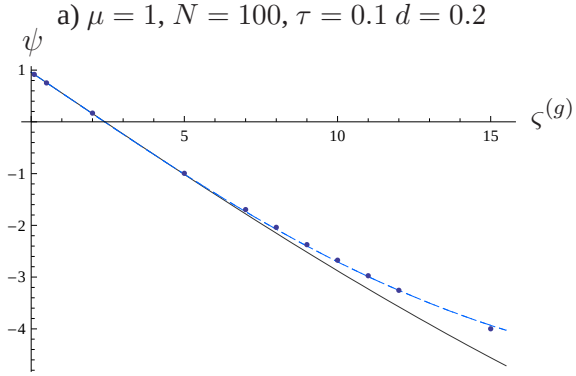
The derivation of the second order self adaptation response is straightforward. The calculations simplify considerably if (C.147) is re-expressed in terms of $(\varsigma - \langle\sigma\rangle)^k$. Since $\varsigma = (\varsigma - \langle\sigma\rangle) + \langle\sigma\rangle$ and



a) $\mu = 1$, $N = 100$, $\tau = 0.1$ $d = 0.2$



b) $\mu = 10$, $N = 100$, $\tau = 0.1$, $d = 0.2$



c) $\mu = 20$, $N = 100$, $\tau = 0.1$ $d = 0.2$

Figure C.17: Comparison of the SAR (C.132) with the results of experiments. Three SARs, ψ_1, ψ_2, and ψ_3 are shown. The solid line represents ψ_1, the dotted ψ_2 (dashed, short dots) and ψ_3 (dashed, longer dots). The results for ψ_2 and ψ_3 cannot be distinguished, since the lines nearly overlap in the case of $\mu = 1$ and $\mu = 30$. Only for $\mu = 20$, greater differences can be observed.

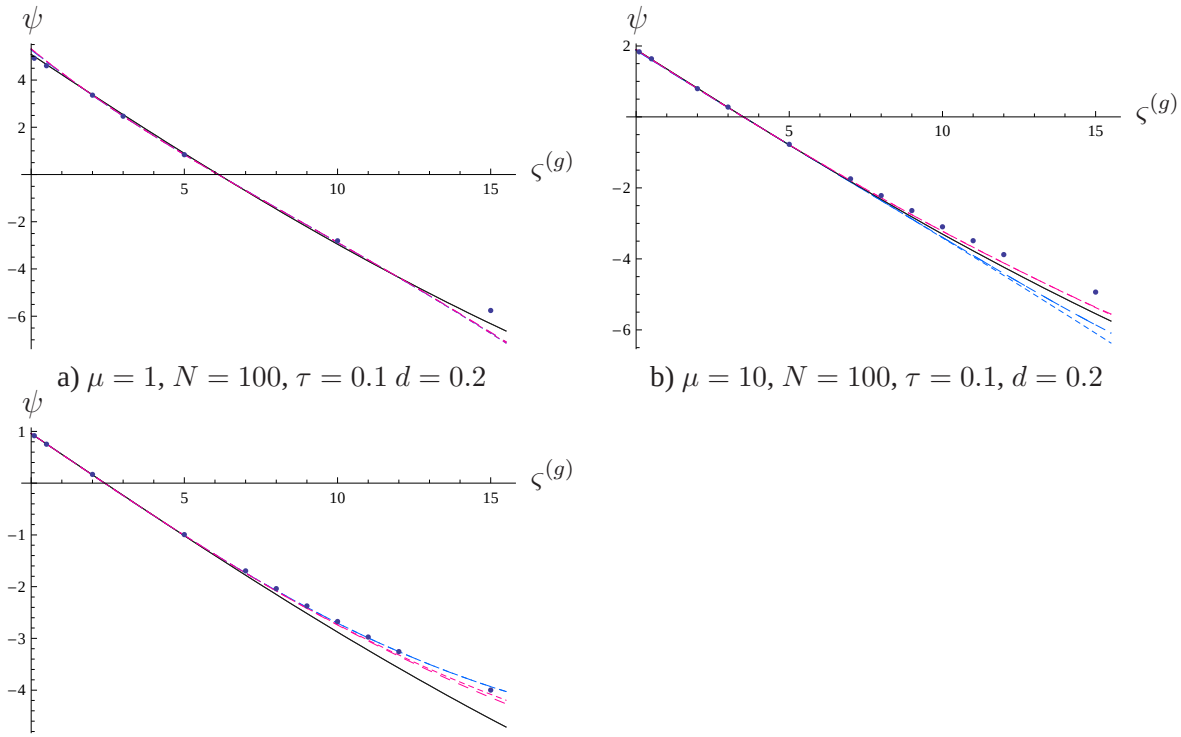


Figure C.18: Comparison of the SAR (C.132) (red lines) with (C.104) (blue lines) and the results of experiments.

$\varsigma^2 = \langle\sigma\rangle^2 - 2\langle\sigma\rangle(\varsigma - \langle\sigma\rangle) + (\varsigma - \langle\sigma\rangle)^2$ hold, the sums in Eq. (C.147) change to

$$\begin{aligned}
\frac{1}{\mu^2}\sum_{m=1}^{\mu}\varsigma_{m;\lambda}^2 &= \frac{1}{\mu^2}\sum_{m=1}^{\mu}\langle\sigma\rangle^2 - 2\langle\sigma\rangle(\varsigma_{m;\lambda} - \langle\sigma\rangle) + (\varsigma_{m;\lambda} - \langle\sigma\rangle)^2 \\
&= \frac{\langle\sigma\rangle^2}{\mu} - \frac{2\langle\sigma\rangle}{\mu^2}\sum_{m=1}^{\mu}(\varsigma_{m;\lambda} - \langle\sigma\rangle) + \frac{1}{\mu^2}\sum_{m=1}^{\mu}(\varsigma_{m;\lambda} - \langle\sigma\rangle)^2 \qquad \text{(C.148)} \\
\frac{2}{\mu^2}\sum_{k=2}^{\mu}\sum_{m=1}^{k-1}\varsigma_{k;\lambda}\varsigma_{m;\lambda} &= \frac{2}{\mu^2}\frac{\mu(\mu-1)}{2}\langle\sigma\rangle^2 + \frac{2}{\mu^2}(\mu-1)\langle\sigma\rangle\sum_{k=1}^{\mu}(\varsigma_{m;\lambda} - \langle\sigma\rangle) \\
&\quad + \frac{2}{\mu^2}\sum_{k=2}^{\mu}\sum_{m=1}^{k-1}(\varsigma_{k;\lambda} - \langle\sigma\rangle)(\varsigma_{m;\lambda} - \langle\sigma\rangle) \\
&= \langle\sigma\rangle^2 - \frac{\langle\sigma\rangle^2}{\mu} + \frac{2}{\mu}\langle\sigma\rangle\sum_{k=1}^{\mu}(\varsigma_{k;\lambda} - \langle\sigma\rangle) \\
&\quad - \frac{2}{\mu^2}\langle\sigma\rangle\sum_{k=1}^{\mu}(\varsigma_{k;\lambda} - \langle\sigma\rangle) \\
&\quad + \frac{2}{\mu^2}\sum_{k=2}^{\mu}\sum_{m=1}^{k-1}(\varsigma_{k;\lambda} - \langle\sigma\rangle)(\varsigma_{m;\lambda} - \langle\sigma\rangle) \qquad \text{(C.149)}
\end{aligned}$$

$$\frac{2\langle\sigma\rangle}{\mu}\sum_{m=1}^{\mu}\varsigma_{m;\lambda} = \frac{2\langle\sigma\rangle}{\mu}\sum_{m=1}^{\mu}(\varsigma_{m;\lambda}-\langle\sigma\rangle)+2\langle\sigma\rangle^2. \tag{C.150}$$

The result (C.149) can be easily obtained by considering

$$\begin{aligned}
\sum_{k=2}^{\mu}\sum_{m=1}^{k-1}(\varsigma_{k;\lambda}-\langle\sigma\rangle)(\varsigma_{m;\lambda}-\langle\sigma\rangle) &= \sum_{k=2}^{\mu}\sum_{m=1}^{k-1}\varsigma_{k;\lambda}\varsigma_{m;\lambda}-\langle\sigma\rangle\sum_{k=2}^{\mu}\sum_{m=1}^{k-1}\varsigma_{k;\lambda}-\langle\sigma\rangle\sum_{k=2}^{\mu}\sum_{m=1}^{k-1}\varsigma_{m;\lambda}\\
&\quad+\langle\sigma\rangle^2\sum_{k=2}^{\mu}\sum_{m=1}^{k-1}1\\
&= \sum_{k=2}^{\mu}\sum_{m=1}^{k-1}\varsigma_{k;\lambda}\varsigma_{m;\lambda}-\langle\sigma\rangle\sum_{k=2}^{\mu}\varsigma_{k;\lambda}\sum_{m=1}^{k-1}1-\langle\sigma\rangle\sum_{m=1}^{\mu-1}\varsigma_{m;\lambda}\sum_{k=m+1}^{\mu}1\\
&\quad+\langle\sigma\rangle^2\sum_{k=2}^{\mu}\sum_{m=1}^{k-1}1\\
&= \sum_{k=2}^{\mu}\sum_{m=1}^{k-1}\varsigma_{k;\lambda}\varsigma_{m;\lambda}-\langle\sigma\rangle\sum_{k=2}^{\mu}(k-1)\varsigma_{k;\lambda}-\langle\sigma\rangle\sum_{k=1}^{\mu-1}(\mu-k)\varsigma_{k;\lambda}\\
&\quad+\langle\sigma\rangle^2\frac{\mu(\mu-1)}{2}\\
&= \sum_{k=2}^{\mu}\sum_{m=1}^{k-1}\varsigma_{k;\lambda}\varsigma_{m;\lambda}-\langle\sigma\rangle\sum_{k=2}^{\mu-1}(k-1+\mu-k)\varsigma_{k;\lambda}\\
&\quad-\langle\sigma\rangle(\mu-1)\varsigma_{1;\lambda}-\langle\sigma\rangle(\mu-1)\varsigma_{\mu;\lambda}+\langle\sigma\rangle^2\frac{\mu(\mu-1)}{2}\\
&= \sum_{k=2}^{\mu}\sum_{m=1}^{k-1}\varsigma_{k;\lambda}\varsigma_{m;\lambda}-\langle\sigma\rangle(\mu-1)\sum_{k=1}^{\mu}(\varsigma_{k;\lambda}-\langle\sigma\rangle)\\
&\quad-\langle\sigma\rangle^2\frac{\mu(\mu-1)}{2}.
\end{aligned} \tag{C.151}$$

As a result, the expression $\langle\varsigma\rangle^2-2\langle\varsigma\rangle+\langle\sigma\rangle^2$ simplifies considerably to

$$\begin{aligned}
\langle\varsigma\rangle^2-2\langle\varsigma\rangle+\langle\sigma\rangle^2 &= \langle\sigma\rangle^2+\frac{\langle\sigma\rangle^2}{\mu}-\frac{2\langle\sigma\rangle}{\mu^2}\sum_{m=1}^{\mu}(\varsigma_{m;\lambda}-\langle\sigma\rangle)+\frac{1}{\mu^2}\sum_{m=1}^{\mu}(\varsigma_{m;\lambda}-\langle\sigma\rangle)^2\\
&\quad\langle\sigma\rangle^2-\frac{\langle\sigma\rangle^2}{\mu}+\frac{2}{\mu}\langle\sigma\rangle\sum_{k=1}^{\mu}(\varsigma_{m;\lambda}-\langle\sigma\rangle)-\frac{2}{\mu^2}\langle\sigma\rangle\sum_{k=1}^{\mu}(\varsigma_{m;\lambda}-\langle\sigma\rangle)\\
&\quad+\frac{2}{\mu^2}\sum_{k=2}^{\mu}\sum_{m=1}^{k-1}(\varsigma_{k;\lambda}-\langle\sigma\rangle)(\varsigma_{m;\lambda}-\langle\sigma\rangle)-\frac{2\langle\sigma\rangle}{\mu}\sum_{m=1}^{\mu}(\varsigma_{m;\lambda}-\langle\sigma\rangle)-2\langle\sigma\rangle^2\\
&= \frac{1}{\mu^2}\sum_{m=1}^{\mu}(\varsigma_{m;\lambda}-\langle\sigma\rangle)^2+\frac{2}{\mu^2}\sum_{k=2}^{\mu}\sum_{m=1}^{k-1}(\varsigma_{k;\lambda}-\langle\sigma\rangle)(\varsigma_{m;\lambda}-\langle\sigma\rangle).
\end{aligned} \tag{C.152}$$

Let us consider the expectation of $1/\mu^2 \sum_{m=1}^{\mu}(\varsigma_{m;\lambda} - \langle\sigma\rangle)^2$

$$\begin{aligned}\mathrm{E}\left[\frac{\langle\sigma\rangle^2}{\mu^2}\sum_{m=1}^{\mu}\Big(\frac{\varsigma_{m;\lambda}-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)^2\right] &= \frac{\lambda!}{\mu^2}\sum_{m=1}^{\mu}\int_{-\infty}^{\infty}\int_0^{\infty}\frac{\Big(\frac{\varsigma_{m;\lambda}-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)^2}{(\lambda-m-1)!(m-1)!}p_\sigma^*(\varsigma|\langle\sigma\rangle)p(Q|\varsigma)\\ &\quad\times\Big(1-P(Q|\langle\sigma\rangle)\Big)^{m-1}\Big(P(Q|\langle\sigma\rangle)\Big)^{\lambda-m}\,\mathrm{d}Q\,\mathrm{d}\varsigma \qquad \text{(C.153)}\end{aligned}$$

first. As in Section C.1, the pdf and the cdf are assumed to be are given by (C.10)

$$P_Q(Q|\langle\sigma\rangle) = \Phi\Big(\frac{Q+h(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}\Big),\quad p_Q(Q|\varsigma) = \frac{1}{g(\varsigma)\sqrt{2\pi}}\mathrm{e}^{-\frac{1}{2}\left(\frac{Q+h(\varsigma)}{g(\varsigma)}\right)^2}.$$

Setting again $z = (Q + h(\langle\sigma\rangle))/g(\langle\sigma\rangle)$,

$$\begin{aligned}\mathrm{E}\left[\frac{\langle\sigma\rangle^2}{\mu^2}\sum_{m=1}^{\mu}\Big(\frac{\varsigma_{m;\lambda}-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)^2\right] &= \frac{\lambda!}{\mu^2}\sum_{m=1}^{\mu}\int_{-\infty}^{\infty}\int_0^{\infty}\frac{\Big(\frac{\varsigma_{m;\lambda}-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)^2}{(\lambda-m-1)!(m-1)!}p_\sigma(\varsigma|\langle\sigma\rangle)g(\langle\sigma\rangle)\\ &\quad\times p(-z|\langle\sigma\rangle)\Big(1-P(-z|\langle\sigma\rangle)\Big)^{m-1}\Big(P(-z|\langle\sigma\rangle)\Big)^{\lambda-m}\,\mathrm{d}z\,\mathrm{d}\varsigma\\ &= \frac{\lambda!}{\mu^2}\sum_{m=1}^{\mu}\int_0^{\infty}\frac{\Big(\frac{\varsigma_{m;\lambda}-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)^2}{(\lambda-m-1)!(m-1)!}p_\sigma(\varsigma|\langle\sigma\rangle)\\ &\quad\times\int_{-\infty}^{\infty}\frac{g(\langle\sigma\rangle)}{g(\varsigma)}\frac{\mathrm{e}^{-\frac{1}{2}\left(\frac{g(\langle\sigma\rangle)}{g(\varsigma)}z-\frac{h(\varsigma)-h(\langle\sigma\rangle)}{g(\varsigma)}\right)^2}}{\sqrt{2\pi}}\\ &\quad\times\Phi(z)^{m-1}\Big(1-\Phi(z)\Big)^{\lambda-m}\,\mathrm{d}z\,\mathrm{d}\varsigma \qquad \text{(C.154)}\end{aligned}$$

is obtained (cf.(C.11)) which is similar to (C.12). The following steps are the same as in the derivation of the first order SAR (C.12)-(C.20). Instead of (C.20), finally

$$\begin{aligned}\mathrm{E}\left[\frac{\langle\sigma\rangle^2}{\mu^2}\sum_{m=1}^{\mu}\Big(\frac{\varsigma_{m;\lambda}-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)^2\right] &= \frac{\langle\sigma\rangle^2}{\mu^2}\int_0^{\infty}\Big(\frac{\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)^2 p_\sigma(\varsigma|\langle\sigma\rangle)(\lambda-\mu)\binom{\lambda}{\mu}\\ &\quad\times\int_{-\infty}^{\infty}\Big(1-\Phi(t)\Big)^{\lambda-\mu-1}\Phi(t)^{\mu}\frac{\mathrm{e}^{-\frac{t^2}{2}}}{\sqrt{2\pi}}\,\mathrm{d}t\,\mathrm{d}\varsigma\\ &\quad+\frac{\langle\sigma\rangle^2}{\mu^2}\int_0^{\infty}\langle\sigma\rangle\Big(\frac{\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)^3 p_\sigma(\varsigma|\langle\sigma\rangle)(\lambda-\mu)\binom{\lambda}{\mu}\\ &\quad\times\int_{-\infty}^{\infty}\Big(1-\Phi(t)\Big)^{\lambda-\mu-1}\Phi(t)^{\mu-1}\frac{\mathrm{e}^{-t^2}}{2\pi}\\ &\quad\times\Big(-\frac{g'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}t-\frac{h'(\langle\sigma\rangle)}{g(\langle\sigma\rangle)}\Big)\,\mathrm{d}t\,\mathrm{d}\varsigma\\ &\quad+\frac{\langle\sigma\rangle^2}{\mu^2}\sum_{k=1}^{\infty}\int_0^{\infty}\frac{\langle\sigma\rangle^{k+1}}{(k+1)!}\Big(\frac{(\varsigma-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)^{k+3}p_\sigma(\varsigma|\langle\sigma\rangle)\\ &\quad\times(\lambda-\mu)\binom{\lambda}{\mu}\int_{-\infty}^{\infty}\Big(1-\Phi(t)\Big)^{\lambda-\mu-1}\Phi(t)^{\mu-1}\frac{\mathrm{e}^{-\frac{t^2}{2}}}{2\pi}\end{aligned}$$

$$\times\frac{\partial^k}{\partial\varsigma^k}\Bigg(\Big(\frac{g'(\varsigma)}{g^2(\varsigma)}t+\frac{g'(\varsigma)(h'(\varsigma)-h(\langle\sigma\rangle))}{g((\langle\sigma\rangle)^2}-\frac{h'(\varsigma)}{g(\varsigma)}\Big)$$
$$\times\exp\Big(-\frac{1}{2}\Big(\frac{g(\langle\sigma\rangle)t-(h(\varsigma)-h(\langle\sigma\rangle))}{g(\varsigma)}\Big)^2\Big)\Bigg)|_{\varsigma=\langle\sigma\rangle}\,\mathrm{d}t\,\mathrm{d}\varsigma \quad \text{(C.155)}$$

is obtained. The same argumentation as in the case of (C.20)f. applies to (C.155). Only the first integral has to be taken into account if $\tau \ll 1$ or $\beta \ll 1$ holds. Let us first consider the log-normal operator. Provided that the learning rate τ is small, higher order terms of τ, i.e., $\mathcal{O}(\tau^4)$, can be neglected. Taking only the value of the first integral into account, (C.155) leads to

$$\mathrm{E}\left[\frac{\langle\sigma\rangle^2}{\mu^2}\sum_{m=1}^{\mu}\Big(\frac{\varsigma_{m;\lambda}-\langle\sigma\rangle}{\langle\sigma\rangle}\Big)^2\right] = \frac{\langle\sigma\rangle^2}{\mu}\tau^2+\mathcal{O}(\tau^4). \quad \text{(C.156)}$$

The expectation of the double sum

$$I_2 = \frac{2}{\mu^2\langle\sigma\rangle^2}\mathrm{E}\left[\sum_{k=2}^{\mu}\sum_{l=1}^{k-1}(\varsigma_{k;\lambda}-\langle\sigma\rangle)(\varsigma_{l;\lambda}-\langle\sigma\rangle)\right] \quad \text{(C.157)}$$

remains to be determined. It will be shown that the contribution of I_2 may be neglected for $\tau \ll 1$. In I_2, the joint distribution of $\varsigma_{l;\lambda}$ and $s_{k;\lambda}$ needs to be taken into account. To this end, the results obtained in [23, 4] are used. W.l.o.g., let us assume that a minimization problem is considered. Using again the concept of induced order statistics, the variable $\varsigma_{l;\lambda}$ denotes the mutation strength that is associated with the apparent lth best offspring, i.e., it leads to the lth smallest apparent fitness in λ trials. Note, the lth smallest apparent fitness is associated with the lth highest quality or fitness change.

Thus, assuming that the offspring are ordered, i.e., $l < k$, $(l-1)$ offspring need to have an apparent fitness change higher than that of the lth individual. In addition, there are $k-l-1$ offspring between the lth and the kth individual. Finally, $\lambda-k$ individuals will have a smaller apparent fitness change than the kth offspring. This leads to

$$\begin{aligned} I_2 = {} & \frac{2\lambda!}{\mu^2\langle\sigma\rangle^2}\sum_{k=2}^{\mu}\sum_{l=1}^{k-1}\frac{1}{(l-1)!(\lambda-k)!(k-l-1)!}\int_0^\infty\int_0^\infty\int_{-\infty}^{\infty}\int_{-\infty}^{w}(\varsigma-\langle\sigma\rangle)(s-\langle\sigma\rangle) \\ & \times p(w|\varsigma)p(v|s)P(v|\langle\sigma\rangle)^{\lambda-k}\Big(P(w|\langle\sigma\rangle)-P(v|\langle\sigma\rangle)\Big)^{k-l-1}\Big(1-P(w|\langle\sigma\rangle)\Big)^{l-1} \\ & \times p_\sigma(\varsigma|\langle\sigma\rangle)p_\sigma(s|\langle\sigma\rangle)\,\mathrm{d}v\,\mathrm{d}w\,\mathrm{d}\varsigma\,\mathrm{d}s. \end{aligned} \quad \text{(C.158)}$$

The key point of the remaining argumentation is that the random variables ς and σ do not depend on each other. If the mutation strengths are log-normally distributed or follow a two-point distribution, similar arguments as before apply. Provided that $\tau \ll 1$ or $\beta \ll 1$, all terms in (C.158) are negligible since finally the expectation of terms of the form $(\varsigma-\langle\sigma\rangle)^k(s-\langle\sigma\rangle)^l$ has to be taken. Considering (C.158), the lowest power of the learning rate τ or β that can appear is four. The contribution of I_2 can therefore be neglected for very small values of τ.

As a result, the second order self-adaptation response is given by

$$\psi^{(2)}(\langle\sigma\rangle) = \frac{\tau^2}{\mu}+\mathcal{O}(\tau^4). \quad \text{(C.159)}$$

Equation (C.159) only holds for small τ-values and due to the derivations for the cdf of the sphere and the ridge it is applicable in large dimensional search spaces only. It is very interesting to note, that there is no influence of the distance to the optimizer (or to the ridge axis) and additionally no influence of potential noise. Furthermore, (C.159) is not influenced by the search space dimension.

In the case of the symmetric two-point operator, a similar result can be obtained. The first integral in (C.155) leads to

$$\mathrm{E}\left[\frac{\langle\sigma\rangle^2}{\mu^2}\sum_{m=1}^{\mu}\left(\frac{\varsigma_{m;\lambda}-\langle\sigma\rangle}{\langle\sigma\rangle}\right)^2\right] \quad = \quad \langle\sigma\rangle^2\frac{\beta^2}{\mu}(1+\beta)+\mathcal{O}(\beta^4) \tag{C.160}$$

and thus to the second-order self-adaptation response

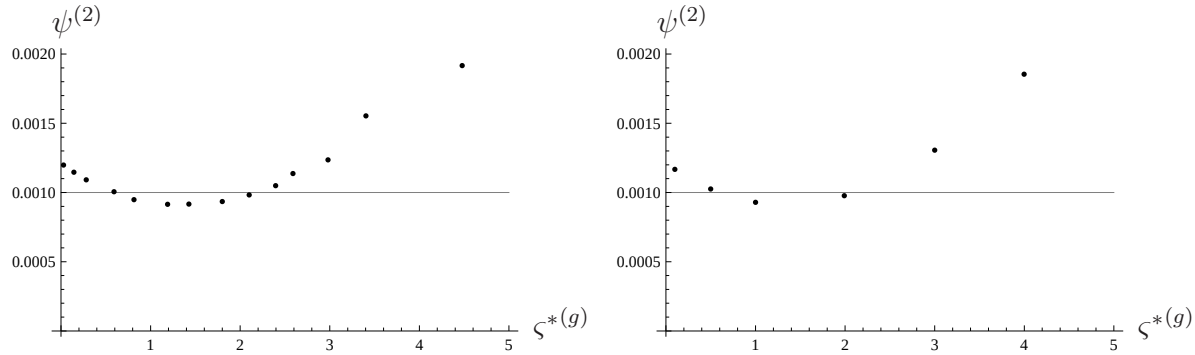
$$\psi^{(2)}(\langle\sigma\rangle) \quad = \quad \frac{\beta^2}{\mu}(1+\beta)+\mathcal{O}(\beta^4). \tag{C.161}$$

All further terms contain only higher order terms of β.

It remains to compare (C.159) with the results of experiments. Recall the fitness function of the sphere model $f(\mathbf{y}) = g(\|\mathbf{y}-\hat{\mathbf{y}}\|)$. The experiments were conducted using $g(\mathbf{y}) = -\|\mathbf{y}\|^2$.

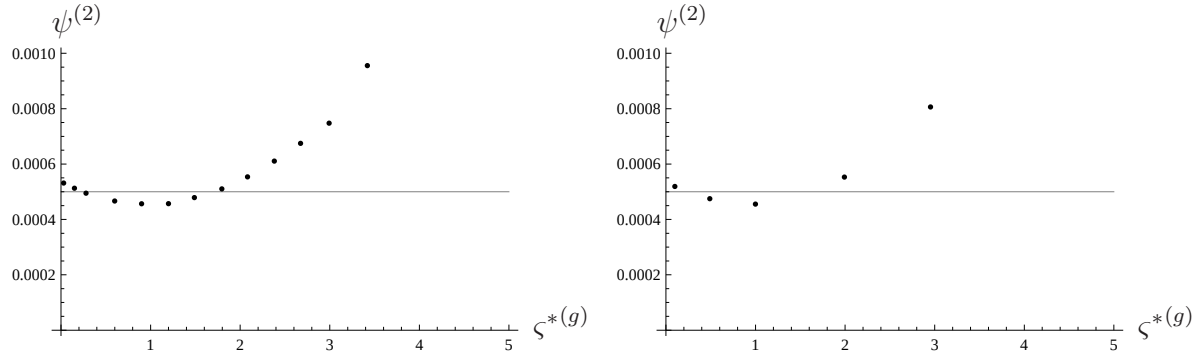
Sphere Model: Experiments for the second order SAR

Equation (C.159) was compared to the results of experiments (see Figures C.19 to C.23). The values were obtained by averaging over the results of $250,000$ one-generation experiments. As predicted, the experiments show no apparent dependency of the second order SAR on the search space dimensionality. But in contrast to the constant value (C.159) predicts, a dependency on the mutation strength can be found in the experimental data. To state it more clearly, the influence can only be neglected for small mutation strengths. The higher the mutation strength, the more the measured second-order SAR deviates from the straight line. This is more pronounced for smaller normalized noise strengths than for larger. Thus, (C.159) is strictly speaking only valid for small mutation strengths. We suspect that the reasons for this can be found in the negligence of the higher order terms of τ. If the mutation strength is increased too far, its contribution seems to outweigh the τ^4 and higher order terms. This could be amended to some extent by choosing smaller τ-values or of course by taking higher order terms of τ into account.



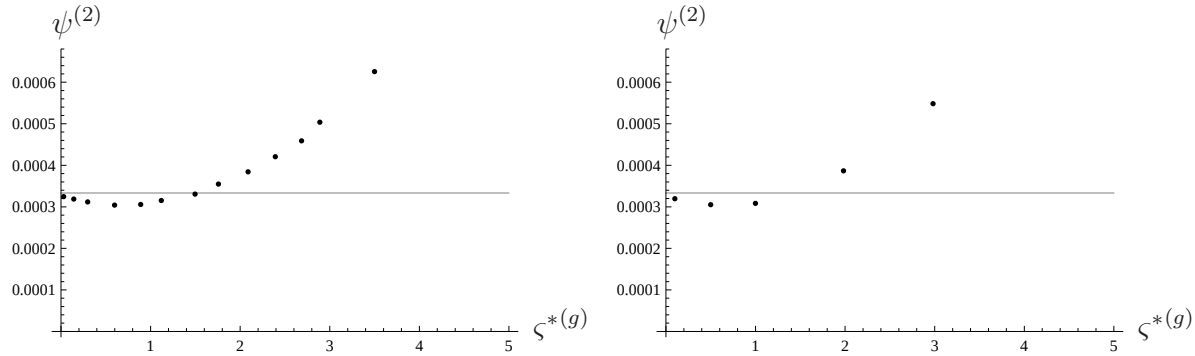
a) $\mu = 10$, $N = 30$, $\tau = 0.1$, $\sigma_\epsilon^* = 0$

b) $\mu = 10$, $N = 100$, $\tau = 0.1$, $\sigma_\epsilon^* = 0$



c) $\mu = 20$, $N = 30$, $\tau = 0.1$, $\sigma_\epsilon^* = 0$

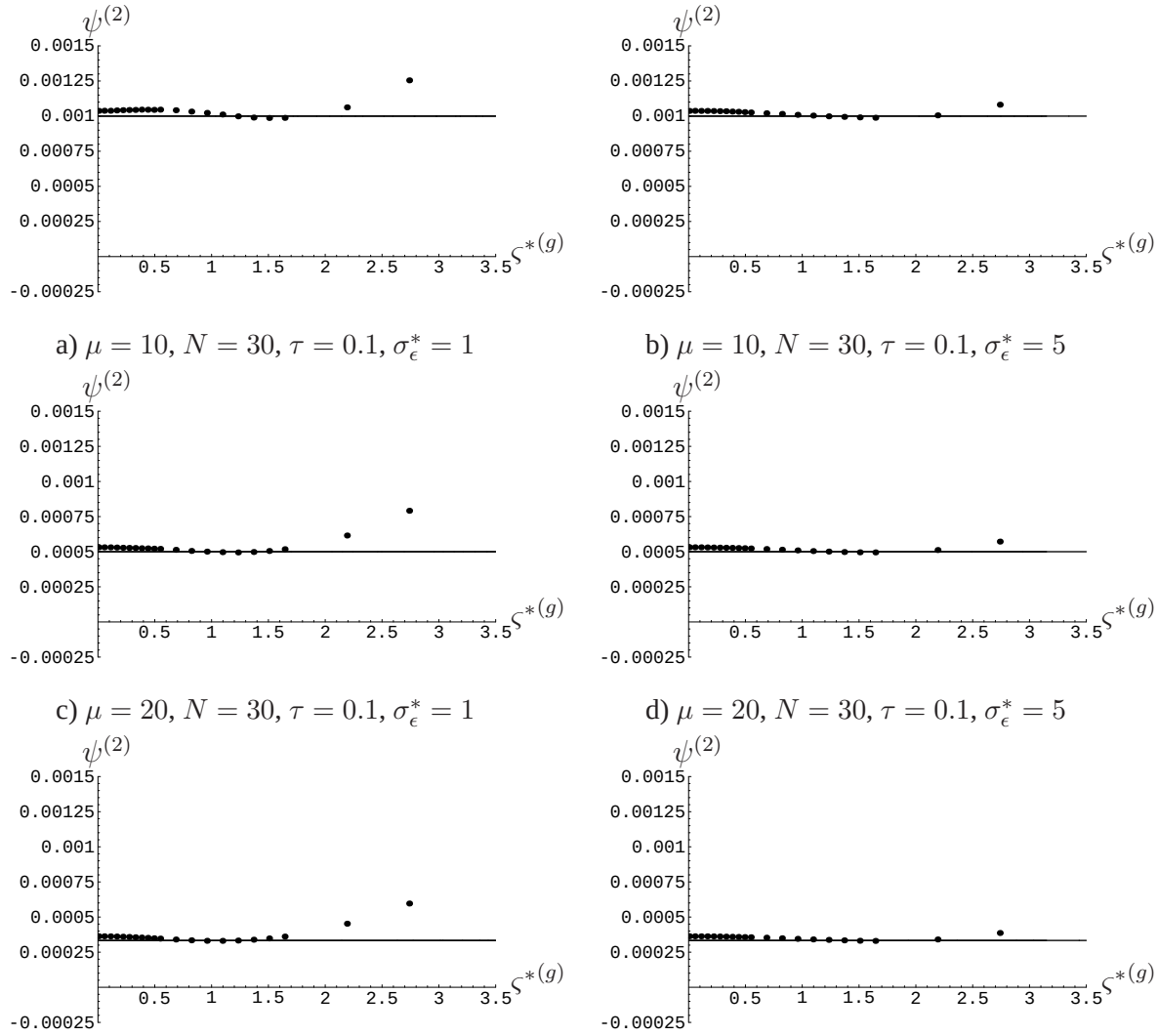
d) $\mu = 20$, $N = 100$, $\tau = 0.1$, $\sigma_\epsilon^* = 0$



c) $\mu = 30$, $N = 30$, $\tau = 0.1$, $\sigma_\epsilon^* = 0$

d) $\mu = 30$, $N = 100$, $\tau = 0.1$, $\sigma_\epsilon^* = 0$

Figure C.19: The second-order self-adaptation response function $\psi^{(2)}$ for some choices of τ and some $(\mu/\mu_I, 100)$-ES. The points denote the results of one-generation experiments and each was obtained by averaging over $250,000$ trials.



a) $\mu = 10$, $N = 30$, $\tau = 0.1$, $\sigma_\epsilon^* = 1$

b) $\mu = 10$, $N = 30$, $\tau = 0.1$, $\sigma_\epsilon^* = 5$

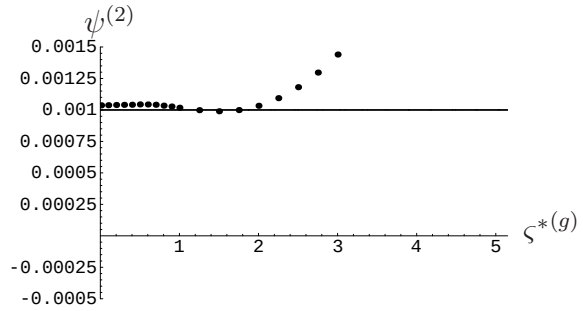
c) $\mu = 20$, $N = 30$, $\tau = 0.1$, $\sigma_\epsilon^* = 1$

d) $\mu = 20$, $N = 30$, $\tau = 0.1$, $\sigma_\epsilon^* = 5$

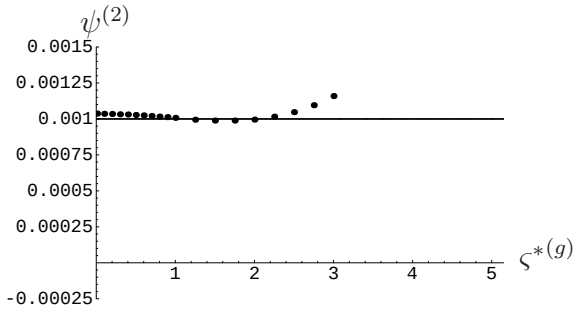
c) $\mu = 30$, $N = 30$, $\tau = 0.1$, $\sigma_\epsilon^* = 1$

d) $\mu = 30$, $N = 30$, $\tau = 0.1$, $\sigma_\epsilon^* = 5$

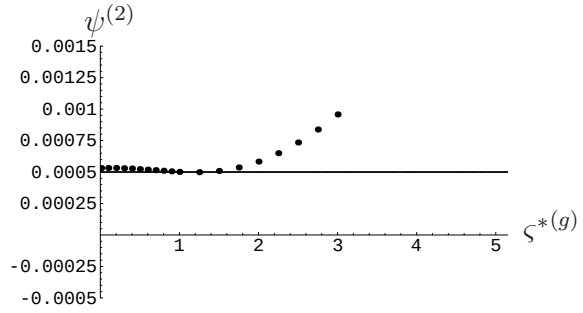
Figure C.20: The second-order self-adaptation response function $\psi^{(2)}$ for some choices of τ and some $(\mu/\mu_I, 100)$-ES. The points denote the results of one-generation experiments and each was obtained by averaging over $250,000$ trials.



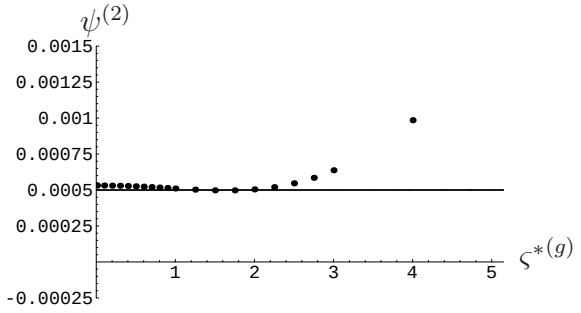
a) $\mu = 10$, $N = 100$, $\tau = 0.1$, $\sigma_\epsilon^* = 1$



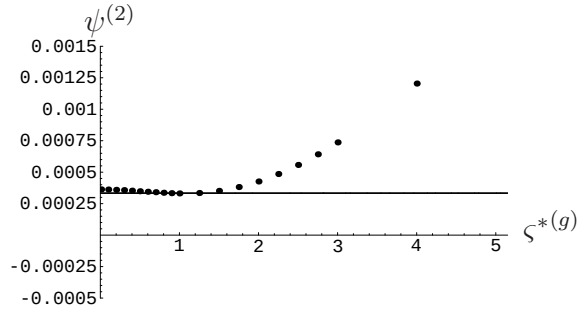
b) $\mu = 10$, $N = 100$, $\tau = 0.1$, $\sigma_\epsilon^* = 5$



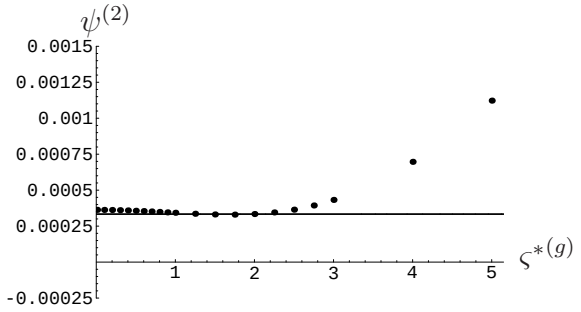
c) $\mu = 20$, $N = 100$, $\tau = 0.1$, $\sigma_\epsilon^* = 1$



d) $\mu = 20$, $N = 100$, $\tau = 0.1$, $\sigma_\epsilon^* = 5$

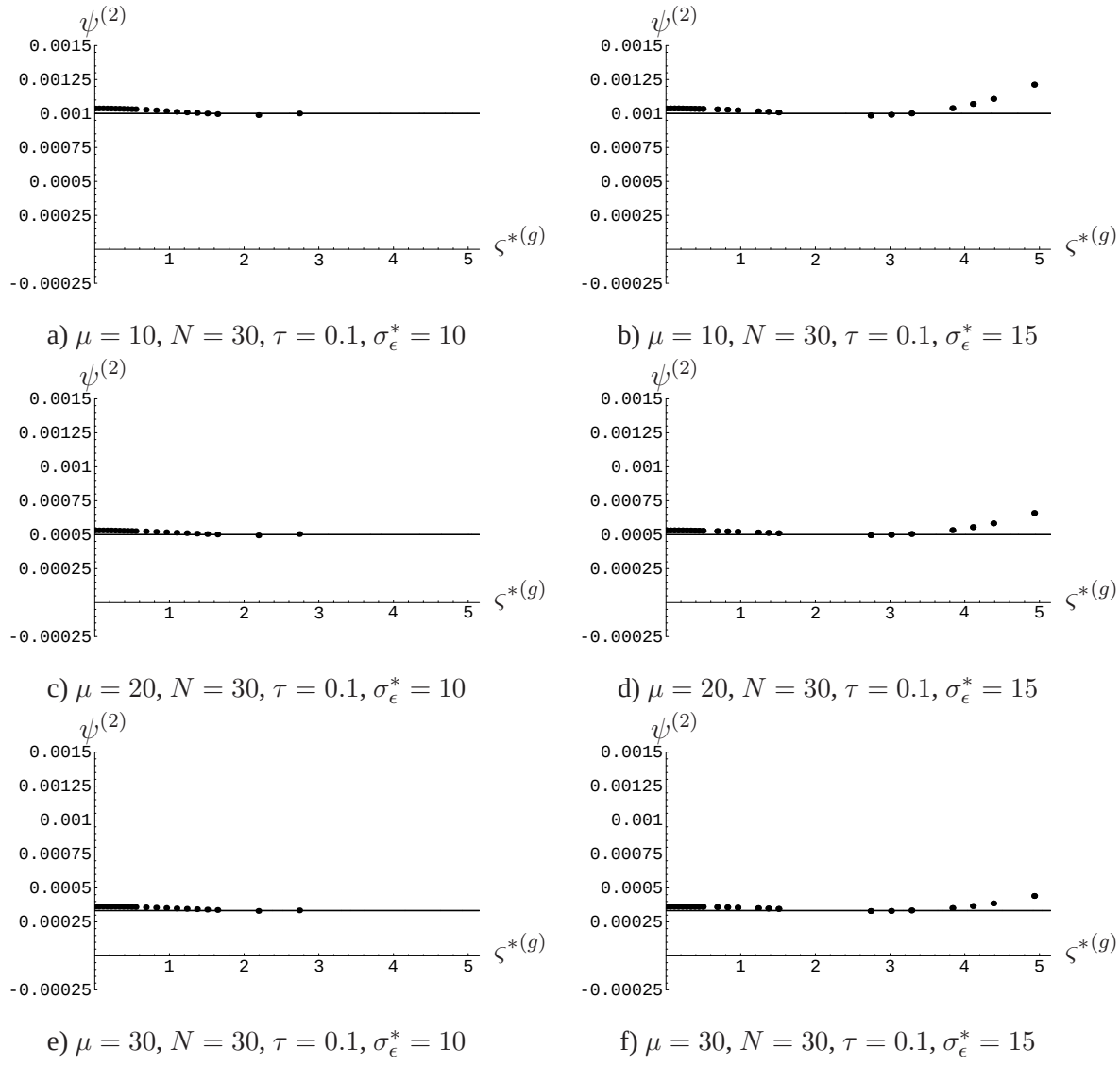


c) $\mu = 30$, $N = 100$, $\tau = 0.1$, $\sigma_\epsilon^* = 1$



d) $\mu = 30$, $N = 100$, $\tau = 0.1$, $\sigma_\epsilon^* = 5$

Figure C.21: The second-order self-adaptation response function $\psi^{(2)}$ for some choices of τ and some $(\mu/\mu_I, 100)$-ES. The points denote the results of one-generation experiments and each was obtained by averaging over $250,000$ trials.



a) $\mu = 10$, $N = 30$, $\tau = 0.1$, $\sigma_\epsilon^* = 10$

b) $\mu = 10$, $N = 30$, $\tau = 0.1$, $\sigma_\epsilon^* = 15$

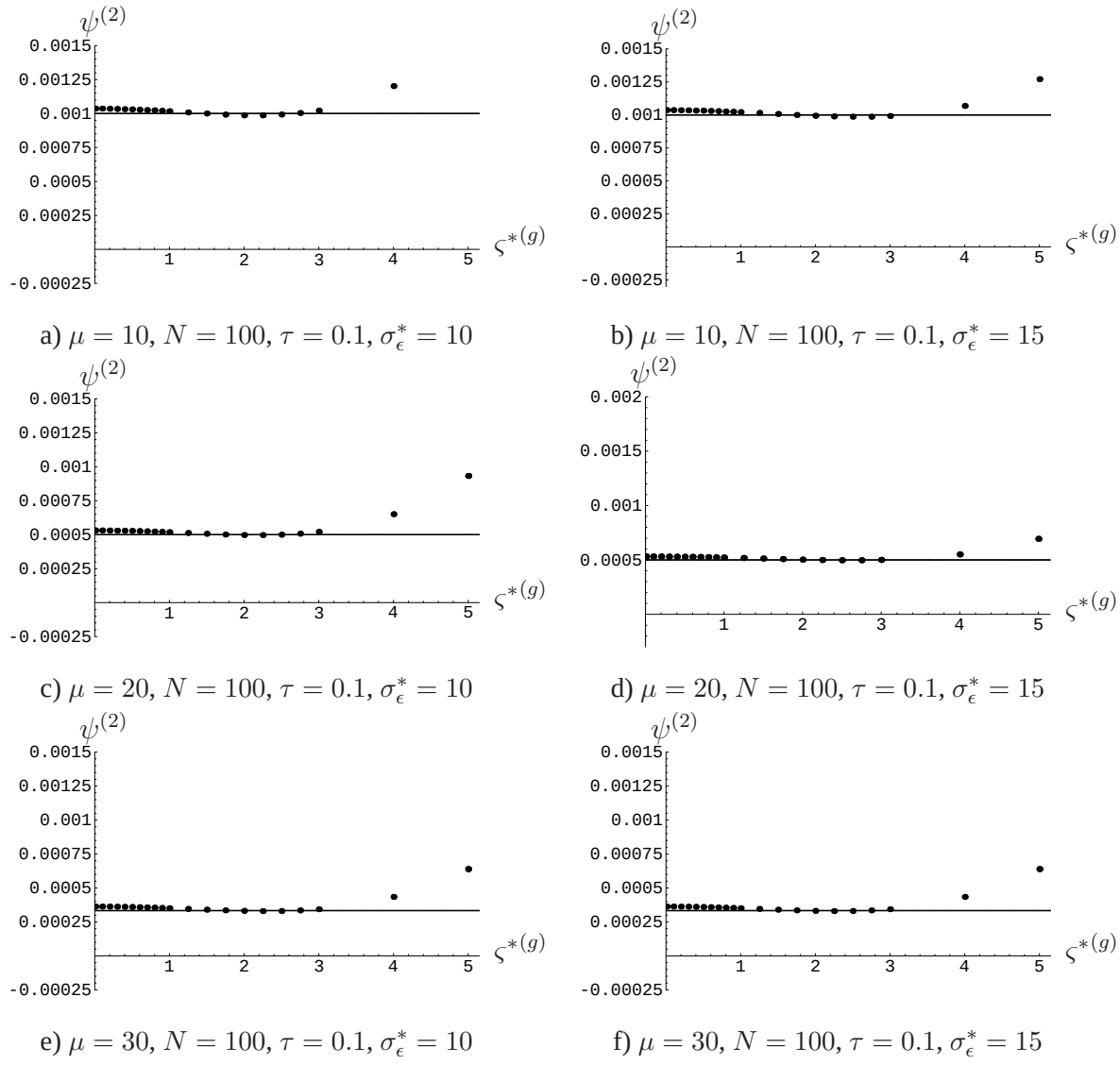
c) $\mu = 20$, $N = 30$, $\tau = 0.1$, $\sigma_\epsilon^* = 10$

d) $\mu = 20$, $N = 30$, $\tau = 0.1$, $\sigma_\epsilon^* = 15$

e) $\mu = 30$, $N = 30$, $\tau = 0.1$, $\sigma_\epsilon^* = 10$

f) $\mu = 30$, $N = 30$, $\tau = 0.1$, $\sigma_\epsilon^* = 15$

Figure C.22: The second-order self-adaptation response function for some choices of τ and some $(\mu/\mu_I, 100)$-ES. The points denote the results of one-generation experiments and each was obtained by averaging over $250,000$ trials.



a) $\mu = 10$, $N = 100$, $\tau = 0.1$, $\sigma_\epsilon^* = 10$

b) $\mu = 10$, $N = 100$, $\tau = 0.1$, $\sigma_\epsilon^* = 15$

c) $\mu = 20$, $N = 100$, $\tau = 0.1$, $\sigma_\epsilon^* = 10$

d) $\mu = 20$, $N = 100$, $\tau = 0.1$, $\sigma_\epsilon^* = 15$

e) $\mu = 30$, $N = 100$, $\tau = 0.1$, $\sigma_\epsilon^* = 10$

f) $\mu = 30$, $N = 100$, $\tau = 0.1$, $\sigma_\epsilon^* = 15$

Figure C.23: The second-order self-adaptation response function for some choices of τ and some $(\mu/\mu_I, 100)$-ES. The points denote the results of one-generation experiments and each was obtained by averaging over $250,000$ trials.

D The Sphere Model: Derivations of the Main Results

This chapter gives the details of the calculations used for obtaining the central results in Chapter 4. Its outline also follows the general outline of Chapter 4. First, $(\mu/\mu_I, \lambda)$-ES on the undisturbed sphere model are addressed – giving the derivations of the results in Section 4.1. Afterwards, the calculations leading to the results of $(1, \lambda)$-ES on the noisy sphere in Section 4.2 are presented. The remaining sections, D.2.2 and D.3, are devoted to intermediate ES on the noisy sphere, i.e., to the results in Sec. 4.3 and to the analysis including the perturbation parts in Sec. 4.4.

D.1 The Sphere Model without Noise

This section illustrates in greater detail how the results of Section 4.1 are obtained. First, the determination of the stationary points of the evolution of the mutation strength is described in D.1.1. The results obtained are then used to derive an optimal learning rate which maximizes the stationary progress rate (see D.1.2). Finally in D.1.3, it is shown that the stationary solution is stable under certain circumstances.

D.1.1 Stationary Points of the Evolution of the Mutation Strength

Consider the deterministic evolution equations (4.7), p. 34,

$$\begin{pmatrix} r \\ \langle \varsigma^{*(g+1)} \rangle \end{pmatrix} = \begin{pmatrix} R\Big(1 - \varphi_R^*(\sigma^*)/N\Big) \\ \sigma^* \Big(\frac{1+\psi(\sigma^*)}{1-\varphi_R^*(\sigma^*)/N}\Big) \end{pmatrix} \tag{D.1}$$

which describe the one-generational change of $(\mu/\mu_I, \lambda)$-ES. The progress rate appearing in (D.1) reads

$$\varphi_R^*(\sigma^*) = c_{\mu/\mu,\lambda}\sigma^* - \frac{\sigma^{*2}}{2\mu} \tag{D.2}$$

(cf. Eq. (4.8)) and the SAR is given by

$$\psi(\sigma^*) = \tau^2\Big(1/2 + e_{\mu,\lambda}^{1,1} - c_{\mu/\mu,\lambda}\sigma^*\Big) \tag{D.3}$$

(cf. (4.9)). In this section, the stationary points of the σ^*-evolution of (4.7) (or (D.1), respectively) are derived. Recall, stationary points are defined by

$$\begin{aligned} \langle \varsigma^{*(g+1)} \rangle = \sigma^* \quad &\Leftrightarrow \quad \sigma^* = 0 \bigvee \frac{1+\psi(\sigma^*)}{1-\frac{\varphi_R^*(\sigma^*)}{N}} = 1 \\ &\Leftrightarrow \quad \sigma^* = 0 \bigvee c_{\mu/\mu,\lambda}\sigma^* - \frac{\sigma^{*2}}{2\mu} = -N\tau^2\Big(1/2 + e_{\mu,\lambda}^{1,1} - c_{\mu/\mu,\lambda}\sigma^*\Big) \end{aligned} \tag{D.4}$$

(see (D.2) and (D.3)). As (D.4) shows either $\varsigma^*_{st_1} = 0$ or

$$\begin{aligned}
1 &= \left(\frac{1+\psi(\sigma^*)}{1-\frac{\varphi^*(\sigma^*)}{N}}\right) \\
\Rightarrow 1 - \frac{\varphi^*(\sigma^*)}{N} &= 1+\psi(\sigma^*) \\
\Leftrightarrow -\varphi^*(\sigma^*) &= N\psi(\sigma^*) \\
\Rightarrow -c_{\mu/\mu,\lambda}\sigma^* + \frac{\sigma^{*2}}{2\mu} &= N\tau^2\Big(1/2 + e^{1,1}_{\mu,\lambda} - c_{\mu/\mu,\lambda}\sigma^*\Big) \text{ (cf. (D.2) and (D.3))} \\
\Leftrightarrow 0 &= -2\mu N\tau^2(1/2+e^{1,1}_{\mu,\lambda}) - 2(1-N\tau^2)\mu c_{\mu/\mu,\lambda}\sigma^* + \sigma^{*2} \qquad \text{(D.5)}
\end{aligned}$$

has to hold. The positive solution of this quadratic equation is given by

$$\varsigma^*_{st_2} = \mu c_{\mu/\mu,\lambda}\left((1-N\tau^2) + \sqrt{(1-N\tau^2)^2 + 2N\tau^2\frac{1/2+e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}}}\right) \qquad \text{(D.6)}$$

which equals the non-zero stationary mutation strength (4.11), p. 35.

D.1.2 The Optimal Learning Rate

In this paragraph, the optimal learning rate for self-adaptive $(\mu/\mu_I, \lambda)$-ES is derived. The starting point is Eq. (4.11), p. 35 or (D.6), respectively,

$$\varsigma^*_{stat_2} = \mu c_{\mu/\mu,\lambda}\left((1-N\tau^2) + \sqrt{(1-N\tau^2)^2 + 2N\tau^2\frac{1/2+e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}}}\right).$$

The optimizer of the progress rate (D.2), $\varphi^*_R = c_{\mu/\mu,\lambda}\varsigma^* - \varsigma^{*2}/(2\mu)$ is given by $\varsigma^*_{\varphi^*_{R\,opt}} = \mu c_{\mu/\mu,\lambda}$. Requiring that $\varsigma^*_{stat_2}(\tau) = \varsigma^*_{\varphi^*_{R\,opt}} = \mu c_{\mu/\mu,\lambda}$ leads to (4.23), since

$$\begin{aligned}
\mu c_{\mu/\mu,\lambda} &= \mu c_{\mu/\mu,\lambda}\left((1-N\tau^2) + \sqrt{(1-N\tau^2)^2 + 2N\tau^2\frac{1/2+e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}}}\right) \\
\Rightarrow 1 &= (1-N\tau^2) + \sqrt{(1-N\tau^2)^2 + 2N\tau^2\frac{1/2+e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}}} \\
\Rightarrow (N\tau^2)^2 &= (1-N\tau^2)^2 + 2N\tau^2\frac{1/2+e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}} \\
\Leftrightarrow 0 &= 1 - 2N\tau^2\Big(1 - \frac{1/2+e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}}\Big) \\
\Rightarrow 0 &= 1 - 2N\tau^2\Big(\frac{\mu c^2_{\mu/\mu,\lambda} - 1/2 - e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}}\Big). \qquad \text{(D.7)}
\end{aligned}$$

Equation (D.7) leads to the optimal learning rate

$$\tau = \frac{1}{\sqrt{2N}}\sqrt{\frac{\mu c^2_{\mu/\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda} - 1/2 - e^{1,1}_{\mu,\lambda}}}. \tag{D.8}$$

D.1.3 Stability of the stationary mutation strength

Consider System (4.7), p. 34. It is shown in the following that the stationary mutation strength (4.11), p. 35, or (D.6), p. 194,

$$\varsigma^*_{st_2} = \mu c_{\mu/\mu,\lambda}\left((1-N\tau^2) + \sqrt{(1-N\tau^2)^2 + N\tau^2\Big(\frac{1/2 + e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}}\Big)}\right)$$

is a stable fixed point of the evolution of the mutation strength

$$\varsigma^* = \sigma^*\Big(\frac{1+\psi(\sigma^*)}{1-\frac{\varphi^*_R(\sigma^*)}{N}}\Big) =: f(\sigma^*). \tag{D.9}$$

Using Lemma 1, p. 36, i.e., showing the stability using the linear approximation, the stability criterion for the fixed point $\varsigma^*_{st_2}$ is given by $|f'(\sigma^*)|_{\sigma^*=\varsigma^*_{st_2}}| < 1$. The function f is given by

$$f(\sigma^*) = \sigma^*\left(\frac{1+\psi(\sigma^*)}{1-\frac{\varphi^*(\sigma^*)}{N}}\right) \tag{D.10}$$

with the progress rate (D.2), p.193,

$$\varphi^*_R(\sigma^*) = c_{\mu/\mu,\lambda}\sigma^* - \frac{\sigma^{*2}}{2\mu}$$

and the self-adaptation response function (D.3), p. 193,

$$\psi(\sigma^*) = \tau^2\Big(1/2 + e^{1,1}_{\mu,\lambda} - c_{\mu/\mu,\lambda}\sigma^*\Big).$$

The derivative of f reads

$$f'(\sigma^*) = \frac{1+\psi(\sigma^*)}{1-\varphi^*_R(\sigma^*)/N} + \sigma^*\Big(\frac{\psi'(\sigma^*)}{1-\varphi^*_R(\sigma^*)/N} + \frac{(1+\psi(\sigma^*))\varphi^{*\prime}_R(\sigma^*)/N}{(1-\varphi^*_R(\sigma^*)/N)^2}\Big). \tag{D.11}$$

First of all, note $\varsigma^*_{st_2}$ (D.6) is a stationary point. Therefore,

$$\frac{1+\psi(\varsigma^*_{st_2})}{1-\varphi^*_R(\varsigma^*_{st_2})/N} = 1 \tag{D.12}$$

holds. The derivative of f at $\sigma^* = \varsigma^*_{st_2}$ simplifies to

$$f'(\sigma^*)|_{\sigma^*=\varsigma^*_{st_2}} = 1 + \frac{\varsigma^*_{st_2}}{1-\varphi^*_R(\varsigma^*_{st_2})/N}\Big(\psi'(\varsigma^*_{st_2}) + \varphi^{*\prime}_R(\varsigma^*_{st_2})/N\Big). \tag{D.13}$$

Note, $\varsigma^*_{st_2} > 0$, $\varphi^*_R(\varsigma^*_{st_2}) \geq 0$, and w.l.o.g. $\varphi^*_R(\varsigma^*_{st_2}) < N$. A necessary condition for the stability of $\varsigma^*_{st_2}$ is therefore

$$\psi'(\varsigma^*_{st_2}) + \varphi^{*\prime}_R(\varsigma^*_{st_2})/N \quad < \quad 0. \tag{D.14}$$

This can be shown very easily. Since $\psi'(\varsigma^*_{st_2}) = -\tau^2 c_{\mu/\mu,\lambda}$ and $\varphi^{*\prime}_R(\varsigma^*_{st_2}) = c_{\mu/\mu,\lambda} - \varsigma^*_{st_2}/\mu$, (D.14) leads to the inequality

$$\begin{aligned}
\psi'(\varsigma^*_{st_2}) + \varphi^{*\prime}_R(\varsigma^*_{st_2})/N &< 0 \\
\Rightarrow c_{\mu/\mu,\lambda}(1/N - \tau^2) - \frac{\varsigma^*_{st_2}}{\mu N} &< 0 \\
\Leftrightarrow (1 - N\tau^2)c_{\mu/\mu,\lambda} &< \frac{\varsigma^*_{st_2}}{\mu} \\
\Leftrightarrow (1 - N\tau^2)\mu c_{\mu/\mu,\lambda} &< \varsigma^*_{st_2} \text{ with the stationary mutation strength (D.6) or (4.11)} \\
\Rightarrow (1 - N\tau^2)\mu c_{\mu/\mu,\lambda} &< \mu c_{\mu/\mu,\lambda}\left((1 - N\tau^2) + \sqrt{(1 - N\tau^2)^2 + 2N\tau^2\Big(\frac{1/2 + e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}}\Big)}\right) \\
\Rightarrow 0 &< \sqrt{(1 - N\tau^2)^2 + 2N\tau^2\Big(\frac{1/2 + e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}}\Big)}
\end{aligned} \tag{D.15}$$

which holds in general. The necessary condition is therefore fulfilled. To prove that $|f'(\sigma^*)|_{\sigma^*=\varsigma^*_{st_2}}| < 1$, it has to be shown that either $0 < f'(\sigma^*)|_{\sigma^*=\varsigma^*_{st_2}} < 1$ or $-1 < f'(\sigma^*)|_{\sigma^*=\varsigma^*_{st_2}} < 0$ holds. Let us start with $f'(\sigma^*)|_{\sigma^*=\varsigma^*_{st_2}} > 0$.

$$\begin{aligned}
f'(\sigma^*)|_{\sigma^*=\varsigma^*_{st_2}} = 1 + \frac{\varsigma^*_{st_2}}{1 - \varphi^*_R(\varsigma^*_{st_2})/N}\Big(\psi'(\varsigma^*_{st_2}) + \varphi^{*\prime}_R(\varsigma^*_{st_2})/N\Big) &> 0 \\
\Rightarrow 1 - \varphi^*_R(\varsigma^*_{st_2})/N + \varsigma^*_{st_2}\Big(\psi'(\varsigma^*_{st_2}) + \varphi^{*\prime}_R(\varsigma^*_{st_2})/N\Big) &> 0 \\
\Rightarrow N - \varphi^*_R(\varsigma^*_{st_2}) + \varsigma^*_{st_2}\Big(N\psi'(\varsigma^*_{st_2}) + \varphi^{*\prime}_R(\varsigma^*_{st_2})\Big) &> 0 \\
\Rightarrow N - c_{\mu/\mu,\lambda}\varsigma^*_{st_2} + \frac{{\varsigma^*_{st_2}}^2}{2\mu} + \varsigma^*_{st_2}\Big(-N\tau^2 c_{\mu/\mu,\lambda} + c_{\mu/\mu,\lambda} - \frac{\varsigma^*_{st_2}}{\mu}\Big) &> 0 \\
\Rightarrow N - c_{\mu/\mu,\lambda}\varsigma^*_{st_2} + \frac{{\varsigma^*_{st_2}}^2}{2\mu} + (1 - N\tau^2)c_{\mu/\mu,\lambda}\varsigma^*_{st_2} - \frac{{\varsigma^*_{st_2}}^2}{\mu} &> 0 \\
\Rightarrow N - N\tau^2 c_{\mu/\mu,\lambda}\varsigma^*_{st_2} - \frac{{\varsigma^*_{st_2}}^2}{2\mu} &> 0.
\end{aligned} \tag{D.16}$$

In order to show the last inequality, the stationary mutation strength (D.6) must be inserted into (D.16) and the result must be evaluated numerically.

Note, though, if the last inequality (D.16) is seen as a function of $\varsigma^*_{st_2}$ it is quite easy to show that (D.16) holds provided that $\tau \propto 1/\sqrt{N}$ and N is large. First of all, the last inequality of (D.16) is monotonously decreasing function of $\varsigma^*_{st_2}$. The maximal value the stationary mutation strength can assume is $\varsigma^*_{st_2} = 2\mu c_{\mu/\mu,\lambda}$. If (D.16) holds for the upper bound, it holds for all other mutation strengths given by (D.6) as well. Inserting $\varsigma^*_{st_2} = 2\mu c_{\mu/\mu,\lambda}$ into (D.16) gives

$$N - N\tau^2 2\mu c^2_{\mu/\mu,\lambda} - \frac{4\mu^2 c^2_{\mu/\mu,\lambda}}{2\mu} \quad > \quad 0$$

$$\begin{aligned} \Leftrightarrow N - (1 + N\tau^2)2\mu c^2_{\mu/\mu,\lambda} &> 0 && \text{(D.17)} \\ \Leftrightarrow N(1 - \tau^2 2\mu c^2_{\mu/\mu,\lambda}) &> 2\mu c^2_{\mu/\mu,\lambda} && \text{(D.18)} \\ \Rightarrow N &> \frac{2\mu c^2_{\mu/\mu,\lambda}}{1 - \tau^2 2\mu c^2_{\mu/\mu,\lambda}} \bigwedge \tau^2 < \frac{1}{2\mu c^2_{\mu/\mu,\lambda}} && \text{(D.19)} \end{aligned}$$

which holds for sufficiently large N provided that τ is sufficiently small or $\tau \propto 1/\sqrt{N}$. In other words, provided that the search space dimensionality is sufficiently large, it can be assumed that $f'(\sigma^*)|_{\sigma^*=\varsigma^*_{st_2}} > 0$. The question that remains is whether $f'(\sigma^*)|_{\sigma^*=\varsigma^*_{st_2}} < 1$. This condition is easily shown since it simplifies to (D.14)

$$\begin{aligned} f'(\sigma^*)|_{\sigma^*=\varsigma^*_{st_2}} = 1 + \frac{\varsigma^*_{st_2}}{1 - \varphi^*_R(\varsigma^*_{st_2})/N}\Big(\psi'(\varsigma^*_{st_2}) + \varphi^{*\prime}_R(\varsigma^*_{st_2})/N\Big) &< 1 \\ \Rightarrow \varsigma^*_{st_2}\Big(\psi'(\varsigma^*_{st_2}) + \varphi^{*\prime}_R(\varsigma^*_{st_2})/N\Big) &< 0 \\ \Rightarrow N\psi'(\varsigma^*_{st_2}) + \varphi^{*\prime}_R(\varsigma^*_{st_2}) &< 0 \end{aligned}$$

which was already shown. Note, the result is only valid in high-dimensional search spaces since $N > \varphi^*_R$ is required. A sufficient but not necessary condition is for example $N > \varphi^*_{max} = \mu c^2_{\mu/\mu,\lambda}/2$.

D.2 The Sphere Model with Noise

In this section, the derivations of the results for $(1, \lambda)$-ES and for $(\mu/\mu_I, \lambda)$-ES on the noisy sphere are given. The fitness evaluations are assumed to be disturbed by noise. The noise model applied is the standard noise model consisting of an additive normally distributed noise term with zero mean and (constant) standard deviation σ_ϵ. The derivations of this and the following sections are restricted to the quadratic sphere.

D.2.1 $(1, \lambda)$-ES on the Noisy Sphere: The Stability of the Stationary Points

This subsection describes the calculations which lead to the determination of the stationary points of the evolution equations (4.46) and (4.47)

$$\begin{aligned} \varsigma^{*(g+1)} &= \sigma^* \frac{1 + \psi(\varsigma^{*(g)}, \sigma^{*(g)}_\epsilon)}{\big(1 - \frac{1}{N}\varphi^*(\varsigma^{*(g)}, \sigma^{*(g)}_\epsilon)\big)} && \text{(D.20)} \\ \sigma^{*(g+1)}_\epsilon &= \frac{\sigma^{*(g)}_\epsilon}{\big(1 - \frac{1}{N}\varphi^*(\varsigma^{*(g)}, \sigma^{*(g)}_\epsilon)\big)^2} && \text{(D.21)} \end{aligned}$$

in Section 4.2. Taking (4.46) and (4.47) into account, there are two different pairs of equilibrium points of the evolution equations: The first with $\mathbf{e}_1 = (0, w)^{\mathrm{T}}$ with $w \in \mathbb{R}$ and ideally $w = 2c_{1,\lambda}$ and the second at $\mathbf{e}_2 = (s_2, w_2)^{\mathrm{T}}$ with s_2 given by (4.51)

$$\varsigma^*_{st} = 2c_{1,\lambda}\frac{1}{\sqrt{2(2c^2_{1,\lambda} + 1 - d^{(2)}_{1,\lambda})}} \quad \text{(D.22)}$$

and w_2 by (4.52)

$$\sigma^*_{\epsilon\, st} = 2c_{1,\lambda}\sqrt{1 - \frac{1}{2(2c^2_{1,\lambda} + 1 - d^{(2)}_{1,\lambda})}}. \quad \text{(D.23)}$$

The question arises which of these pairs is locally stable, i.e., stable w.r.t. small disturbances.

This will be shown again using a linear approximation in the vicinity of the equilibrium solution. Recall, if the general map $\mathbf{x}^{(g+1)} = f(\mathbf{x}^{(g)})$ is considered, the stability of hyperbolic fixed points can be shown via the Jacobian

$$Df(\mathbf{x})|_{\mathbf{x}=\mathbf{x}_S} = \begin{pmatrix} \frac{\partial}{\partial x_1} f_1 & \cdots & \frac{\partial}{\partial x_N} f_1 \\ \vdots & & \vdots \\ \frac{\partial}{\partial x_1} f_N & \cdots & \frac{\partial}{\partial x_N} f_N \end{pmatrix}. \tag{D.24}$$

The question, whether $\mathbf{y}$ is a stable fixed point can be solved by determining the eigenvalues of $Df|_{\mathbf{x}=\mathbf{y}}$. If an eigenvalue λ_i exists with $|\lambda_i| > 1$, then $\mathbf{y}$ is unstable [71]. Thus, the solutions of $\det(Df|_{\mathbf{x}=\mathbf{y}} - \lambda^{\mathrm{T}}\mathbf{E}) = 0$, with $\mathbf{E}$ the unity matrix, have to be determined. Considering the evolution equations (4.46) and (4.47), first the Jacobian matrix at $(\varsigma^*_\infty, \sigma^*_{\epsilon\,\infty})^{\mathrm{T}}$ of

$$f\begin{pmatrix}\varsigma^* \\ \sigma^*_\epsilon\end{pmatrix} = \begin{pmatrix} \varsigma^* \frac{1+\psi(\varsigma^*,\sigma^*_\epsilon)}{1-\varphi^*(\varsigma^*,\sigma^*_\epsilon)/N} \\ \sigma^*_\epsilon \frac{1}{(1-\varphi^*(\varsigma^*,\sigma^*_\epsilon)/N)^2} \end{pmatrix} \tag{D.25}$$

must be obtained. In general, the Jacobian of a function $f : \mathbb{R}^2 \to \mathbb{R}^2$ is given by

$$Df\begin{pmatrix}\varsigma^* \\ \sigma^*_\epsilon\end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial \varsigma^*} f_1 & \frac{\partial}{\partial \sigma^*_\epsilon} f_1 \\ \frac{\partial}{\partial \varsigma^*} f_2 & \frac{\partial}{\partial \sigma^*_\epsilon} f_2 \end{pmatrix}. \tag{D.26}$$

In the special case of the evolution equations,

$$\begin{aligned}
\frac{\partial}{\partial \varsigma^*} f_1 &= \frac{1+\psi(\varsigma^*,\sigma^*_\epsilon)}{1-\varphi^*(\varsigma^*,\sigma^*_\epsilon)/N} + \varsigma^* \Bigg(\frac{\frac{\partial}{\partial \varsigma^*}\psi(\varsigma^*,\sigma^*_\epsilon)}{1-\varphi^*(\varsigma^*,\sigma^*_\epsilon)/N} \\
&\quad + \frac{\partial}{\partial \varsigma^*}\varphi^*(\varsigma^*,\sigma^*_\epsilon) \frac{1+\psi(\varsigma^*,\sigma^*_\epsilon)}{N(1-\varphi^*(\varsigma^*,\sigma^*_\epsilon)/N)^2} \Bigg) \\
\frac{\partial}{\partial \varsigma^*} f_2 &= \sigma^*_\epsilon \frac{2\frac{\partial}{\partial \varsigma^*}\varphi^*(\varsigma^*,\sigma^*_\epsilon)}{N(1-\varphi^*(\varsigma^*,\sigma^*_\epsilon)/N)^{2+1}} \\
\frac{\partial}{\partial \sigma^*_\epsilon} f_1 &= \varsigma^* \Bigg(\frac{\frac{\partial}{\partial \sigma^*_\epsilon}\psi(\varsigma^*,\sigma^*_\epsilon)}{1-\varphi^*(\varsigma^*,\sigma^*_\epsilon)/N} \\
&\quad + \frac{\partial}{\partial \sigma^*_\epsilon}\varphi^*(\varsigma^*,\sigma^*_\epsilon) \frac{1+\psi(\varsigma^*,\sigma^*_\epsilon)}{N(1-\varphi^*(\varsigma^*,\sigma^*_\epsilon)/N)^2} \Bigg) \\
\frac{\partial}{\partial \sigma^*_\epsilon} f_2 &= \frac{1}{(1-\varphi^*(\varsigma^*,\sigma^*_\epsilon)/N)^2} + \sigma^*_\epsilon \frac{2\frac{\partial}{\partial \sigma^*_\epsilon}\varphi^*(\varsigma^*,\sigma^*_\epsilon)}{N(1-\varphi^*(\varsigma^*,\sigma^*_\epsilon)/N)^3}
\end{aligned} \tag{D.27}$$

have to be determined. The derivations of the progress rate (4.42)

$$\varphi^*_R(\sigma^*, R, \sigma^*_\epsilon) = c_{1,\lambda} \frac{\sigma^{*2}}{\sqrt{\sigma^{*2} + \sigma^{*2}_\epsilon}} - \frac{\sigma^{*2}}{2\mu} \tag{D.28}$$

and the SAR (4.43)

$$\psi(\sigma^*) = \tau^2 \Big((d^{(2)}_{1,\lambda} - 1) \frac{\sigma^{*2}}{\sigma^{*2} + \sigma^{*2}_\epsilon} - c_{1,\lambda} \frac{\sigma^{*2}}{\sqrt{\sigma^{*2} + \sigma^{*2}_\epsilon}} \Big) \tag{D.29}$$

are given by

$$\begin{aligned}
\frac{\partial}{\partial \varsigma^*}\varphi^*(\varsigma^*,\sigma_\epsilon^*) &= \frac{c_{1,\lambda}\varsigma^*}{\sqrt{\varsigma^{*2}+\sigma_\epsilon^{*2}}}\Big(2-\frac{\varsigma^{*2}}{\varsigma^{*2}+\sigma_\epsilon^{*2}}\Big)-\varsigma^* \\
\frac{\partial}{\partial \varsigma^*}\psi^*(\varsigma^*,\sigma_\epsilon^*) &= \frac{\tau^2\varsigma^*}{\sqrt{\varsigma^{*2}+\sigma_\epsilon^{*2}}}\Bigg(\frac{2(d_{1,\lambda}^{(2)}-1)}{\sqrt{\varsigma^{*2}+\sigma_\epsilon^{*2}}}(1-\frac{\varsigma^{*2}}{\varsigma^{*2}+\sigma_\epsilon^{*2}}) \\
&\quad +\frac{c_{1,\lambda}\varsigma^{*2}}{\varsigma^{*2}+\sigma_\epsilon^{*2}}-2c_{1,\lambda}\Bigg) \\
\frac{\partial}{\partial \sigma_\epsilon^*}\varphi^*(\varsigma^*,\sigma_\epsilon^*) &= -\frac{c_{1,\lambda}\sigma_\epsilon^*\varsigma^{*2}}{\sqrt{\varsigma^{*2}+\sigma_\epsilon^{*2}}^3} \\
\frac{\partial}{\partial \sigma_\epsilon^*}\psi^*(\varsigma^*,\sigma_\epsilon^*) &= \frac{\tau^2\sigma_\epsilon^*\varsigma^{*2}}{\sqrt{\varsigma^{*2}+\sigma_\epsilon^{*2}}^3}\Big(c_{1,\lambda}-2\frac{d_{1,\lambda}^{(2)}-1}{\sqrt{\varsigma^{*2}+\sigma_\epsilon^{*2}}}\Big).
\end{aligned} \tag{D.30}$$

Let us now consider the first equilibrium point $\mathbf{e}_1 = (0,w)^{\mathrm{T}}$ with $w \in \mathbb{R}$. The Jacobian at $\mathbf{e}_1$ is easily calculated as

$$Df = \begin{pmatrix} 1+\frac{\tau^2}{2} & 0 \\ 0 & 1 \end{pmatrix} \tag{D.31}$$

leading to the equation $(1+\tau^2/2-\lambda_1)(1-\lambda_2)=0$ for the eigenvalues. Unfortunately strictly speaking, a problem occurs, since one of the eigenvalues is exactly one – leading to a non-hyperbolic fixed point. Therefore in general, the stability for the fixed point cannot be examined using the linear approximation. The reason is that the nature of the eigenspace cannot be used to infer the nature of the center manifold[1] W_c of the non-linear system. Note, though, the first eigenvalue $\lambda_1 = 1+\tau^2/2$ leads to an unstable manifold W_u. The nature of the solution in W_c does not matter anymore. Any disturbance close to zero but entirely in W_u does not converge to zero: The fixed point is not stable.

The stability of the second equilibrium point can be determined by inserting (4.51) and (4.52) into the Jacobian. The expression obtained is rather clumsy, therefore, a numerically obtained plot of the eigenvalues and a range of λ-values is provided in Fig. D.1. As one can see, the larger of both eigenvalues is less than the critical value of one. Generally, the larger eigenvalue approaches 1 if $\tau \to 0$ and decreases if the learning parameter increases. This is a reasonable result: If $\tau = 0$, the mutation operator

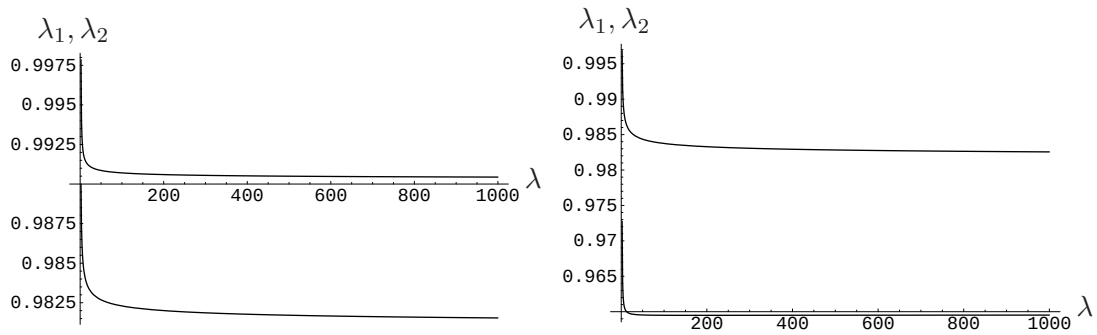
$$\sigma' = \sigma \mathrm{e}^{\tau\mathcal{N}(0,1)},$$

Eqs. (2.2) and (2.4), p. 11, does not change the mutation strength. That is, the mapping is neither contracting nor expanding. In finite dimensional search spaces and for $\tau > 0$, one can conclude that the second fixed point, where the mutation strength is given by (4.51) and the noise strength by (4.52), is locally stable – at least for the quadratic sphere.

D.2.2 Intermediate Recombination and Noisy Fitness Evaluations

This section presents the derivations of the results used in Section 4.3. That is, it describes how the stationary mutation strength (4.63), noise strength (4.64), and residual location error (4.65) are

[1] In short, a manifold can be assumed to have locally the structure of an Euclidean subspace[103]



a) $\tau = 0.1$ b) $\tau = 0.2$

Figure D.1: Numerically obtained eigenvalues of the Jacobian for the fixed point $\mathbf{e}_2$, i.e., the mutation strength given by (4.51) and the noise strength by (4.52). The search space dimension was set to $N = 100$.

derived. Since the approach is analogous to the approach used for $(1, \lambda)$-ES, only the main points are given for the sake of completeness. Starting point is the stationarity of the σ_ϵ^*-evolution in (4.60)

$$\begin{aligned} \langle \varsigma^{*(g+1)} \rangle &= \sigma^* \left(\frac{1 + \psi(\sigma^*, \sigma_\epsilon^*)}{1 - \frac{\varphi_R^*(\sigma^*, \sigma_\epsilon^*)}{N}} \right) \\ \sigma_\epsilon^{*(g+1)} &= \frac{\sigma_\epsilon^*}{(1 - \frac{\varphi_R^*(\sigma^*, \sigma_\epsilon^*)}{N})^2} \end{aligned}$$

demanding either a zero noise strength or a vanishing progress rate (4.61)

$$\begin{aligned} 0 &= \frac{\sigma^{*2}}{\sqrt{\sigma^{*2} + \sigma_\epsilon^{*2}}} c_{\mu/\mu,\lambda} - \frac{\sigma^{*2}}{2\mu} \\ \Rightarrow \sigma_{st_1}^* = 0 \bigvee 0 &= \frac{1}{\sqrt{\sigma^{*2} + \sigma_\epsilon^{*2}}} c_{\mu/\mu,\lambda} - \frac{1}{2\mu} \\ \Leftrightarrow \varsigma_{stat_1}^* = 0 \bigvee 4\mu^2 c_{\mu/\mu,\lambda}^2 &= \sigma^{*2} + \sigma_\epsilon^{*2}. \end{aligned} \tag{D.32}$$

Equation (D.32) gives a stationarity condition for the R-evolution which can be used in two ways. First of all, a maximal noise strength and with it a minimal residual location error can be obtained by setting $\sigma^* = 0$

$$\sigma_{\epsilon\, max}^* = 2\mu c_{\mu/\mu,\lambda} \tag{D.33}$$

$$\Rightarrow R_{min} = \sqrt{\frac{\sigma_\epsilon N}{4\mu c_{\mu/\mu,\lambda}}} \tag{D.34}$$

since $\sigma_\epsilon^* = \sigma_\epsilon N/(2R^2)$. Second, Eq. (D.32) can be used together with the stationarity condition of the $\langle \varsigma^{*(g)} \rangle$-evolution to derive the stationary state values of the mutation strength, distance, and noise strength. The $\langle \varsigma^{*(g)} \rangle$-evolution (4.60) is stationary if the normalized mutation strength is zero or if the SAR (4.62) vanishes, i.e., if

$$0 = \tau^2 \Big(\frac{1}{2} + e_{\mu,\lambda}^{1,1} \frac{\sigma^{*2}}{\sigma^{*2} + \sigma_\epsilon^{*2}} - c_{\mu/\mu,\lambda} \frac{\sigma^{*2}}{\sqrt{\sigma^{*2} + \sigma_\epsilon^{*2}}} \Big)$$

$$\Rightarrow 0 = \frac{1}{2} + e^{1,1}_{\mu,\lambda}\frac{\sigma^{*2}}{\sigma^{*2}+\sigma^{*2}_{\epsilon}} - c_{\mu/\mu,\lambda}\frac{\sigma^{*2}}{\sqrt{\sigma^{*2}+\sigma^{*2}_{\epsilon}}}. \tag{D.35}$$

Equation (D.35) can be used together with (D.32) to obtain the stationary mutation strength (4.63)

$$\begin{aligned} 0 &= \frac{1}{2} + e^{1,1}_{\mu,\lambda}\frac{\sigma^{*2}}{4\mu^2 c^2_{\mu/\mu,\lambda}} - c_{\mu/\mu,\lambda}\frac{\sigma^{*2}}{2\mu c_{\mu/\mu,\lambda}} \\ \Leftrightarrow -\frac{1}{2} &= \sigma^{*2}\Big(\frac{e^{1,1}_{\mu,\lambda}}{4\mu^2 c^2_{\mu/\mu,\lambda}} - \frac{c_{\mu/\mu,\lambda}}{2\mu c_{\mu/\mu,\lambda}}\Big) \\ \Leftrightarrow \sigma^{*2} &= \frac{4\mu^2 c^2_{\mu/\mu,\lambda}}{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}} \end{aligned}$$

$$\Rightarrow \varsigma^*_{st} = \sqrt{\frac{4\mu^2 c^2_{\mu/\mu,\lambda}}{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}}}. \tag{D.36}$$

The remaining stationary values are obtained by inserting (D.36) (or (4.63), respectively) into the stationarity condition (D.32). Solving the result for $\sigma^*_{\epsilon\, st}$ leads to (4.64), since

$$\begin{aligned} \sigma^*_{\epsilon\, st} &= \sqrt{4\mu^2 c^2_{\mu/\mu,\lambda} - \varsigma^{*\,2}_{st}} \\ &= \sqrt{4\mu^2 c^2_{\mu/\mu,\lambda} - \frac{4\mu^2 c^2_{\mu/\mu,\lambda}}{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}}} \end{aligned} \tag{D.37}$$

gives

$$\sigma^*_{\epsilon\, st} = 2\mu c_{\mu/\mu,\lambda}\sqrt{\frac{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1}{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}}} \tag{D.38}$$

Equation (D.38) can be used to derive the residual location error R_{st}, (4.65),

$$R_{st} = \sqrt{\frac{N\sigma_{\epsilon}}{2\sigma^*_{\epsilon\, st}}} = \sqrt{\frac{N\sigma_{\epsilon}}{4\mu c_{\mu/\mu,\lambda}}}\sqrt[4]{\frac{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1}} \tag{D.39}$$

since $\sigma^*_{\epsilon} = \sigma_{\epsilon}[N/(2R^2)]$.

D.3 The Sphere Model: A Second Order Approach

This section presents the calculations in the case of the second order approach which takes the fluctuation terms into account (cf. Section 4.4). In this section, the mean value dynamics are considered. Since the distribution cannot be obtained analytically, an alternative approach must be applied.

Following [20], a log-normal distribution is used to model the distribution of the mutation strength. This is described in D.3.2. Since the results obtained tend to differ in some case from the results observed in experiments, an alternative approach using a normal distribution is evaluated and compared to the approach using the log-normal distribution (see D.3.3).

D.3.1 Mean Value Dynamics of the Mutation Strength in the Stationary State

This section is devoted to determining the mean value dynamics of the mutation strength using the second order approach. The starting point is the evolution equation of the mutation strength (4.70), p. 59

$$\langle \varsigma^{*(g+1)} \rangle = \sigma^* \Big(\frac{1 + \psi(\sigma^*) + \epsilon_\sigma^*(R, \sigma^*)}{1 - \frac{\varphi_R^*(\sigma^*)}{N} + \epsilon_R^*(R, \sigma^*)} \Big). \tag{D.40}$$

Using the N-independent variants, the progress rate and the self-adaptation response are given by Eqs. (4.8)

$$\varphi_R(\sigma^*) = c_{\mu/\mu,\lambda}\sigma^* - \frac{\sigma^{*2}}{2\mu} \tag{D.41}$$

and (4.9)

$$\psi(\sigma^*) = \tau^2 \Big(\frac{1}{2} + e_{\mu,\lambda}^{1,1} - c_{\mu/\mu,\lambda}\sigma^* \Big). \tag{D.42}$$

The fluctuation parts are modeled using Gaussian distributions with zero mean. The variance can be determined using the evolution equations (4.67) and (4.70) (cf. Chapter 3). In the case of the R-evolution (4.67), the variance is given by

$$D_{\varphi_R}^2 = \varphi_R^{(2)} - \varphi_R^2. \tag{D.43}$$

Since the assumptions that were made during the derivation of the progress rate lead to $\varphi_R^{(2)} = \varphi_R^2$, the variance of the R-evolution is zero. Provided that these assumptions (see, e.g., [6]) are valid, deviations from the deterministic equations are mainly due to the evolution of the mutation strength. Its variance is given by

$$D_\psi^2 = \psi^{(2)} - \psi^2. \tag{D.44}$$

The second order SAR $\psi^{(2)}$ was obtained in Appendix C.5 as

$$\psi^{(2)} = \frac{\tau^2}{\mu} \tag{D.45}$$

if the higher order terms of τ are neglected. Plugging (D.45) and (D.42) into (D.44) and dropping all terms of higher than quadratic order leads to

$$D_\psi^2 = \frac{\tau^2}{\mu}. \tag{D.46}$$

The evolution equation (D.40) changes to

$$\langle \varsigma^{*(g+1)} \rangle = \sigma^* \Big(\frac{1 + \psi(\sigma^*) + \frac{\tau}{\sqrt{\mu}}\mathcal{N}(0,1)}{1 - \frac{\varphi_R^*(\sigma^*)}{N}} \Big).$$

Since $\varphi^* \ll N$ is assumed, the term $1/(1-\varphi_R^*/n)$ can be simplified to

$$\frac{1}{1-\frac{\varphi_R^*(\sigma^*)}{N}} = 1+\frac{\varphi_R^*(\sigma^*)}{N}\Big(\frac{1}{1-\frac{\varphi_R^*(\sigma^*)}{N}}\Big) = 1+\frac{\varphi_R^*(\sigma^*)}{N}+\mathcal{O}\Big((\frac{\varphi_R^*(\sigma^*)}{N})^2\Big). \quad \text{(D.47)}$$

Accordingly, Eq. (D.47) changes to

$$\langle \varsigma^{*g+1}\rangle = \sigma^*\Big(1+\psi(\sigma^*)+\frac{\tau^2}{\sqrt{\mu}}\mathcal{N}(0,1)\Big)\Big(1+\frac{\varphi_R^*(\sigma^*)}{N}\Big). \quad \text{(D.48)}$$

Under the conditions that $\psi\varphi_R^* \ll N$ and that the realizations of $\frac{\tau^2}{\sqrt{\mu}}\mathcal{N}(0,1)\varphi_R^*$ are considerably smaller than N, Eq. (D.48) can be further simplified yielding

$$\langle \varsigma^{*g+1}\rangle = \sigma^*\Big(1+\psi(\sigma^*)+\frac{\tau^2}{\sqrt{\mu}}\mathcal{N}(0,1)+\frac{\varphi_R^*(\sigma^*)}{N}\Big). \quad \text{(D.49)}$$

Equation (D.49) serves as a starting point for the determination of the moments. Plugging Eqs. (4.8) and (4.9) into (D.49) leads to the expectation

$$\begin{aligned}\mathrm{E}[\varsigma^*] &= \overline{\sigma^*}+\tau^2\Big(\frac{\overline{\sigma^*}}{2}-c_{\mu/\mu,\lambda}\overline{\sigma^{*2}}+e_{\mu,\lambda}^{1,1}\overline{\sigma^*}\Big)+\frac{c_{\mu/\mu,\lambda}}{N}\overline{\sigma^{*2}}-\frac{\overline{\sigma^{*3}}}{2\mu N}\\ &= \overline{\sigma^*}\Big(1+\tau^2\Big(\frac{1}{2}+e_{\mu,\lambda}^{1,1}\Big)\Big)-\overline{\sigma^{*2}}c_{\mu/\mu,\lambda}\tau^2\Big(1-\frac{1}{N\tau^2}\Big)-\tau^2\frac{\overline{\sigma^{*3}}}{2\mu N\tau^2}. \quad \text{(D.50)}\end{aligned}$$

As can be seen (D.50) depends on the previous values through higher-order terms. As a result, the expectations of ς^{*2} and ς^{*3} are needed. It will be shown that they in turn depend on the past through higher-order terms. The expectation of the square of (D.49) is given by

$$\begin{aligned}\mathrm{E}[\varsigma^{*2}] &= \mathrm{E}\Big[\Big(\sigma^*\Big(1+\psi(\sigma^*)+\frac{\varphi^*(\sigma^*)}{N}\Big)+\sigma^*\frac{\tau}{\sqrt{\mu}}\mathcal{N}(0,1)\Big)^2\Big]\\ &= \mathrm{E}\Big[\sigma^{*2}\Big(1+\psi(\sigma^*)+\frac{\varphi^*(\sigma^*)}{N}\Big)^2\Big]+\mathrm{E}\Big[\sigma^{*2}\frac{\tau^2}{\mu}\Big]\\ &= \mathrm{E}\Big[\sigma^{*2}\Big(1+\psi^2(\sigma^*)+\frac{\varphi^{*2}(\sigma^*)}{N^2}+2\psi(\sigma^*)+\frac{2\varphi^*(\sigma^*)}{N}+2\psi(\sigma^*)\frac{\varphi^*(\sigma^*)}{N}\Big)\Big]+\mathrm{E}\Big[\sigma^{*2}\frac{\tau^2}{\mu}\Big]\\ &= \mathrm{E}\Big[\sigma^{*2}\Big(1+2\psi(\sigma^*)+\frac{2\varphi^*(\sigma^*)}{N}\Big)\Big]+\mathrm{E}\Big[\sigma^{*2}\frac{\tau^2}{\mu}\Big]\end{aligned}$$

if $\psi^2 \ll 1$ and ${\varphi_R^*}^2 \ll N$ hold. Inserting (4.8) and (4.9) leads then to

$$\begin{aligned}\mathrm{E}[\varsigma^{*2}] &= \mathrm{E}\Big[\sigma^{*2}\Big(1+2\tau^2[\frac{1}{2}+e_{\mu,\lambda}^{1,1}-c_{\mu/\mu,\lambda}\sigma^*]\Big)\Big]+\overline{\sigma^{*2}}\frac{\tau^2}{\mu}+\mathrm{E}\Big[\sigma^{*2}\frac{2}{N}\Big(c_{\mu/\mu,\lambda}\sigma^*-\frac{\sigma^{*2}}{2\mu}\Big)\Big]\\ &= \overline{\sigma^{*2}}\Big(1+\tau^2[1+2e_{\mu,\lambda}^{1,1}+\frac{1}{\mu}]\Big)-2\overline{\sigma^{*3}}\tau^2c_{\mu/\mu,\lambda}\Big(1-\frac{1}{N\tau^2}\Big)-\overline{\sigma^{*4}}\frac{\tau^2}{\mu}\frac{1}{N\tau^2} \quad \text{(D.51)}\end{aligned}$$

which again depends on higher order terms. Similarly to (D.51), the expectation $\mathrm{E}[\varsigma^{*3}]$ can be approximated with

$$\mathrm{E}[\varsigma^{*3}] = \overline{\sigma^{*3}}\Big(1+3\frac{\tau^2}{\mu}\Big)+3\Big(1+\frac{\tau^2}{\mu}\Big)\mathrm{E}\Big[\sigma^{*3}\Big(\psi(\sigma^*)+\frac{\varphi^*(\sigma^*)}{N}\Big)\Big] \quad \text{(D.52)}$$

if $\tau \ll 1$ and $\varphi^* \ll N$ are assumed. Inserting again (4.8) and (4.9) gives

$$\begin{aligned}
\mathrm{E}[\varsigma^{*3}] &= \overline{\sigma^{*3}}\Big(1+3\frac{\tau^2}{\mu}\Big)+3\Big(1+\frac{\tau^2}{\mu}\Big)\mathrm{E}\Big[\sigma^{*3}\tau^2\Big[\frac{1}{2}+e^{1,1}_{\mu,\lambda}-c_{\mu/\mu,\lambda}\sigma^*\Big]+\frac{\sigma^{*3}}{N}\Big[c_{\mu/\mu,\lambda}\sigma^*-\frac{\sigma^{*2}}{2\mu}\Big]\Big)\Big] \\
&= \overline{\sigma^{*3}}\Big(1+3\frac{\tau^2}{\mu}\Big)+3\Big(1+\frac{\tau^2}{\mu}\Big)\tau^2\Big[\frac{\overline{\sigma^{*3}}}{2}+\overline{\sigma^{*3}}e^{1,1}_{\mu,\lambda}-c_{\mu/\mu,\lambda}\overline{\sigma^{*4}}+\frac{c_{\mu/\mu,\lambda}}{N\tau^2}\overline{\sigma^{*4}}-\frac{\overline{\sigma^{*5}}}{2\mu N\tau^2}\Big] \\
&= \overline{\sigma^{*3}}\Big(1+3\frac{\tau^2}{\mu}+3\Big(1+\frac{\tau^2}{\mu}\Big)\tau^2\Big(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\Big)\Big)-3\Big(1+\frac{\tau^2}{\mu}\Big)\tau^2 c_{\mu/\mu,\lambda}\overline{\sigma^{*4}}\Big(1-\frac{1}{N\tau^2}\Big) \\
&\quad -3\Big(1+\frac{\tau^2}{\mu}\Big)\tau^2\frac{\overline{\sigma^{*5}}}{2\mu N\tau^2} \qquad \text{(D.53)}
\end{aligned}$$

which in turn depends on higher order moments. Let us now address the stationary state behavior. As the result, $\mathrm{E}[\varsigma^*]=\mathrm{E}[\sigma^*]=\mathrm{E}[\sigma^*_\infty]$ holds. Equations (D.50), (D.51), and (D.53) lead to the non-linear equations

$$\begin{aligned}
\overline{\sigma^*_\infty} &= \overline{\sigma^*_\infty}\Big(1+\tau^2\Big(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\Big)\Big)-\overline{\sigma^{*\,2}_\infty}\tau^2 c_{\mu/\mu,\lambda}\Big(1-\frac{1}{N\tau^2}\Big)-\tau^2\frac{\overline{\sigma^{*\,3}_\infty}}{2\mu N\tau^2} \\
\Rightarrow 0 &= \overline{\sigma^*_\infty}\Big(\tau^2\Big(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\Big)\Big)-\overline{\sigma^{*\,2}_\infty}\tau^2 c_{\mu/\mu,\lambda}\Big(1-\frac{1}{N\tau^2}\Big)-\tau^2\frac{\overline{\sigma^{*\,3}_\infty}}{2\mu N\tau^2} \\
\Rightarrow 0 &= \overline{\sigma^*_\infty}\Big(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\Big)-\overline{\sigma^{*\,2}_\infty}c_{\mu/\mu,\lambda}\Big(1-\frac{1}{N\tau^2}\Big)-\frac{\overline{\sigma^{*\,3}_\infty}}{2\mu N\tau^2}, \qquad \text{(D.54)}
\end{aligned}$$

$$\begin{aligned}
\overline{\sigma^{*\,2}_\infty} &= \overline{\sigma^{*\,2}_\infty}\Big(1+\tau^2[1+2e^{1,1}_{\mu,\lambda}+\frac{1}{\mu}]\Big)-2\overline{\sigma^{*\,3}_\infty}\tau^2 c_{\mu/\mu,\lambda}\Big(1-\frac{1}{N\tau^2}\Big)-\overline{\sigma^{*\,4}_\infty}\frac{\tau^2}{\mu}\frac{1}{N\tau^2} \\
\Rightarrow 0 &= \overline{\sigma^{*\,2}_\infty}\Big(\tau^2[1+2e^{1,1}_{\mu,\lambda}+\frac{1}{\mu}]\Big)-2\overline{\sigma^{*\,3}_\infty}\tau^2 c_{\mu/\mu,\lambda}\Big(1-\frac{1}{N\tau^2}\Big)-\overline{\sigma^{*\,4}_\infty}\frac{\tau^2}{\mu}\frac{1}{N\tau^2} \\
\Rightarrow 0 &= \overline{\sigma^{*\,2}_\infty}\Big(1+2e^{1,1}_{\mu,\lambda}+\frac{1}{\mu}\Big)-2\overline{\sigma^{*\,3}_\infty}c_{\mu/\mu,\lambda}\Big(1-\frac{1}{N\tau^2}\Big)-\overline{\sigma^{*\,4}_\infty}\frac{1}{\mu N\tau^2}, \qquad \text{(D.55)}
\end{aligned}$$

and

$$\begin{aligned}
\overline{\sigma^{*\,3}_\infty} &= \overline{\sigma^{*\,3}_\infty}\Big(1+3\frac{\tau^2}{\mu}+3\Big(1+\frac{\tau^2}{\mu}\Big)\tau^2\Big(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\Big)\Big)-3\Big(1+\frac{\tau^2}{\mu}\Big)\tau^2 c_{\mu/\mu,\lambda}\overline{\sigma^{*\,4}_\infty}\Big(1-\frac{1}{N\tau^2}\Big) \\
&\quad -3\Big(1+\frac{\tau^2}{\mu}\Big)\tau^2\frac{\overline{\sigma^{*\,5}_\infty}}{2\mu N\tau^2} \\
\Rightarrow 0 &= \overline{\sigma^{*\,3}_\infty}\Big(3\frac{\tau^2}{\mu}+3\Big(1+\frac{\tau^2}{\mu}\Big)\tau^2\Big(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\Big)\Big)-3\Big(1+\frac{\tau^2}{\mu}\Big)\tau^2 c_{\mu/\mu,\lambda}\overline{\sigma^{*\,4}_\infty}\Big(1-\frac{1}{N\tau^2}\Big) \\
&\quad -3\Big(1+\frac{\tau^2}{\mu}\Big)\tau^2\frac{\overline{\sigma^{*\,5}_\infty}}{2\mu N\tau^2} \\
\Rightarrow 0 &= \overline{\sigma^{*\,3}_\infty}\Big(\frac{1}{\mu}+\Big(1+\frac{\tau^2}{\mu}\Big)\Big(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\Big)\Big)-\Big(1+\frac{\tau^2}{\mu}\Big)c_{\mu/\mu,\lambda}\overline{\sigma^{*\,4}_\infty}\Big(1-\frac{1}{N\tau^2}\Big) \qquad \text{(D.56)} \\
&\quad -\Big(1+\frac{\tau^2}{\mu}\Big)\frac{\overline{\sigma^{*\,5}_\infty}}{2\mu N\tau^2}
\end{aligned}$$

which could be solved if a solution of the eigenvalue problem (3.19) (Ch. 3, p. 30) can be given. In general this is not the case. Therefore, the *ansatz* introduced in [21] is applied.

D.3.2 A Log-Normal Distribution in the Steady State

Instead of determining the stationary distribution, the log-normal distribution is used as a placeholder. The raw moments of a log-normal distribution are of the form $\overline{\sigma_\infty^{*\,k}} = S\exp(k^2t^2/2)$. The parameters S and t remain to be determined. To this end, Eqs. (D.54) to (D.57) are used. Plugging $\overline{\sigma_\infty^{*\,k}} = S\exp(k^2t^2/2)$ into Eqs. (D.54)-(D.57) leads to

$$\begin{aligned} 0 &= S\mathrm{e}^{\frac{t^2}{2}}\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right) - S^2\mathrm{e}^{2t^2}c_{\mu/\mu,\lambda}\left(1-\frac{1}{N\tau^2}\right) - S^3\mathrm{e}^{\frac{9}{2}t^2}\frac{1}{2\mu N\tau^2} \\ \Rightarrow 0 &= \frac{1}{2}+e^{1,1}_{\mu,\lambda} - S\mathrm{e}^{\frac{3}{2}t^2}c_{\mu/\mu,\lambda}\left(1-\frac{1}{N\tau^2}\right) - S^2\mathrm{e}^{4t^2}\frac{1}{2\mu N\tau^2}, \end{aligned} \tag{D.57}$$

$$\begin{aligned} 0 &= S^2\mathrm{e}^{2t^2}\left(1+2e^{1,1}_{\mu,\lambda}+\frac{1}{\mu}\right) - S^3\mathrm{e}^{\frac{9}{2}t^2}2c_{\mu/\mu,\lambda}\left(1-\frac{1}{N\tau^2}\right) - S^4\mathrm{e}^{8t^2}\frac{1}{\mu N\tau^2} \\ \Rightarrow 0 &= 1+2e^{1,1}_{\mu,\lambda}+\frac{1}{\mu} - S\mathrm{e}^{\frac{5}{2}t^2}c_{\mu/\mu,\lambda}2\left(1-\frac{1}{N\tau^2}\right) - S^2\mathrm{e}^{6t^2}\frac{1}{\mu N\tau^2}, \end{aligned} \tag{D.58}$$

and

$$\begin{aligned} 0 &= S^3\mathrm{e}^{\frac{9}{2}t^2}\left(\frac{1}{\mu}+\left(1+\frac{\tau^2}{\mu}\right)\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)\right) - S^4\mathrm{e}^{8t^2}\left(1+\frac{\tau^2}{\mu}\right)c_{\mu/\mu_I,\lambda}\left(1-\frac{1}{N\tau^2}\right) \\ &\quad -S^5\mathrm{e}^{\frac{25}{2}t^2}\frac{1+\frac{\tau^2}{\mu}}{2\mu N\tau^2} \\ \Rightarrow 0 &= \frac{1}{\mu}+\left(1+\frac{\tau^2}{\mu}\right)\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right) - \left(1+\frac{\tau^2}{\mu}\right)S\mathrm{e}^{\frac{7}{2}t^2}c_{\mu/\mu_I,\lambda}\left(1-\frac{1}{N\tau^2}\right) \\ &\quad -\left(1+\frac{\tau^2}{\mu}\right)S^2\mathrm{e}^{8t^2}\frac{1}{2\mu N\tau^2} \end{aligned} \tag{D.59}$$

with the unknown parameters S and t to be determined. Note that the equations above lead to a nonlinear system the general solution of which cannot be provided analytically. It is possible, though, to obtain numerical solutions using MATHEMATICA (see the discussion in Paragraph 4.4.3 and Fig. 4.15, p. 61).

In some special cases, analytical solutions are obtainable. Before proceeding, Eqs. (D.57) to (D.59) are rewritten in terms of $s^*_\infty := S\,\mathrm{e}^{t^2/2}$

$$0 = \frac{1}{2}+e^{1,1}_{\mu,\lambda} - s^*_\infty\mathrm{e}^{t^2}c_{\mu/\mu,\lambda}\left(1-\frac{1}{N\tau^2}\right) - {s^*_\infty}^2\mathrm{e}^{3t^2}\frac{1}{2\mu N\tau^2}, \tag{D.60}$$

$$0 = 1+2e^{1,1}_{\mu,\lambda}+\frac{1}{\mu} - s^*_\infty\mathrm{e}^{2t^2}c_{\mu/\mu,\lambda}2\left(1-\frac{1}{N\tau^2}\right) - {s^*_\infty}^2\mathrm{e}^{5t^2}\frac{1}{\mu N\tau^2}, \text{and} \tag{D.61}$$

$$\begin{aligned} 0 &= \frac{1}{\mu}+\left(1+\frac{\tau^2}{\mu}\right)\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right) - s^*_\infty\mathrm{e}^{3\tau^2}c_{\mu/\mu,\lambda}\left(1-\frac{1}{N\tau^2}\right)\left(1+\frac{\tau^2}{\mu}\right) \\ &\quad -\frac{1+\frac{\tau^2}{\mu}}{2\mu N\tau^2}{s^*_\infty}^2\mathrm{e}^{7t^2} \end{aligned} \tag{D.62}$$

similarly to [23, p. 319]. As said (D.60) - (D.62) lead to analytical solutions in some cases. These are the limit cases of $N\tau^2 \to \infty$ and $N\tau^2 = 1$. The calculations are given in the following paragraphs of this appendix. A discussion of the results and a comparison with experiments can be found in Section 4.4.3, p. 63.

Limit Case of $N\tau^2 \to \infty$ Let us first consider the limit case of $N\tau^2 \to \infty$. The system Eqs. (D.60) and (D.61), i.e.,

$$\begin{aligned} 0 &= \frac{1}{2} + e^{1,1}_{\mu,\lambda} - s^*_\infty \mathrm{e}^{t^2} c_{\mu/\mu,\lambda}\left(1 - \frac{1}{N\tau^2}\right) - {s^*_\infty}^2 \mathrm{e}^{3t^2} \frac{1}{2\mu N\tau^2} \\ 0 &= 1 + 2e^{1,1}_{\mu,\lambda} + \frac{1}{\mu} - s^*_\infty \mathrm{e}^{2t^2} c_{\mu/\mu,\lambda} 2\left(1 - \frac{1}{N\tau^2}\right) - {s^*_\infty}^2 \mathrm{e}^{5t^2} \frac{1}{\mu N\tau^2} \end{aligned} \tag{D.63}$$

can be easily solved for $N\tau^2 \to \infty$. Taking the limit gives

$$\begin{aligned} 0 &= \frac{1}{2} + e^{1,1}_{\mu,\lambda} - s^*_\infty \mathrm{e}^{t^2} c_{\mu/\mu,\lambda} \\ 0 &= 1 + 2e^{1,1}_{\mu,\lambda} + \frac{1}{\mu} - s^*_\infty \mathrm{e}^{2t^2} 2c_{\mu/\mu,\lambda} \end{aligned} \tag{D.64}$$

leading to two equations describing s^*_∞

$$s^*_\infty = \frac{\frac{1}{2} + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}} \mathrm{e}^{-t^2} \text{ and} \tag{D.65}$$

$$s^*_\infty = \mathrm{e}^{-2t^2} \left(\frac{\frac{1}{2} + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}} + \frac{1}{2\mu c_{\mu/\mu,\lambda}} \right). \tag{D.66}$$

Setting (D.65) equal (D.66) leads to an expression for $\exp(-t^2)$

$$\begin{aligned} \frac{\frac{1}{2} + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}} \mathrm{e}^{-t^2} &= \left(\frac{\frac{1}{2} + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}} + \frac{1}{2\mu c_{\mu/\mu,\lambda}} \right) \mathrm{e}^{-2t^2} \\ \Rightarrow \frac{\frac{1}{2} + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}} \mathrm{e}^{-t^2} &= \left(\frac{2\mu\left(\frac{1}{2} + e^{1,1}_{\mu,\lambda}\right) + 1}{2\mu c_{\mu/\mu,\lambda}} \right) \mathrm{e}^{-2t^2} \\ \Rightarrow \mathrm{e}^{-t^2} &= \left(\frac{\frac{1}{2} + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}} \right) \left(\frac{2\mu c_{\mu/\mu,\lambda}}{2\mu\left(\frac{1}{2} + e^{1,1}_{\mu,\lambda}\right) + 1} \right) \\ \Rightarrow \mathrm{e}^{-t^2} &= \frac{2\mu\left(\frac{1}{2} + e^{1,1}_{\mu,\lambda}\right)}{2\mu\left(\frac{1}{2} + e^{1,1}_{\mu,\lambda}\right) + 1}. \end{aligned} \tag{D.67}$$

Equation (D.67) can be used to obtain the stationary mutation strength for $N\tau^2 \to \infty$ by inserting (D.67) into (D.65) or (D.66)

$$s^*_\infty = \left(\frac{\frac{1}{2} + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}} \right) \left(\frac{2\mu\left(\frac{1}{2} + e^{1,1}_{\mu,\lambda}\right)}{2\mu\left(\frac{1}{2} + e^{1,1}_{\mu,\lambda}\right) + 1} \right). \tag{D.68}$$

The stationary progress rate can be determined in turn by plugging (D.68) into the progress rate (B.24), $\varphi^*(\varsigma^*) = c_{\mu/\mu,\lambda}\varsigma^* - {\varsigma^*}^2/(2\mu)$, leads to

$$\varphi^*_\infty = c_{\mu/\mu,\lambda}\overline{\sigma^*_\infty} - \frac{\overline{(\sigma^*_\infty)^2}}{2\mu}$$

$$
\begin{aligned}
&= c_{\mu/\mu,\lambda} s^*_\infty - \frac{(s^*_\infty)^2 \mathrm{e}^{t^2}}{2\mu} \\
&= c_{\mu/\mu,\lambda}\left(\frac{\frac{1}{2}+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\right)\left(\frac{2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)}{2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)+1}\right) \\
&\quad -\frac{1}{2\mu}\left(\left(\frac{\frac{1}{2}+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\right)\left(\frac{2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)}{2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)+1}\right)\right)^2\left(\frac{2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)+1}{2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)}\right) \\
&= \left(c_{\mu/\mu,\lambda}\left(\frac{\frac{1}{2}+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\right)-\frac{1}{2\mu}\left(\frac{\frac{1}{2}+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\right)^2\right)\left(\frac{2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)}{2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)+1}\right) \\
&= \varphi^*\left(\frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\right)\left(\frac{2\mu(1/2+e^{1,1}_{\mu,\lambda})}{2\mu(1/2+e^{1,1}_{\mu,\lambda})+1}\right). \qquad \text{(D.69)}
\end{aligned}
$$

An Analytical Solution for $N\tau^2 = 1$ As a second special case, $N\tau^2 = 1$ is considered. Again, analytical solutions can be easily obtained. Equations (D.60)-(D.62) describing s^*_∞ change to

$$
\begin{aligned}
0 &= \frac{1}{2} + e^{1,1}_{\mu,\lambda} - {s^*_\infty}^2 \mathrm{e}^{3t^2}\frac{1}{2\mu} && \text{(D.70)}\\
0 &= 1 + 2e^{1,1}_{\mu,\lambda} + \frac{1}{\mu} - {s^*_\infty}^2 \mathrm{e}^{5t^2}\frac{1}{\mu} && \text{(D.71)}\\
0 &= \frac{1}{\mu} + \left(1+\frac{\tau^2}{\mu}\right)\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right) - \frac{1+\frac{\tau^2}{\mu}}{2\mu}{\sigma^*_\infty}^2 \mathrm{e}^{7t^2}. && \text{(D.72)}
\end{aligned}
$$

Only the first two equations are needed to determine s^*_∞. Rewriting Eqs. (D.70) and (D.71) gives

$$
\begin{aligned}
{s^*_\infty}^2 &= 2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)\mathrm{e}^{-3t^2} && \text{(D.73)}\\
{s^*_\infty}^2 &= \left(2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)+1\right)\mathrm{e}^{-5t^2}. && \text{(D.74)}
\end{aligned}
$$

Thus, setting Eq. (D.73) equal to (D.74), we can derive an expression for e^{-2t^2}

$$
\begin{aligned}
2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)\mathrm{e}^{-3t^2} &= \left(2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)+1\right)\mathrm{e}^{-5t^2} \\
\Rightarrow \mathrm{e}^{-2t^2} &= \frac{2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)+1}{2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)}. \qquad \text{(D.75)}
\end{aligned}
$$

Equation (D.75) is then inserted into (D.73) leading to

$$
{s^*_\infty}^2 = 2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)\left(\frac{2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)+1}{2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)}\right)^{\frac{3}{2}}. \qquad \text{(D.76)}
$$

The resulting expected mutation strength

$$s^*_\infty = \left(2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)\right)^{\frac{1}{2}}\left(\frac{2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)+1}{2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)}\right)^{\frac{3}{4}} \tag{D.77}$$

differs from the deterministic result (4.11) by

$$\left(\frac{2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)+1}{2\mu\left(\frac{1}{2}+e^{1,1}_{\mu,\lambda}\right)}\right)^{\frac{3}{4}} \tag{D.78}$$

as can be seen by inserting $N\tau^2 = 1$ into (4.11)

$$\begin{aligned}
\varsigma^*_{stat_2} &= \mu c_{\mu/\mu,\lambda}\left((1-N\tau^2)+\sqrt{(1-N\tau^2)^2+2N\tau^2\frac{1/2+e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}}}\right)\\
&= \mu c_{\mu/\mu,\lambda}\left(\sqrt{2\frac{1/2+e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}}}\right)\\
&= \sqrt{2\mu\left(1/2+e^{1,1}_{\mu,\lambda}\right)}.
\end{aligned} \tag{D.79}$$

D.3.3 A Normal Distribution in the Stationary State

In this subsection the normal distribution $\mathcal{N}(m, s^2)$ is used to model the distribution of the stationary mutation strength. The subsection is devoted to determining the equations to describe the stationary state, to obtain some special analytical solutions and to compare the results with that of the approach using the log-normal distribution. Since a normal distribution with mean m and standard deviation s is used, the raw moments can be obtained easily over

$$\begin{aligned}
\overline{x^k} &= \frac{1}{s\sqrt{2\pi}}\int_{-\infty}^{\infty} x^k \mathrm{e}^{-\frac{1}{2}\left(\frac{x-m}{s}\right)^2}\,\mathrm{d}x\\
&= \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}(st+m)^k\mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t\\
&= \frac{1}{\sqrt{2\pi}}\sum_{l=0}^{k}\binom{k}{l}m^{k-l}s^l\int_{-\infty}^{\infty}t^l\mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t\\
&= \frac{1}{\sqrt{2\pi}}\sum_{l=0}^{k}\binom{k}{l}m^{k-l}s^l\int_{-\infty}^{\infty}t^l\mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t\\
&= \begin{cases} m & \text{for } k=1\\ m^2+s^2 & \text{for } k=2\\ \frac{1}{\sqrt{2\pi}}\left(m^k+\sum_{l=1}^{k}\binom{k}{l}m^{k-l}s^l\int_{-\infty}^{\infty}t^l\mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t\right) & \text{for } k>2\end{cases}.
\end{aligned} \tag{D.80}$$

The integral can be calculated using partial integration. The result is a recursive equation for all even l

$$\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty}t^l\mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t = -[t^{l-1}e^{-\frac{t^2}{2}}]_{-\infty}^{\infty}+\frac{(l-1)}{\sqrt{2\pi}}\int_{-\infty}^{\infty}t^{l-2}\mathrm{e}^{-\frac{t^2}{2}}\,\mathrm{d}t$$

$$
\begin{aligned}
&= \frac{(l-1)}{\sqrt{2\pi}} \int_{-\infty}^{\infty} t^{l-2} \mathrm{e}^{-\frac{t^2}{2}} \, \mathrm{d}t \\
&= \begin{cases} (l-1)(l-3)\ldots \int_{-\infty}^{\infty} \frac{\mathrm{e}^{-\frac{t^2}{2}}}{\sqrt{2\pi}} \, \mathrm{d}t & \text{if } l = 2j \\ (l-1)(l-3)\ldots \int_{-\infty}^{\infty} t \frac{\mathrm{e}^{-\frac{t^2}{2}}}{\sqrt{2\pi}} \, \mathrm{d}t & \text{if } l = 2j+1 \end{cases} \\
&= \begin{cases} (l-1)(l-3)\ldots 1 & \text{if } l = 2j \\ 0 & \text{if } l = 2j+1 \end{cases} . \qquad \text{(D.81)}
\end{aligned}
$$

The first raw moments are therefore

$$
\begin{aligned}
\overline{\sigma_\infty} &= m && \text{(D.82)} \\
\overline{\sigma_\infty^2} &= m^2 + s^2 && \text{(D.83)} \\
\overline{\sigma_\infty^3} &= m^3 + 3ms^2 && \text{(D.84)} \\
\overline{\sigma_\infty^4} &= m^4 + 6m^2 s^2 + 3s^4 && \text{(D.85)} \\
\overline{\sigma_\infty^5} &= m^5 + 10ms^2 + 15s^4. && \text{(D.86)}
\end{aligned}
$$

The starting point to determine the steady state values are Eqs. (4.80)-(4.82), p. 61,

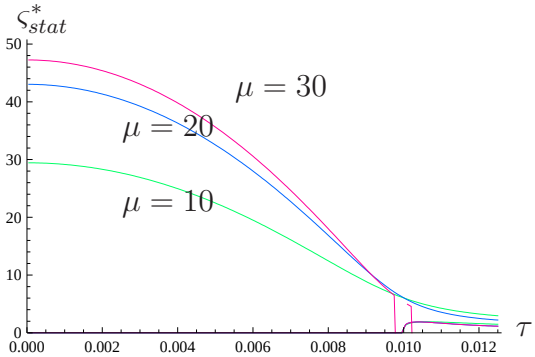
$$
\begin{aligned}
0 &= \overline{\sigma_\infty^*}\Big(1/2 + e_{\mu,\lambda}^{1,1} - \overline{\sigma_\infty^{*\,2}} c_{\mu/\mu,\lambda}\Big(1 - \frac{1}{N\tau^2}\Big) - \frac{\overline{\sigma_\infty^{*\,3}}}{2\mu N\tau^2} && \text{(D.87)} \\
0 &= \overline{\sigma_\infty^{*\,2}}\Big(1 + 2e_{\mu,\lambda}^{1,1} + \frac{1}{\mu}\Big) - \overline{\sigma_\infty^{*\,3}} 2c_{\mu/\mu,\lambda}\Big(1 - \frac{1}{N\tau^2}\Big) - \frac{\overline{\sigma_\infty^{*\,4}}}{\mu N\tau^2} && \text{(D.88)} \\
0 &= \overline{\sigma_\infty^{*\,3}}\left(\frac{1}{\mu} + \Big(1 + \frac{\tau^2}{\mu}\Big)\Big(\frac{1}{2} + e_{\mu,\lambda}^{1,1}\Big)\right) - \Big(1 + \frac{\tau^2}{\mu}\Big) c_{\mu/\mu,\lambda} \overline{\sigma_\infty^{*\,4}}\Big(1 - \frac{1}{N\tau^2}\Big) \\
&\quad - \Big(1 + \frac{\tau^2}{\mu}\Big) \frac{\overline{\sigma_\infty^{*\,5}}}{2\mu N\tau^2}. && \text{(D.89)}
\end{aligned}
$$

Using the normal distribution $\sigma_\infty \sim \mathcal{N}(m, s^2)$, a system of nonlinear equations in m and s is obtained (cf. (D.82)-(D.86))

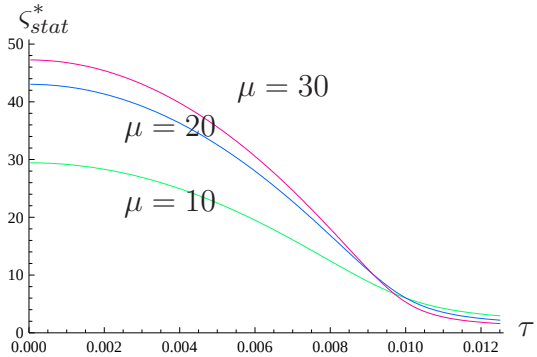
$$
\begin{aligned}
0 &= m\Big(1/2 + e_{\mu,\lambda}^{1,1}\Big) - (s^2 + m^2) c_{\mu/\mu,\lambda}\Big(1 - \frac{1}{N\tau^2}\Big) - \frac{m(m^2 + 3s^2)}{2\mu N\tau^2} && \text{(D.90)} \\
0 &= (s^2 + m^2)\Big(1 + 2e_{\mu,\lambda}^{1,1} + \frac{1}{\mu}\Big) - m(3s^2 + m^2) 2c_{\mu/\mu,\lambda}\Big(1 - \frac{1}{N\tau^2}\Big) \\
&\quad - \frac{3s^4 + 6s^2 m^2 + m^4}{\mu N\tau^2} && \text{(D.91)} \\
0 &= m(m^2 + 3s^2)\left(\frac{1}{\mu} + \Big(1 + \frac{\tau^2}{\mu}\Big)\Big(\frac{1}{2} + e_{\mu,\lambda}^{1,1}\Big)\right) \\
&\quad - \Big(1 + \frac{\tau^2}{\mu}\Big) c_{\mu/\mu,\lambda}(3s^4 + 6s^2 m^2 + m^4)\Big(1 - \frac{1}{N\tau^2}\Big) \\
&\quad - \Big(1 + \frac{\tau^2}{\mu}\Big) \frac{m(15s^4 + 10s^2 m + m^5)}{2\mu N\tau^2}. && \text{(D.92)}
\end{aligned}
$$

Again, analytical results can only be derived for some special cases which is done in the following paragraphs for the sake of completeness. First of all: Equations (D.90) and (D.91) allow up to two

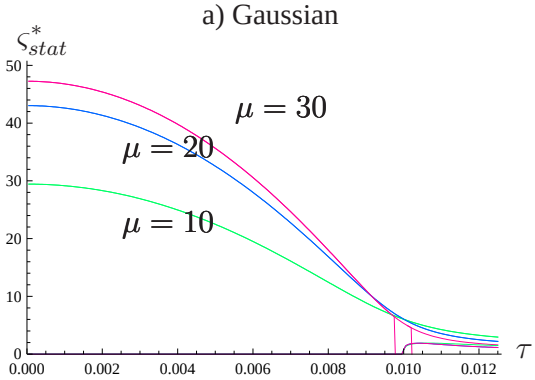
solutions for the stationary mutation strength, expecially if τ is relatively high, i.e., $N\tau^2 \geq 1$. One solution is small, the other nearly coincides with the solution obtained using the log-normal distribution (and is more in accordance with the results of experiments). Figure D.2 shows the results for some $(\mu/\mu_I, 60)$-ES ($N = 10,000$). As one can see the curves for the greater solution of (D.90) and (D.91) and the one obtained using the log-normal ansatz and Eqs. (D.62) - (D.62) cannot be distinguished. If the learning rate increases, the numerical determination of the mutation strength seems to be problematic in some cases as seen in Fig. D.2. In the following, some special analytical solutions



a) Gaussian



b) Log-Normal



c) Comparison of Gaussian and Log-Normal Distribution

Figure D.2: Stationary normalized mutation strength and progress rate as a function of τ for some $(\mu/\mu_I, 60)$-ES and $N = 10,000$.

are provided.

The special case of $N\tau^2 \to \infty$ In this paragraph, the stationary mutation strength is derived $N\tau^2 \to \infty$.

$$0 = m\left(1/2 + e^{1,1}_{\mu,\lambda}\right) - (s^2 + m^2)c_{\mu/\mu,\lambda} \tag{D.93}$$

$$0 = (s^2 + m^2)\left(1 + 2e^{1,1}_{\mu,\lambda} + \frac{1}{\mu}\right) - m(3s^2 + m^2)2c_{\mu/\mu,\lambda} \tag{D.94}$$

$$0 = m(m^2 + 3s^2)\left(\frac{1}{\mu} + \left(1 + \frac{\tau^2}{\mu}\right)\left(\frac{1}{2} + e^{1,1}_{\mu,\lambda}\right)\right) \tag{D.95}$$

Equation (D.93) leads to

$$m\left(\frac{1/2 + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\right) = s^2 + m^2 \tag{D.96}$$

which, following insertion into (D.94) and solving for $m = \overline{\sigma_\epsilon}$, gives

$$\begin{aligned}
(s^2+m^2)\Big(\Big(\frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\Big)+\frac{1}{2\mu c_{\mu/\mu,\lambda}}\Big) &= m\Big(3s^2+m^2\Big)\\
\Rightarrow m\Big(\frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\Big)\Big(\Big(\frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\Big)+\frac{1}{2\mu c_{\mu/\mu,\lambda}}\Big) &= m\Big(3s^2+m^2\Big)\\
\Leftrightarrow \Big(\frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\Big)\Big(\Big(\frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\Big)+\frac{1}{2\mu c_{\mu/\mu,\lambda}}\Big) &= 3m\Big(\frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\Big)-3m^2+m^2\\
\Rightarrow \Big(\frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\Big)^2+\frac{1/2+e^{1,1}_{\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda}} &= 3m\Big(\frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\Big)-2m^2. \qquad \text{(D.97)}
\end{aligned}$$

Equation (D.97) leads to two positive solutions

$$\begin{aligned}
m_{1,2} &= \frac{3}{4}\Big(\frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\Big)\\
&\quad \pm\sqrt{\frac{9}{16}\Big(\frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\Big)^2-\frac{1}{2}\Big(\frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\Big)^2-\frac{1}{2}\Big(\frac{1/2+e^{1,1}_{\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda}}\Big)}\\
\Rightarrow m_{1,2} &= \Big(\frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\Big)\\
&\quad \Big(\frac{3}{4}\pm\sqrt{\frac{9}{16}-\frac{1}{2}-\frac{1}{4\mu(1/2+e^{1,1}_{\mu,\lambda})}}\Big)\\
\Rightarrow m_{1,2} &= \Big(\frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\Big)\\
&\quad \Big(\frac{3}{4}\pm\frac{1}{4}\sqrt{\frac{2\mu(1/2+e^{1,1,}_{\mu,\lambda})-8}{2\mu(1/2+e^{1,1}_{\mu,\lambda})}}\Big) \qquad \text{(D.98)}
\end{aligned}$$

which are defined for $\mu > 4/(1/2+e^{1,1}_{\mu,\lambda})$ provided that $1/2+e^{1,1}_{\mu,\lambda} > 0$. Again, the solution of the deterministic result (4.11) for $N\tau^2 \to \infty$

$$\lim_{N\tau^2\to\infty}\varsigma^*_{stat} = \frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}} \qquad \text{(D.99)}$$

reappears with a correction factor.

The special case of $N\tau^2 = 1$ In this section, the stationary mutation strength for $N\tau^2 = 1$ is determined. Setting $N\tau^2 = 1$ changes Eqs. (D.90) and (D.91) to

$$0 = m(1/2+e^{1,1}_{\mu,\lambda}) - m\frac{m^2+3s^2}{2\mu} \qquad \text{(D.100)}$$

$$0 = (s^2+m^2)\Big(1+2e^{1,1}_{\mu,\lambda}+\frac{1}{\mu}\Big)-\frac{1}{\mu}\Big(3s^4+6s^2m^2+m^4\Big). \qquad \text{(D.101)}$$

Solving (D.100) for s^2 leads to

$$s^2 = \frac{1}{3}\Big(2\mu(1/2 + e^{1,1}_{\mu,\lambda}) - m^2\Big) \tag{D.102}$$

$$m^4 = \frac{1}{9}\bigg(\Big((2\mu(1/2 + e^{1,1}_{\mu,\lambda})\Big)^2 - 2m^2 2\mu(1/2 + e^{1,1}_{\mu,\lambda}) + m^4\bigg). \tag{D.103}$$

Plugging (D.102) and (D.103) into (D.101) leads to the replacement of the terms by

$$\begin{aligned}
s^2 + m^2 &= \frac{1}{3}\Big(2\mu(1/2 + e^{1,1}_{\mu,\lambda}) + 2m^2\Big) \\
(s^2 + m^2)\Big(2\mu(1/2 + e^{1,1}_{\mu,\lambda}) + 1\Big) &= \frac{1}{3}\Big(2\mu(1/2 + e^{1,1}_{\mu,\lambda}) + 2m^2\Big)\Big(2\mu(1/2 + e^{1,1}_{\mu,\lambda}) + 1\Big) \\
&= \frac{1}{3}\Big(2\mu(1/2 + e^{1,1}_{\mu,\lambda})^2 + 2\mu(1/2 + e^{1,1}_{\mu,\lambda}) \\
&\quad +2m^2 2\mu(1/2 + e^{1,1}_{\mu,\lambda}) + 2m^2\Big) \\
3s^4 &= \frac{1}{3}\bigg(\Big((2\mu(1/2 + e^{1,1}_{\mu,\lambda})\Big)^2 - 2m^2 2\mu(1/2 + e^{1,1}_{\mu,\lambda}) + m^4\bigg) \\
6m^2 s^2 &= 2m^2 2\mu(1/2 + e^{1,1}_{\mu,\lambda}) - 2m^4 \\
3s^4 + 6m^2 s^2 + m^4 &= \frac{1}{3}\bigg(\Big(2\mu(1/2 + e^{1,1}_{\mu,\lambda})\Big)^2 - 2m^2 2\mu(1/2 + e^{1,1}_{\mu,\lambda}) + m^4\bigg) \\
&\quad 2m^2 2\mu(1/2 + e^{1,1}_{\mu,\lambda}) - 2m^4 + m^4 \\
&= \frac{1}{3}\Big(2\mu(1/2 + e^{1,1}_{\mu,\lambda})\Big)^2 + \frac{4}{3}m^2 2\mu(1/2 + e^{1,1}_{\mu,\lambda}) \\
&\quad -\frac{2}{3}m^4
\end{aligned} \tag{D.104}$$

and therefore to

$$\begin{aligned}
3s^4 + 6s^2 2m^2 + m^4 &= (s^2 + m^2)\Big(2\mu(1/2 + e^{1,1}_{\mu,\lambda}) + 1\Big) \\
0 &= m^4 - m^2(2\mu(1/2 + e^{1,1}_{\mu,\lambda}) - 1) + \frac{1}{2}2\mu(1/2 + e^{1,1}_{\mu,\lambda})
\end{aligned} \tag{D.105}$$

Solving (D.105) for m^2 gives

$$\begin{aligned}
m^2_{1,2} &= \frac{1}{2}\Big(2\mu(1/2 + e^{1,1}_{\mu,\lambda}) - 1\Big) \\
&\quad \pm\sqrt{\frac{1}{4}\Big(2\mu(1/2 + e^{1,1}_{\mu,\lambda}) - 1\Big)^2 - \frac{1}{2}\Big(2\mu(1/2 + e^{1,1}_{\mu,\lambda})\Big)} \\
&= \frac{1}{2}\Big(2\mu(1/2 + e^{1,1}_{\mu,\lambda}) - 1\Big) \\
&\quad \pm\sqrt{\frac{1}{4}\Big(2\mu(1/2 + e^{1,1}_{\mu,\lambda})^2 - 4\mu(1/2 + e^{1,1}_{\mu,\lambda}) + 1\Big) - \frac{1}{2}\Big(2\mu(1/2 + e^{1,1}_{\mu,\lambda})\Big)} \\
&= \frac{1}{2}\Big(2\mu(1/2 + e^{1,1}_{\mu,\lambda}) - 1\Big) \\
&\quad \pm\sqrt{\frac{1}{4}\Big(2\mu(1/2 + e^{1,1}_{\mu,\lambda})^2\Big) - 2\mu(1/2 + e^{1,1}_{\mu,\lambda}) + \frac{1}{4}}
\end{aligned}$$

$$= \frac{1}{2}\Big(2\mu(1/2 + e^{1,1}_{\mu,\lambda}) - 1 \pm\sqrt{\Big(2\mu(1/2 + e^{1,1}_{\mu,\lambda})^2\Big) - 4(2\mu(1/2 + e^{1,1}_{\mu,\lambda})) + 1}\Big). \quad \text{(D.106)}$$

Again, two positive solutions are obtained for the root of (D.106). As stated, the larger stationary mutation strength is more in accordance with the results of experiments. As before, the result of the deterministic approach $2\mu(1/2 + e^{1,1}_{\mu,\lambda})$ appears coupled with a correction term.

E Ridge Functions: Derivation of the Main Results

In this section, the derivations of main equations in Chapter 5 are presented. In Subsection E.1.1, the undisturbed sharp ridge is considered. In particular, it is shown how the stationary mutation strength can be derived. Afterwards in Subsection E.1.2, the derivation of the stationary state values is given in the case of the parabolic ridge. The next section addresses noisy ridge functions. Subsection E.2.1 is devoted to the noisy sharp ridge. First, the stationary points are derived. This concluded, the non-normalized stationary values of the mutation strength and progress rate are determined. Subsection E.2.2 is devoted to the noise parabolic ridge. The main point is the determination of the stationary distance to the ridge.

E.1 The Noise Free Case

As stated before, this section is devoted to the derivation of the main results of Section 5.1 starting with the sharp ridge before presenting the calculations in the case of the parabolic ridge.

E.1.1 The Sharp Ridge: The Stationary Normalized Mutation Strength

In this section, the derivation of the stationary state of the evolution equation (5.23)

$$\begin{aligned}\varsigma^*_{st} &= \sqrt{\frac{d^2}{1+d^2}}\mu c_{\mu/\mu,\lambda}\Big((1-N\tau^2) \\ &\quad +\sqrt{(1+N\tau^2)^2+2N\tau^2(1+d^2)\frac{1/2+e^{1,1}_{\mu,\lambda}}{d^2\mu c^2_{\mu/\mu,\lambda}}}\Big)\end{aligned} \tag{E.1}$$

is given. The evolution equation (5.46)

$$\varsigma^* = \sigma^*\left(\frac{1+\tau^2(1/2+e^{1,1}_{\mu,\lambda}-\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}\sigma^*)}{1-\frac{1}{N}(\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}\sigma^*-\frac{\sigma^{*2}}{2\mu})}\right) \tag{E.2}$$

serves as the starting point of the derivation. After having derived the stationary mutation strength, it is shown that this stationary solution is stable with respect to the linear approximation.

Deriving the Stationary Mutation Strength

Demanding stationarity of the ς^*-evolution, (5.46), the mutation strength must either be zero $\sigma^* = 0$ or

$$1 = \frac{1+\tau^2\Big(1/2+e^{1,1}_{\mu,\lambda}-\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}\sigma^*)}{1-\frac{1}{N}(\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}\sigma^*-\frac{\sigma^{*2}}{2\mu})} \tag{E.3}$$

has to hold. The latter condition leads to

$$\begin{aligned}
1-\frac{1}{N}(\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}\sigma^* - \frac{\sigma^{*2}}{2\mu}) &= 1+\tau^2\Big(1/2+e^{1,1}_{\mu,\lambda}-\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}\sigma^*\Big)\\
\Rightarrow \frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}\sigma^* - \frac{\sigma^{*2}}{2\mu} &= -N\tau^2\Big(\frac{1}{2}+e^{1,1}_{\mu,\lambda}-\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}\sigma^*\Big)\\
\Leftrightarrow (1-N\tau^2)\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}\sigma^* - \frac{\sigma^{*2}}{2\mu} &= -N\tau^2(\frac{1}{2}+e^{1,1}_{\mu,\lambda})\\
\Rightarrow \sigma^{*2}-(1-N\tau^2)\frac{d}{\sqrt{1+d^2}}2\mu c_{\mu/\mu,\lambda}\sigma^* &= N\tau^2 2\mu(\frac{1}{2}+e^{1,1}_{\mu,\lambda}) \qquad \text{(E.4)}
\end{aligned}$$

which has two solutions

$$\begin{aligned}
\varsigma^*_{1,2} &= (1-N\tau^2)\frac{d}{\sqrt{1+d^2}}\mu c_{\mu/\mu,\lambda}\\
&\quad \pm\sqrt{\mu^2 c^2_{\mu/\mu,\lambda}\frac{d^2}{1+d^2}(1-N\tau^2)^2+N\tau^2 2\mu(1/2+e^{1,1}_{\mu,\lambda})}. \qquad \text{(E.5)}
\end{aligned}$$

As can be inferred from (E.5), the positive solution and the stationary mutation strength is given by (5.23)

$$\begin{aligned}
\varsigma^*_{st} &= \sqrt{\frac{d^2}{1+d^2}}\mu c_{\mu/\mu,\lambda}\Bigg((1-N\tau^2)\\
&\quad +\sqrt{(1-N\tau^2)^2+2N\tau^2\Big(\frac{1+d^2}{d^2}\Big)\Big(\frac{\frac{1}{2}+e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}}\Big)}\Bigg). \qquad \text{(E.6)}
\end{aligned}$$

The stationary quality change (5.24) can be derived by inserting (E.6) into (5.17)

$$\overline{\Delta Q^*} = \sqrt{1+d^2}c_{\mu/\mu,\lambda}\varsigma^* - \frac{d}{2\mu}\varsigma^{*2}. \qquad \text{(E.7)}$$

This leads to

$$\begin{aligned}
\overline{\Delta Q^*_{st}} &= \sqrt{1+d^2}c_{\mu/\mu,\lambda}\sqrt{\frac{d^2}{1+d^2}}\mu c_{\mu/\mu,\lambda}\Bigg((1-N\tau^2)\\
&\quad +\sqrt{(1+N\tau^2)^2+2N\tau^2\Big(\frac{1+d^2}{d^2}\Big)\Big(\frac{\frac{1}{2}+e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}}\Big)}\Bigg)\\
&\quad -\frac{d}{2\mu}\mu^2 c^2_{\mu/\mu,\lambda}\Big(\frac{d^2}{1+d^2}\Big)\Bigg((1-N\tau^2)\\
&\quad +\sqrt{(1+N\tau^2)^2+2N\tau^2\Big(\frac{1+d^2}{d^2}\Big)\Big(\frac{\frac{1}{2}+e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}}\Big)}\Bigg)^2\\
&= d\mu c^2_{\mu/\mu,\lambda}\Bigg((1-N\tau^2)
\end{aligned}$$

$$
\begin{aligned}
&+\sqrt{(1+N\tau^2)^2+2N\tau^2\Big(\frac{1+d^2}{d^2}\Big)\Big(\frac{\frac{1}{2}+e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}}\Big)}\,\Bigg)\\
&\times\Bigg(1-\frac{d^2}{2(1+d^2)}\Bigg((1-N\tau^2)\\
&+\sqrt{(1+N\tau^2)^2+2N\tau^2\Big(\frac{1+d^2}{d^2}\Big)\Big(\frac{\frac{1}{2}+e^{1,1}_{\mu,\lambda}}{\mu c^2_{\mu/\mu,\lambda}}\Big)}\,\Bigg)\Bigg).
\end{aligned}
\tag{E.8}
$$

Stability of the Stationary Mutation Strength

The stability of the stationary mutation strength (E.6) remains to be shown. To this end, the linear approximation is used again. Therefore, the first derivative of

$$
f(s) \;=\; s\Big(\frac{1+\psi(s)}{1-\frac{\varphi_R(s)}{N}}\Big) = s\Big(\frac{1+\tau^2(1/2+e^{1,1}_{\mu,\lambda}-\frac{d}{\sqrt{1+d^2}}s)}{1-\frac{1}{N}(\frac{d}{\sqrt{1+d^2}}s-\frac{s^2}{2\mu})}\Big) \tag{E.9}
$$

which appears in the evolution equation (E.2) needs to be determined. Since the derivative at $s=\sigma^*_{st}$ is required, the calculations simplify. More specifically, the derivative given by

$$
f'(s) \;=\; \frac{1+\psi(s)}{1-\frac{\varphi_R(s)}{N}}+\frac{s}{1-\frac{\varphi_R(s)}{N}}\Bigg(\psi'(s)+\frac{\varphi'_R(s)}{N}\Big(\frac{1+\psi(s)}{1-\frac{\varphi_R(s)}{N}}\Big)\Bigg) \tag{E.10}
$$

changes to

$$
f'(s)|_{s=\sigma^*_{st}} \;=\; 1+\frac{\sigma^*_{st}}{1-\frac{\varphi_R(\sigma^*_{st})}{N}}\Bigg(\psi'(\sigma^*_{st})+\frac{\varphi'_R(\sigma^*_{st})}{N}\Bigg) \tag{E.11}
$$

since the ES is in the stationary state. It remains to show that $|f'(\sigma^*_{st})|<1$ holds. Let us start with $f'(\sigma^*_{st})>-1$. It has to be shown that

$$
1+\frac{\sigma^*_{st}}{1-\frac{1}{N}\Big(\frac{dc_{\mu/\mu,\lambda}}{\sqrt{1+d^2}}\sigma^*_{st}-\frac{{\sigma^*_{st}}^2}{2\mu}\Big)}\Bigg(-\frac{d\tau^2c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}}+\frac{dc_{\mu/\mu,\lambda}}{N\sqrt{1+d^2}}-\frac{\sigma^*_{st}}{N\mu}\Bigg)>-1 \tag{E.12}
$$

holds. Inequality (E.12) can be simplified to

$$
\begin{aligned}
1 \;+\; &\frac{\sigma^*_{st}}{1-\frac{1}{N}\Big(\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}\sigma^*_{st}-\frac{{\sigma^*_{st}}^2}{2\mu}\Big)}\Bigg(\Big(\frac{1}{N}-\tau^2\Big)\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}-\frac{\sigma^*_{st}}{N\mu}\Bigg)>-1\\
&\Rightarrow \frac{\sigma^*_{st}}{1-\frac{1}{N}\Big(\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}\sigma^*_{st}-\frac{{\sigma^*_{st}}^2}{2\mu}\Big)}\Bigg(\Big(\frac{1}{N}-\tau^2\Big)\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}-\frac{\sigma^*_{st}}{N\mu}\Bigg)>-2\\
&\Longleftrightarrow \Big(\frac{1}{N}-\tau^2\Big)\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}\sigma^*_{st}-\frac{{\sigma^*_{st}}^2}{N\mu}>-2+\frac{2}{N}\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}\sigma^*_{st}-\frac{{\sigma^*_{st}}^2}{N\mu}\\
&\Longleftrightarrow \Big(\frac{1}{N}-\tau^2\Big)\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}\sigma^*_{st}>-2+\frac{2}{N}\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}\sigma^*_{st}\\
&\Longleftrightarrow -\Big(\frac{1}{N}+\tau^2\Big)\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}\sigma^*_{st}>-2\\
&\Rightarrow \Big(\frac{1}{N}+\tau^2\Big)\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}\sigma^*_{st}<2.
\end{aligned}
\tag{E.13}
$$

Inequality (E.13) requires inserting the normalized mutation strength (E.6) into (E.13) and determining whether (E.13) holds or not. In the following, a different approach is followed. Instead of inserting (E.6), the highest stationary mutation strength that may occur is considered. The conditions derived in this manner are therefore sufficient but not necessary. The highest mutation strength is the zero of the SAR, $\varsigma^*_{\psi_0} = (1/2 + e^{1,1}_{\mu,\lambda})/c_{\mu/\mu,\lambda}\sqrt{1+d^2}/d$ which is obtained for $N\tau^2 \to \infty$. If (E.13) holds for $\varsigma^*_{\psi_0}$, it holds in general. Inserting $\varsigma^*_{\psi_0}$ into (E.13) leads to a sufficient condition for $f'(\sigma^*_{st}) > -1$

$$\begin{aligned}\Big(\frac{1}{N}+\tau^2\Big)\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}\Big(\frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\Big)\frac{\sqrt{1+d^2}}{d} &< 2\\ \Rightarrow \Big(\frac{1}{N}+\tau^2\Big)\Big(1/2+e^{1,1}_{\mu,\lambda}\Big) &< 2\\ \Rightarrow \tau^2 &< \frac{2}{1/2+e^{1,1}_{\mu,\lambda}}-\frac{1}{N}. \end{aligned} \tag{E.14}$$

Provided that the learning rate τ is sufficiently small with respect to the choices of μ and λ, $f'(\sigma^*_{st}) > -1$ can be ensured. The case of $f'(\sigma^*_{st}) < 1$ remains to be shown. Consider

$$1+\frac{\sigma^*_{st}}{1-\frac{1}{N}\Big(\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}\sigma^*_{st}-\frac{{\sigma^*_{st}}^2}{2\mu}\Big)}\Big(-\frac{\tau^2 d c_{\mu/\mu,\lambda}}{\sqrt{1+d^2}}+\frac{d c_{\mu/\mu,\lambda}}{N\sqrt{1+d^2}}-\frac{\sigma^*_{st}}{N\mu}\Big)<1. \tag{E.15}$$

Since $d < d_{crit}$, the progress rate for the stationary mutation strength is negative. Inequality (E.15) gives

$$\begin{aligned}\Big(\frac{1}{N}-\tau^2\Big)\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda}-\frac{\sigma^*_{st}}{N\mu} &< 0\\ \Rightarrow \mu\Big(1-N\tau^2\Big)\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda} &< \sigma^*_{st}. \end{aligned} \tag{E.16}$$

If $N\tau^2 \geq 1$, nothing remains to be shown. Otherwise, for $N\tau^2 < 1$, a similar approach as before is followed. This time it is shown that (E.16) is valid for the smallest stationary mutation strength. The smallest stationary mutation strength is the zero of the progress rate $\varsigma^*_{\varphi_{R_0}} = 2\mu c_{\mu/\mu,\lambda} d/\sqrt{1+d^2}$. Inserting the zero into (E.16) leads to

$$\begin{aligned}\mu\Big(1-N\tau^2\Big)\frac{d}{\sqrt{1+d^2}}c_{\mu/\mu,\lambda} &< 2\mu c_{\mu/\mu,\lambda}\frac{d}{\sqrt{1+d^2}}\\ \Rightarrow 1-N\tau^2 &< 2 \end{aligned} \tag{E.17}$$

which is generally fulfilled. In this section, a sufficient condition for the stability of the stationary mutation strength could be derived. Provided that the learning rate is sufficiently small, the stationary mutation strength (E.6) is stable with respect to the linear approximation.

E.1.2 The Parabolic Ridge: The Stationary State

In this section, the calculations leading to the stationary points (5.37), p. 89

$$\begin{pmatrix} R_{st} \\ \varsigma^\star_{st} \end{pmatrix} = \begin{pmatrix} \frac{1}{2d}\sqrt{\frac{1/2+e^{1,1}_{\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda}-1/2-e^{1,1}_{\mu,\lambda}}} \\ \frac{\sqrt{2\mu}}{2d}\frac{1/2+e^{1,1}_{\mu,\lambda}}{\sqrt{2\mu c^2_{\mu/\mu,\lambda}-1/2-e^{1,1}_{\mu,\lambda}}} \end{pmatrix} \tag{E.18}$$

and (5.38), p. 90

$$\begin{pmatrix} R_{st} \\ \varsigma^{\star}_{st} \end{pmatrix} = \begin{pmatrix} \left(\frac{1/2+e^{1,1}_{\mu,\lambda}}{\alpha^2 d^2 \left(2\mu c^2_{\mu/\mu,\lambda} - 1/2 - e^{1,1}_{\mu,\lambda}\right)} \right)^{1/(2\alpha-2)} \\ \sqrt{2\mu(1/2+e^{1,1}_{\mu,\lambda})} \left(\frac{1/2+e^{1,1}_{\mu,\lambda}}{\alpha^2 d^2 \left(2\mu c^2_{\mu/\mu,\lambda} - 1/2 - e^{1,1}_{\mu,\lambda}\right)} \right)^{1/(2\alpha-2)} \end{pmatrix} \tag{E.19}$$

are given. Furthermore, some numerical evidence is provided for the claim that the stationary state (E.18) is stable.

Determining the Stationary State

In this subsection, the stationary states are determined. To this end, the evolution equations (5.26)

$$\begin{aligned} r &= R - \frac{1}{N}\varphi^{\star}_R(\sigma^{\star}, R) \\ \varsigma^{\star} &= \sigma^{\star}\Big(1 + \psi(\sigma^{\star}, R)\Big) \end{aligned} \tag{E.20}$$

are needed. Stationary solutions of (E.20) require either a zero mutation strength of $\varphi^{\star}_R(\sigma^{\star}, R) = 0$ in the case of the R-evolution and $\psi(\sigma^{\star}, R) = 0$ in the case of the $\sigma^{\star}$-evolution. The progress rate (5.27)

$$\varphi^{\star}_R(\sigma^*, R) = \frac{d\alpha R^{\alpha-1} c_{\mu/\mu,\lambda}}{\sqrt{1 + d^2\alpha^2 R^{2\alpha-2}}}\sigma^* - \frac{\sigma^{\star 2}}{2R\mu} \tag{E.21}$$

leads to two zeros $\varsigma^{\star}_{\varphi R0_1} = 0$ and (5.33)

$$\varsigma^{\star}_{\varphi R0} = 2R\mu c_{\mu/\mu,\lambda}\sqrt{\frac{\alpha^2 d^2 R^{2\alpha-2}}{1 + \alpha^2 d^2 R^{2\alpha-2}}}. \tag{E.22}$$

The zero of SAR (5.29)

$$\psi(\sigma^{\star}) = \tau^2\left(\frac{1}{2} + e^{1,1}_{\mu,\lambda} - \frac{c_{\mu/\mu,\lambda}}{R}\sqrt{\frac{d^2\alpha^2 R^{2\alpha-2}}{1 + d^2\alpha^2 R^{2\alpha-2}}}\,\sigma^{\star}\right) \tag{E.23}$$

is given by (5.30)

$$\varsigma^{\star}_{\psi 0} = R\frac{1/2 + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\sqrt{\frac{1 + \alpha^2 d^2 R^{2\alpha-2}}{\alpha^2 d^2 R^{2\alpha-2}}}. \tag{E.24}$$

If stationarity of both evolution equations is demanded, either the mutation strength must be zero or $\varsigma^{\star}_{\psi 0} = \varsigma^{\star}_{\varphi R0}$ has to hold, i.e.,

$$2R\mu c_{\mu/\mu,\lambda}\sqrt{\frac{\alpha^2 d^2 R^{2\alpha-2}}{1 + \alpha^2 d^2 R^{2\alpha-2}}} = R\frac{1/2 + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\sqrt{\frac{1 + \alpha^2 d^2 R^{2\alpha-2}}{\alpha^2 d^2 R^{2\alpha-2}}} \tag{E.25}$$

(cf. (E.22) and (E.23)). Solving (E.25) for R, a stationary distance to the axis

$$2R\mu c_{\mu/\mu,\lambda}\sqrt{\frac{\alpha^2 d^2 R^{2\alpha-2}}{1 + \alpha^2 d^2 R^{2\alpha-2}}} = R\frac{1/2 + e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\sqrt{\frac{1 + \alpha^2 d^2 R^{2\alpha-2}}{\alpha^2 d^2 R^{2\alpha-2}}}$$

$$
\begin{aligned}
\Rightarrow 2\mu c_{\mu/\mu,\lambda}\alpha^2 d^2 R^{2\alpha-2} &= \frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}(1+\alpha^2 d^2 R^{2\alpha-2}) \\
\Leftrightarrow \Big(2\mu c_{\mu/\mu,\lambda} - \frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\Big)\alpha^2 d^2 R^{2\alpha-2} &= \frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}} \\
\Leftrightarrow \alpha^2 d^2 R^{2\alpha-2} &= \frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}\Big(2\mu c_{\mu/\mu,\lambda} - \frac{1/2+e^{1,1}_{\mu,\lambda}}{c_{\mu/\mu,\lambda}}\Big)} \\
\Leftrightarrow \alpha^2 d^2 R^{2\alpha-2} &= \frac{1/2+e^{1,1}_{\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda} - 1/2 - e^{1,1}_{\mu,\lambda}} \\
\Rightarrow R_{st,\alpha} &= \sqrt[2\alpha-2]{\frac{1}{\alpha^2 d^2}\left(\frac{1/2+e^{1,1}_{\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda} - 1/2 - e^{1,1}_{\mu,\lambda}}\right)}
\end{aligned}
\tag{E.26}
$$

is obtained for general $\alpha \geq 2$ and

$$
R_{st} = \frac{1}{2d}\sqrt{\frac{1/2+e^{1,1}_{\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda} - 1/2 - e^{1,1}_{\mu,\lambda}}} \tag{E.27}
$$

for $\alpha = 2$. The stationary distance can be used to determine the stationary mutation strength in (5.38) by plugging (E.26) into (E.22) or (E.23). In the following, (E.22) is used. Let us first consider

$$
\begin{aligned}
\frac{\alpha^2 d^2 R^{2\alpha-2}_{st}}{1+\alpha^2 d^2 R^{2\alpha-2}_{st}} &= \frac{1/2+e^{1,1}_{\mu,\lambda}}{\Big(2\mu c^2_{\mu/\mu,\lambda} - 1/2 - e^{1,1}_{\mu,\lambda}\Big)\Big(1+\frac{1/2+e^{1,1}_{\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda}-1/2-e^{1,1}_{\mu,\lambda}}\Big)} \\
&= \frac{1/2+e^{1,1}_{\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda}}.
\end{aligned}
\tag{E.28}
$$

Plugging (E.28) and (E.26) into (E.22) leads to

$$
\begin{aligned}
\varsigma^\star_{st} &= 2R_{st}\mu c_{\mu/\mu,\lambda}\sqrt{\frac{\alpha^2 d^2 R^{2\alpha-2}_{st}}{1+\alpha^2 d^2 R^{2\alpha-2}_{st}}} \\
&= R_{st} 2\mu c_{\mu/\mu,\lambda}\sqrt{\frac{1/2+e^{1,1}_{\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda}}} \\
&= R_{st}\sqrt{2\mu(1/2+e^{1,1}_{\mu,\lambda})} \\
&= \sqrt[2\alpha-2]{\frac{1}{\alpha^2 d^2}\left(\frac{1/2+e^{1,1}_{\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda} - 1/2 - e^{1,1}_{\mu,\lambda}}\right)}\sqrt{2\mu(1/2+e^{1,1}_{\mu,\lambda})}.
\end{aligned}
\tag{E.29}
$$

Thus, the stationary mutation strength in (5.38) is obtained. Setting $\alpha = 2$ gives the mutation strength in (5.37)

$$
\varsigma^\star_{st} = \frac{\sqrt{2\mu}}{2d}\frac{1/2+e^{1,1}_{\mu,\lambda}}{\sqrt{2\mu c^2_{\mu/\mu,\lambda} - 1/2 - e^{1,1}_{\mu,\lambda}}}. \tag{E.30}
$$

Now the stationary progress rate $\varphi^{\star}_{x\,st}$ can be derived. Plugging (5.38) into (5.28)

$$\varphi^{\star}_{x}(\sigma^{\star},R) \;=\; \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+d^2\alpha^2R^{2\alpha-2}}}\sigma^{\star} \tag{E.31}$$

leads to

$$\begin{aligned}
\varphi^{\star}_{x\,st} &= \frac{c_{\mu/\mu,\lambda}}{\sqrt{1+\frac{1/2+e^{1,1}_{\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda}-1/2-e^{1,1}_{\mu,\lambda}}}}\sqrt{2\mu(1/2+e^{1,1}_{\mu,\lambda})} \\
&\quad\times\sqrt[2\alpha-2]{\frac{1}{\alpha^2d^2}\left(\frac{1/2+e^{1,1}_{\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda}-1/2-e^{1,1}_{\mu,\lambda}}\right)} \\
&= \frac{c_{\mu/\mu,\lambda}}{\sqrt{\frac{2\mu c^2_{\mu/\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda}-1/2-e^{1,1}_{\mu,\lambda}}}}\sqrt{2\mu(1/2+e^{1,1}_{\mu,\lambda})} \\
&\quad\times\sqrt[2\alpha-2]{\frac{1}{\alpha^2d^2}\left(\frac{1/2+e^{1,1}_{\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda}-1/2-e^{1,1}_{\mu,\lambda}}\right)} \\
&= \sqrt{\frac{2\mu c^2_{\mu/\mu,\lambda}-1/2-e^{1,1}_{\mu,\lambda}}{2\mu}}\sqrt{2\mu(1/2+e^{1,1}_{\mu,\lambda})} \\
&\quad\times\sqrt[2\alpha-2]{\frac{1}{\alpha^2d^2}\left(\frac{1/2+e^{1,1}_{\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda}-1/2-e^{1,1}_{\mu,\lambda}}\right)} \\
&= \sqrt{2\mu c^2_{\mu/\mu,\lambda}-1/2-e^{1,1}_{\mu,\lambda}}\sqrt{1/2+e^{1,1}_{\mu,\lambda}} \\
&\quad\times\sqrt[2\alpha-2]{\frac{1}{\alpha^2d^2}\left(\frac{1/2+e^{1,1}_{\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda}-1/2-e^{1,1}_{\mu,\lambda}}\right)}
\end{aligned} \tag{E.32}$$

for general $\alpha \geq 2$ and to

$$\begin{aligned}
\varphi^{\star}_{x\,st} &= \sqrt{2\mu c^2_{\mu/\mu,\lambda}-1/2-e^{1,1}_{\mu,\lambda}}\sqrt{1/2+e^{1,1}_{\mu,\lambda}} \\
&\quad\times\sqrt{\frac{1}{\alpha^2d^2}\left(\frac{1/2+e^{1,1}_{\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda}-1/2-e^{1,1}_{\mu,\lambda}}\right)} \\
&= \frac{1/2+e^{1,1}_{\mu,\lambda}}{2d}
\end{aligned} \tag{E.33}$$

for $\alpha = 2$.

Stability of the Stationary State

In this paragraph, some numerical evidence is provided for the claim that the stationary solution is asymptotically stable. To this end, system

$$\begin{pmatrix} r \\ \varsigma^{\star} \end{pmatrix} \;=\; \begin{pmatrix} R-\frac{1}{N}\varphi^{\star}_{R}(\sigma^{\star},R) \\ \sigma^{\star}\Big(1+\psi(\sigma^{\star},R)\Big) \end{pmatrix} = f\begin{pmatrix} R \\ \sigma^{\star} \end{pmatrix} = \begin{pmatrix} f_1(R,\sigma^{\star}) \\ f_2(R,\sigma^{\star}) \end{pmatrix} \tag{E.34}$$

is reconsidered. The progress rate (E.21) reads

$$\varphi^\star_R(\sigma^\star, R) = \frac{d\alpha R^{\alpha-1} c_{\mu/\mu,\lambda}}{\sqrt{1+d^2\alpha^2R^{2\alpha-2}}}\sigma^\star - \frac{\sigma^{\star 2}}{2R\mu}$$

and the SAR (E.23) is given as

$$\psi_\infty(\sigma^\star) = \tau^2\left(\frac{1}{2} + e^{1,1}_{\mu,\lambda} - \frac{c_{\mu/\mu,\lambda}}{R}\sqrt{\frac{d^2\alpha^2R^{2\alpha-2}}{1+d^2\alpha^2R^{2\alpha-2}}}\,\sigma^\star\right).$$

The eigenvalues of the Jacobian of f at the stationary point must be determined. The Jacobian reads

$$Df\begin{pmatrix} R \\ \sigma^\star \end{pmatrix} = \begin{pmatrix} \frac{\partial}{\partial R} f_1(R,\sigma^\star) & \frac{\partial}{\partial \sigma^\star} f_1(R,\sigma^\star) \\ \frac{\partial}{\partial R} f_2(R,\sigma^\star) & \frac{\partial}{\partial \sigma^\star} f_2(R,\sigma^\star) \end{pmatrix}. \tag{E.35}$$

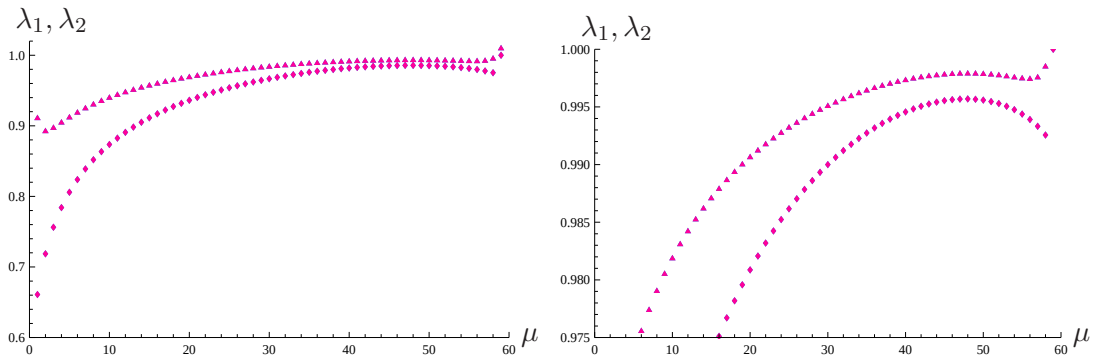
The derivatives can be obtained as follows

$$\begin{aligned}
\frac{\partial}{\partial R} f_1(R,\sigma^\star) &= 1 - \frac{1}{N}\frac{\partial}{\partial R}\varphi^\star_R(\sigma^\star, R) \\
\frac{\partial}{\partial \sigma^\star} f_1(R,\sigma^\star) &= -\frac{1}{N}\frac{\partial}{\partial \sigma^\star}\varphi^\star_R(\sigma^\star, R) \\
\frac{\partial}{\partial R} f_2(R,\sigma^\star) &= \sigma^\star\frac{\partial}{\partial R}\psi(\sigma^\star, R) \\
\frac{\partial}{\partial \sigma^\star} f_2(R,\sigma^\star) &= 1 + \psi(\sigma^\star, R) + \sigma^\star\frac{\partial}{\partial \sigma^\star}\psi(\sigma^\star, R)
\end{aligned}$$

with

$$\begin{aligned}
\frac{\partial}{\partial R}\varphi^\star_R(\sigma^\star, R) &= c_{\mu/\mu,\lambda}\sigma^\star\Bigg(\frac{d\alpha(\alpha-1)R^{\alpha-2}}{\sqrt{1+d^2\alpha^2R^{2\alpha-2}}} \\
&\quad - \frac{d^3\alpha^3(2\alpha-2)R^{(\alpha-1)(2\alpha-3)}}{2\sqrt{1+d^2\alpha^2R^{2\alpha-2}^3}}\Bigg) \\
&\quad + \frac{\sigma^{\star 2}}{2R^2\mu} \qquad \text{(E.36)} \\
\frac{\partial}{\partial \sigma^\star}\varphi^\star_R(\sigma^\star, R) &= \frac{d\alpha R^{\alpha-1}c_{\mu/\mu,\lambda}}{\sqrt{1+d^2\alpha^2R^{2\alpha-2}}} - \frac{\sigma^\star}{R\mu} \\
\frac{\partial}{\partial R}\psi(\sigma^\star, R) &= \tau^2 c_{\mu/\mu,\lambda}\sigma^\star\Bigg(-\frac{d\alpha(\alpha-2)R^{\alpha-3}}{\sqrt{1+d^2\alpha^2R^{2\alpha-2}}}\,\sigma^\star) \\
&= +\frac{d^3\alpha^3(2\alpha-2)R^{(\alpha-2)(2\alpha-3)}}{2\sqrt{1+d^2\alpha^2R^{2\alpha-2}^3}}\,\sigma^\star\Bigg) \qquad \text{(E.37)} \\
\frac{\partial}{\partial \sigma^\star}\psi(\sigma^\star, R) &= -\tau^2\frac{c_{\mu/\mu,\lambda}}{R}\sqrt{\frac{d^2\alpha^2R^{2\alpha-2}}{1+d^2\alpha^2R^{2\alpha-2}}}. \qquad \text{(E.38)}
\end{aligned}$$

At this point, a further analytical analysis is not carried out. Instead, the eigenvalues will be obtained numerically using MATHEMATICA. Therefore, only some numerical evidence can be provided to support the claim of stability. Figure E.1 shows some numerically obtained eigenvalues for $(\mu/\mu_I, 60)$-ES as functions of the parent number μ for some choices of N. The learning rate is set to $1/\sqrt{N}$. As it can be seen, the eigenvalues are smaller than one as long as μ is not too close to λ. In these cases, the larger eigenvalue exceeds one indicating instability. Again, decreasing τ increases the eigenvalues (see the discussion in Appendix D.2.1).



a) $N = 30$ b) $N = 100$

Figure E.1: Numerically obtained eigenvalues for $(\mu/\mu_I, 60)$-ES. The search space dimensionalities examined were $N = 30$ and $N = 100$. The learning rate τ was set to $\tau = 1/\sqrt{N}$. Two values of the ridge parameter d were analyzed. The results for $d = 1$ are indicated using red-colored symbols, whereas blue symbols denote the results obtained for $d = 5$. However, the graphs for both values overlap. The smaller eigenvalue is indicated using diamond-shaped symbols. Triangles stand for the higher eigenvalue.

E.2 The Noisy Ridge

In this section, the derivation of the main results for Chapter 5.2, i.e., for ES on the noisy sharp and parabolic ridge are presented. Again, the noise is modeled using the standard approach with an additive normally distributed noise term. Subsection E.2.1 describes how to obtain the main results for the sharp ridge, whereas Subsection E.2.2 addresses the parabolic ridge.

E.2.1 The Sharp Ridge

This subsection is devoted to the noisy sharp ridge. The noise is modeled using the standard approach with an additive normally distributed noise term. First, the stationary points are derived. Afterwards, the local stability of these fixed points is investigated. As the next step, it is shown that the same d-constant as in the undisturbed case is the decisive parameter deciding the main behavior of the ES. Finally the non-normalized stationary values are derived.

The Derivation of the Stationary Points In this paragraph, it is shown that the stationary state of the system (5.55)

$$\begin{pmatrix} \sigma_\epsilon^{*(g+1)} \\ \langle \varsigma^{*(g+1)} \rangle \end{pmatrix} = \begin{pmatrix} \frac{\sigma_\epsilon^*}{1-\varphi_R^*(\sigma^*,\sigma_\epsilon^*)/N} \\ \sigma^* \left(\frac{1+\psi(\sigma^*,\sigma_\epsilon^*)}{1-\varphi_R^*(\sigma^*,\sigma_\epsilon^*)/N} \right) \end{pmatrix} \tag{E.39}$$

with

$$\varphi_R^\star(\sigma^\star, R) = \frac{d\sigma^{\star 2}}{\sqrt{(1+d^2)\sigma^{\star 2} + \sigma_\epsilon^{\star 2}}} c_{\mu/\mu,\lambda} - \frac{\sigma^{\star 2}}{2R\mu} \tag{E.40}$$

(cf. (5.47)) and

$$\psi(\sigma^\star, R) = \tau^2 \left(1/2 + e_{\mu,\lambda}^{1,1} \frac{(1+d^2)\sigma^{\star 2}}{(1+d^2)\sigma^{\star 2} + \sigma_\epsilon^{\star 2}} \right.$$

$$-c_{\mu/\mu,\lambda}\frac{d\sigma^{\star 2}}{R\sqrt{(1+d^2)\sigma^{\star 2}+\sigma_\epsilon^{\star 2}}}\Bigg) \tag{E.41}$$

(cf. (5.48)) is given by either (5.56)

$$\begin{pmatrix}\sigma^*_{\epsilon\, stat_1}\\ \varsigma^*_{stat_1}\end{pmatrix} = \begin{pmatrix} c\\ 0\end{pmatrix} \tag{E.42}$$

with $c \in \mathbb{R}, c \geq 0$ or by (5.57)

$$\begin{pmatrix}\sigma^*_{\epsilon\, stat}\\ \varsigma^*_{stat}\end{pmatrix} = \begin{pmatrix} 2d\mu c_{\mu/\mu,\lambda}\sqrt{\frac{d^2(4\mu c^2_{\mu/\mu,\lambda}-2e^{1,1}_{\mu,\lambda}-1)-2e^{1,1}_{\mu,\lambda}-1}{d^2(4\mu c^2_{\mu/\mu,\lambda}-2e^{1,1}_{\mu,\lambda})-2e^{1,1}_{\mu,\lambda}}}\\ \frac{2d\mu c_{\mu/\mu,\lambda}}{\sqrt{d^2(4\mu c^2_{\mu/\mu,\lambda}-2e^{1,1}_{\mu,\lambda})-2e^{1,1}_{\mu,\lambda}}}\end{pmatrix} \tag{E.43}$$

Considering (5.55), stationarity of the σ^*_ϵ-evolution requires $\sigma^*_\epsilon = 0$ or

$$\begin{aligned}
\varphi^*_R(\sigma^*,\sigma^*_\epsilon) &= 0\\
\Rightarrow 0 &= \frac{d\sigma^{*2}}{\sqrt{(1+d^2)\sigma^{*2}+\sigma^{*2}_\epsilon}}c_{\mu/\mu,\lambda} - \frac{\sigma^{*2}}{2\mu}\ \text{cf.(E.40)\&(E.41))}\\
\Rightarrow \varsigma^*_{stat_1} = 0 \ \bigvee\ & \frac{d}{\sqrt{(1+d^2)\sigma^{*2}+\sigma^{*2}_\epsilon}}c_{\mu/\mu,\lambda} = \frac{1}{2\mu}\\
\Rightarrow \varsigma^*_{stat_1} = 0 \ \bigvee\ & (1+d^2)\sigma^{*2}+\sigma^{*2}_\epsilon = 4d^2\mu^2 c^2_{\mu/\mu,\lambda}.
\end{aligned} \tag{E.44}$$

This relation between the mutation and the noise strength can be used to determine the stationary mutation strength. Demanding stationarity of the ς^*-evolution

$$\begin{aligned}
\varsigma^*_{stat_1} = 0 \ \bigvee\ & \frac{1+\psi(\sigma^*,\sigma^*_\epsilon)}{1-\varphi^*_R(\sigma^*,\sigma^*_\epsilon)/N} = 1\\
\Rightarrow \varsigma^*_{stat_1} = 0 \ \bigvee\ & N\psi(\sigma^*,\sigma^*_\epsilon) = -\varphi^*_R(\sigma^*,\sigma^*_\epsilon)\\
\Rightarrow \varsigma^*_{stat_1} = 0 \ \bigvee\ & N\tau^2\left(\frac{1}{2}+\frac{e^{1,1}_{\mu,\lambda}(1+d^2)\sigma^{*2}}{(1+d^2)\sigma^{*2}+\sigma^*_\epsilon} - \frac{c_{\mu/\mu,\lambda}d\sigma^{*2}}{\sqrt{(1+d^2)\sigma^{*2}+\sigma^{*2}_\epsilon}}\right) = 0\ \text{cf. (5.48)}\\
\Rightarrow \varsigma^*_{stat_1} = 0 \ \bigvee\ & \frac{1}{2}+\frac{e^{1,1}_{\mu,\lambda}(1+d^2)\sigma^{*2}}{(1+d^2)\sigma^{*2}+\sigma^*_\epsilon} - \frac{c_{\mu/\mu,\lambda}d\sigma^{*2}}{\sqrt{(1+d^2)\sigma^{*2}+\sigma^{*2}_\epsilon}} = 0.
\end{aligned} \tag{E.45}$$

As (E.44) and (E.45) show, $\varsigma^*_{stat_1} = 0$ is a stationary state of the deterministic evolution equations (5.55).

A further stationary solution is obtained by inserting the second condition in (E.44) into (E.45) which eliminates the noise strength

$$\begin{aligned}
0 &= \frac{1}{2}+e^{1,1}_{\mu,\lambda}\frac{(1+d^2)\sigma^{*2}}{4d^2\mu^2c^2_{\mu/\mu,\lambda}} - c_{\mu/\mu,\lambda}\frac{d\sigma^{*2}}{2\mu d c_{\mu/\mu,\lambda}}\\
\Rightarrow -\frac{1}{2} &= \sigma^{*2}\Big(e^{1,1}_{\mu,\lambda}\frac{1+d^2}{4d^2\mu^2c^2_{\mu/\mu,\lambda}} - \frac{1}{2\mu}\Big)\\
\Leftrightarrow -\frac{1}{2} &= \sigma^{*2}\Big(\frac{e^{1,1}_{\mu,\lambda}(1+d^2)-2d^2\mu c^2_{\mu/\mu,\lambda}}{4d^2\mu^2c^2_{\mu/\mu,\lambda}}\Big)\\
\Leftrightarrow \sigma^{*2} &= \frac{4d^2\mu^2c^2_{\mu/\mu,\lambda}}{-2e^{1,1}_{\mu,\lambda}(1+d^2)+4d^2\mu c^2_{\mu/\mu,\lambda}}
\end{aligned}$$

$$\Rightarrow \varsigma^*_{stat_2} = \frac{2d\mu c_{\mu/\mu,\lambda}}{\sqrt{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) - 2e^{1,1}_{\mu,\lambda}}}. \tag{E.46}$$

The stationary mutation strength is only defined if the constant d is sufficiently high

$$\begin{aligned} d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) &> 2e^{1,1}_{\mu,\lambda} \\ \Rightarrow d &> \sqrt{\frac{e^{1,1}_{\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda} - e^{1,1}_{\mu,\lambda}}}. \end{aligned} \tag{E.47}$$

The stationary distance to the axis remains to be determined. To this end, the stationary mutation strength (E.46) is plugged into the second condition of (E.44)

$$\begin{aligned} 4d^2\mu^2 c^2_{\mu/\mu,\lambda} &= (1+d^2){\varsigma^*_{stat_2}}^2 + \sigma^{*2}_\epsilon \\ \Rightarrow 4d^2\mu^2 c^2_{\mu/\mu,\lambda} &= (1+d^2)\frac{4d^2\mu^2 c^2_{\mu/\mu,\lambda}}{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) - 2e^{1,1}_{\mu,\lambda}} + \sigma^{*2}_\epsilon \\ \Leftrightarrow \sigma^{*2}_\epsilon &= 4d^2\mu^2 c^2_{\mu/\mu,\lambda}\left(1 - \frac{1+d^2}{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) - 2e^{1,1}_{\mu,\lambda}}\right) \\ \Rightarrow \sigma^*_\epsilon &= 2d\mu c_{\mu/\mu,\lambda}\sqrt{1 - \frac{1+d^2}{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) - 2e^{1,1}_{\mu,\lambda}}} \end{aligned}$$

$$\Rightarrow \sigma^*_\epsilon = 2d\mu c_{\mu/\mu,\lambda}\sqrt{\frac{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1) - 2e^{1,1}_{\mu,\lambda} - 1}{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) - 2e^{1,1}_{\mu,\lambda}}}. \tag{E.48}$$

The normalized stationary noise strength that is obtained in this way gives the stationary distance to the axis

$$\begin{aligned} \sigma^*_\epsilon &= 2d\mu c_{\mu/\mu,\lambda}\sqrt{\frac{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1) - 2e^{1,1}_{\mu,\lambda} - 1}{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) - 2e^{1,1}_{\mu,\lambda}}} \\ \Rightarrow \frac{N}{R_{stat_2}}\sigma_\epsilon &= 2d\mu c_{\mu/\mu,\lambda}\sqrt{\frac{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1) - 2e^{1,1}_{\mu,\lambda} - 1}{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) - 2e^{1,1}_{\mu,\lambda}}} \end{aligned}$$

$$\Rightarrow R_{stat_2} = \frac{N\sigma_\epsilon}{2d\mu c_{\mu/\mu,\lambda}}\sqrt{\frac{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) - 2e^{1,1}_{\mu,\lambda}}{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1) - 2e^{1,1}_{\mu,\lambda} - 1}}. \tag{E.49}$$

Note, the stationary state in (E.49) is only defined for positive arguments of the square root

$$0 \quad \leq \quad \frac{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) - 2e^{1,1}_{\mu,\lambda}}{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1) - 2e^{1,1}_{\mu,\lambda} - 1}. \tag{E.50}$$

Condition (E.50) is fulfilled if the numerator and denominator are both positive or both negative. The latter situation is not allowed, though, since in this case the normalized mutation strength (E.47) is not defined. In the following paragraph, it is shown that the denominator in (E.50) is decisive leading to the same critical d-constant as in the case of the undisturbed sharp ridge.

The Critical d-Constant Let $\mu \leq \lambda/2$. This paragraph is devoted to showing that (5.59)

$$d \quad > \quad d_{crit} := \sqrt{\frac{2e^{1,1}_{\mu,\lambda} + 1}{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1}} \tag{E.51}$$

has to hold for the existence of a stationary state. Note, if $\mu \approx \lambda$, d_{crit} can assume negative values. In the usual range of $\mu : \lambda$-ratios of $\mu \leq \lambda/2$, it is positive, though. The starting point is (E.50)

$$0 \quad \leq \quad \frac{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) - 2e^{1,1}_{\mu,\lambda}}{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1) - 2e^{1,1}_{\mu,\lambda} - 1}.$$

Under the condition of (E.48), (E.50) holds if

$$0 \quad < \quad d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) - 2e^{1,1}_{\mu,\lambda} \Leftrightarrow d^2 > \frac{e^{1,1}_{\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda} - e^{1,1}_{\mu,\lambda}} \text{ and} \tag{E.52}$$

$$0 \quad < \quad d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1) - 2e^{1,1}_{\mu,\lambda} - 1 \Leftrightarrow d^2 > \frac{2e^{1,1}_{\mu,\lambda} + 1}{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1} \tag{E.53}$$

are true. The decisive bound is (E.53), since

$$\begin{aligned}
\frac{2e^{1,1}_{\mu,\lambda} + 1}{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1} &> \frac{e^{1,1}_{\mu,\lambda}}{2\mu c^2_{\mu/\mu,\lambda} - e^{1,1}_{\mu,\lambda}} \\
\Rightarrow \left(2e^{1,1}_{\mu,\lambda} + 1\right)\left(2\mu c^2_{\mu/\mu,\lambda} - e^{1,1}_{\mu,\lambda}\right) &> e^{1,1}_{\mu,\lambda}\left(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1\right) \\
\Leftrightarrow e^{1,1}_{\mu,\lambda}\left(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}\right) + 2\mu c^2_{\mu/\mu,\lambda} - e^{1,1}_{\mu,\lambda} &> e^{1,1}_{\mu,\lambda}\left(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1\right) \\
\Leftrightarrow 2\mu c^2_{\mu/\mu,\lambda} - e^{1,1}_{\mu,\lambda} &> -e^{1,1}_{\mu,\lambda} \\
2\mu c^2_{\mu/\mu,\lambda} &> 0
\end{aligned} \tag{E.54}$$

which holds in general. The bound (E.53) is therefore always greater than (E.52) and the argument in (E.50) fulfilled if (E.51) holds. In other words, the stationary state exists if and only if the axis is approached and (E.51) or (5.59), respectively, is again the decisive parameter.

Stability of the Stationary Point In this paragraph, the stability of the stationary points is analyzed. Let $\mu \leq \lambda/2$. If $d \geq d_{crit}$, numerical evidence is provided that (5.57)

$$\begin{pmatrix} \sigma^*_{\epsilon\, st} \\ \varsigma^*_{st} \end{pmatrix} = \begin{pmatrix} 2d\mu c_{\mu/\mu,\lambda} \sqrt{\frac{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1) - 2e^{1,1}_{\mu,\lambda} - 1}{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) - 2e^{1,1}_{\mu,\lambda}}} \\ \frac{2d\mu c_{\mu/\mu,\lambda}}{\sqrt{d^2 4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}(1+d^2)}} \end{pmatrix} \tag{E.55}$$

is a locally stable fix-point of (5.55)

$$\begin{pmatrix} \sigma^{*(g+1)}_{\epsilon} \\ \langle \varsigma^{*(g+1)} \rangle \end{pmatrix} = \begin{pmatrix} \frac{\sigma^*_\epsilon}{1 - \varphi^*_R(\sigma^*, \sigma^*_\epsilon)/N} \\ \sigma^* \left(\frac{1 + \psi(R, \sigma^*, \sigma^*_\epsilon)}{1 - \varphi^*_R/N} \right) \end{pmatrix} \tag{E.56}$$

whereas (5.56)

$$\begin{pmatrix} \sigma^*_{\epsilon\, stat_1} \\ \varsigma^*_{stat_1} \end{pmatrix} = \begin{pmatrix} c \\ 0 \end{pmatrix} \tag{E.57}$$

with $c \in \mathbb{R}, c \geq 0$ is instable. The latter can be show relatively easily using again the linear approximation. The approach in this section follows closely the one introduced in Appendix D.2.1. The deterministic evolution equations are of the general form $\mathbf{y}^{(g+1)} = f(\mathbf{y}^{(g)}) = (f_1(\mathbf{y}^{(g)}), f_2(\mathbf{y}^{(g)}))^{\mathrm{T}}$. The stability of hyperbolic fixed points can be established via the eigenvalues of the Jacobian of f

$$Df(\mathbf{y}) = \begin{pmatrix} \frac{\partial}{\partial y_1} f_1(\mathbf{y}^{(g)}) & \frac{\partial}{\partial y_2} f_1(\mathbf{y}^{(g)}) \\ \frac{\partial}{\partial y_1} f_2(\mathbf{y}^{(g)}) & \frac{\partial}{\partial y_2} f_2(\mathbf{y}^{(g)}) \end{pmatrix} \tag{E.58}$$

at the fixed point $\mathbf{y} = \mathbf{y_s}$. Provided the absolute values of all eigenvalues are smaller than one, the fixed point is stable. To this end, the partial derivatives must be obtained

$$\begin{aligned}
\frac{\partial}{\partial y_1} f_1(\mathbf{y}^{(g)}) &= \frac{1}{1 - \varphi^*_R(\sigma^*, \sigma^*_\epsilon)/N} + \frac{\sigma^*_\epsilon \frac{\partial}{\partial \sigma^*_\epsilon} \varphi^*_R(\sigma^*, \sigma^*_\epsilon)/N}{\left(1 - \varphi^*_R(\sigma^*, \sigma^*_\epsilon)/N\right)^2} \\
\frac{\partial}{\partial y_1} f_2(\mathbf{y}^{(g)}) &= \sigma^* \Bigg(\frac{\frac{\partial}{\partial \sigma^*_\epsilon} \psi(\sigma^*, \sigma^*_\epsilon)}{1 - \varphi^*_R(\sigma^*, \sigma^*_\epsilon)/N} \\
&\quad + \frac{\frac{\partial}{\partial \sigma^*_\epsilon} \varphi^*_R(\sigma^*, \sigma^*_\epsilon)}{N} \left[\frac{1 + \psi(\sigma^*, \sigma^*_\epsilon)}{\left(1 - \varphi^*_R(\sigma^*, \sigma^*_\epsilon)/N\right)^2} \right] \Bigg) \\
\frac{\partial}{\partial y_2} f_1(\mathbf{y}^{(g)}) &= \frac{\sigma^*_\epsilon}{N \Big(1 - \varphi^*_R(\sigma^*, \sigma^*_\epsilon)/N\Big)^2} \frac{\partial}{\partial \sigma^*} \varphi^*_R(\sigma^*, \sigma^*_\epsilon) \\
\frac{\partial}{\partial y_2} f_2(\mathbf{y}^{(g)}) &= \frac{1 + \psi(\sigma^*, \sigma^*_\epsilon)}{1 - \varphi^*_R(\sigma^*, \sigma^*_\epsilon)/N} \\
&\quad + \sigma^* \Bigg(\frac{\frac{\partial}{\partial \sigma^*} \psi(\sigma^*, \sigma^*_\epsilon)}{1 - \varphi^*_R(\sigma^*, \sigma^*_\epsilon)/N} \\
&\quad + \frac{\frac{\partial}{\partial \sigma^*} \varphi^*_R(\sigma^*, \sigma^*_\epsilon)}{N} \frac{1 + \psi(\sigma^*, \sigma^*_\epsilon)}{\left(1 - \varphi^*_R(\sigma^*, \sigma^*_\epsilon)/N\right)^2} \Bigg).
\end{aligned} \tag{E.59}$$

We need the values of these derivatives at the fix-points. Therefore, $\varphi_R^* = 0$ and $\psi = 0$ hold giving

$$\begin{aligned}
\frac{\partial}{\partial y_1} f_1(\mathbf{y}^{(g)}) &= 1 + \sigma_\epsilon^* \frac{\partial}{\partial \sigma_\epsilon^*} \varphi_R^*(\sigma^*, \sigma_\epsilon^*)/N \\
\frac{\partial}{\partial y_1} f_2(\mathbf{y}^{(g)}) &= \sigma^* \left(\frac{\partial}{\partial \sigma_\epsilon^*} \psi(\sigma^*, \sigma_\epsilon^*) + \frac{\frac{\partial}{\partial \sigma_\epsilon^*} \varphi_R^*(\sigma^*, \sigma_\epsilon^*)}{N} \right) \\
\frac{\partial}{\partial y_2} f_1(\mathbf{y}^{(g)}) &= \frac{\sigma_\epsilon^*}{N} \frac{\partial}{\partial \sigma^*} \varphi_R^*(\sigma^*, \sigma_\epsilon^*) \\
\frac{\partial}{\partial y_2} f_2(\mathbf{y}^{(g)}) &= 1 + \sigma^* \left(\frac{\partial}{\partial \sigma^*} \psi(\sigma^*, \sigma_\epsilon^*) + \frac{\frac{\partial}{\partial \sigma^*} \varphi_R^*(\sigma^*, \sigma_\epsilon^*)}{N} \right).
\end{aligned} \tag{E.60}$$

For continuing, the derivatives of (5.51) and (5.48) are needed

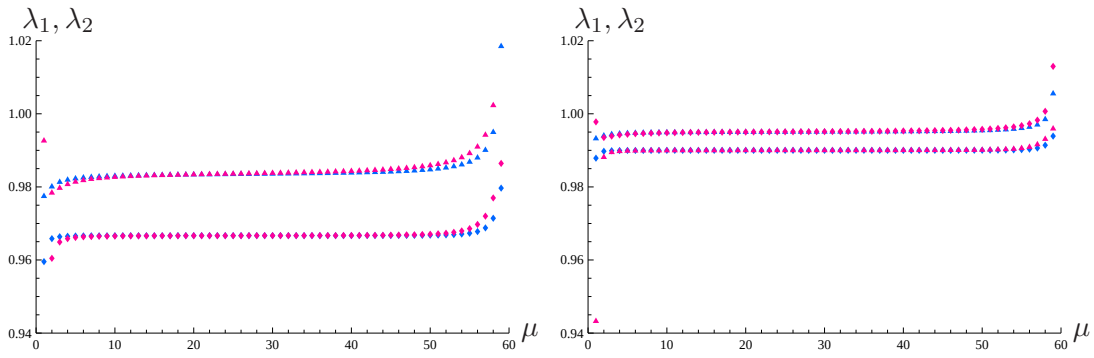
$$\begin{aligned}
\frac{\partial}{\partial \sigma_\epsilon^*} \varphi_R^*(\sigma^*, \sigma_\epsilon^*) &= -c_{\mu/\mu,\lambda} \frac{d\sigma^{*2}}{\sqrt{(1+d^2)\sigma^{*2} + \sigma_\epsilon^{*2}}^3} \sigma_\epsilon^* \\
\frac{\partial}{\partial \sigma_\epsilon^*} \psi(\sigma^*, \sigma_\epsilon^*) &= \tau^2 \sigma_\epsilon^* \left(-2 e_{\mu,\lambda}^{1,1} \frac{(1+d^2)\sigma^{*2}}{\left((1+d^2)\sigma^{*2} + \sigma_\epsilon^{*2}\right)^2} + \frac{c_{\mu/\mu,\lambda} d\sigma^{*2}}{\sqrt{(1+d^2)\sigma^{*2} + \sigma_\epsilon^{*2}}^3} \right) \\
\frac{\partial}{\partial \sigma^*} \varphi_R^*(\sigma^*, \sigma_\epsilon^*) &= \frac{2\sigma^* c_{\mu/\mu,\lambda}}{\sqrt{(1+d^2)\sigma^{*2} + \sigma_\epsilon^{*2}}} - \frac{(1+d^2)\sigma^{*3}}{\sqrt{(1+d^2)\sigma^{*2} + \sigma_\epsilon^{*2}}^3} c_{\mu/\mu,\lambda} - \frac{\sigma^*}{\mu} \\
\frac{\partial}{\partial \sigma^*} \psi(\sigma^*, \sigma_\epsilon^*) &= 2\tau^2 \sigma^* \left(\frac{e_{\mu,\lambda}^{1,1}}{(1+d^2)\sigma^{*2} + \sigma_\epsilon^{*2}} - \frac{(1+d^2) c_{\mu/\mu,\lambda} \sigma^{*2}}{\sqrt{(1+d^2)\sigma^{*2} + \sigma_\epsilon^{*2}}} \right) \\
&\quad + \tau^2 \sigma^* \left(-\frac{2(1+d^2) e_{\mu,\lambda}^{1,1}}{\left((1+d^2)(\sigma^{*2} + \sigma_\epsilon^{*2})\right)^2} + \frac{(1+d^2)^2 c_{\mu/\mu,\lambda} \sigma^{*2}}{\sqrt{(1+d^2)\sigma^{*2} + \sigma_\epsilon^{*2}}^3} \right).
\end{aligned}$$

In the case of the first fixed point, the calculations can be stopped at this point. The equilibrium solution (5.56) with $(\sigma^*_{\epsilon\, stat_1}, \varsigma^*_{stat_1})^{\mathrm{T}} = (c, 0)^{\mathrm{T}}$ is unstable. The eigenvalues of the Jacobian read $\lambda_1 = 1$ and $\lambda_2 = 1 + \tau^2/2$. As seen, (5.56) admits an unstable manifold for $\tau > 0$ leading to a general local instability.

The second fixed point (5.56) requires more effort. In the following, numerical evaluations using MATHEMATICA (R) are provided since inserting the fixed point into the equations above results in complicated expressions. The drawback of this approach is of course that the stability of the stationary point cannot be proven anymore. Instead of a proof, only some numerical evidence can be given that it is stable for the conditions tested. Figure E.2 shows both eigenvalues for $(\mu/\mu_I, 60)$-ES. As before, it is observed that the eigenvalues approach one for $\tau \to 0$. Since there are not any changes of the mutation strength for $\tau = 0$, this behavior was to be expected. If μ approaches λ, the larger eigenvalue may exceed one, indicating instability for $\mu \approx \lambda$. The influence of the constant d appears to be minor.

Non-Normalized Stationary Values In this paragraph, the non-normalized stationary mutation strength and progress rate are obtained. The non-normalized stationary mutation strength can be derived from (5.57) using the stationary distance (5.58). Since

$$\varsigma_{st} = \frac{R_{st}}{N-1} \varsigma^*_{st}, \tag{E.61}$$



a) $N = 30$ b) $N = 100$

Figure E.2: Numerically obtained eigenvalues for $(\mu/\mu_I, 60)$-ES. The search space dimensionalities examined were $N = 30$ and $N = 100$. The learning rate τ was set to $\tau = 1/\sqrt{N}$. Two values of the ridge parameter d were analyzed. The results for $d = 1$ are indicated using red-colored symbols, whereas blue symbols denote the results obtained for $d = 5$. The graphs for both values are close together, however, and cannot be distinguished easily. The smaller eigenvalue is indicated using diamond-shaped symbols. Triangles stand for the higher eigenvalue.

Eq. (5.62) follows

$$\begin{aligned} \varsigma_{st} &= \frac{R_{st}}{N-1} \frac{2d\mu c_{\mu/\mu,\lambda}}{\sqrt{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) - 2e^{1,1}_{\mu,\lambda}}} \\ &= \frac{\sigma_\epsilon (N-1)}{(N-1)2d\mu c_{\mu/\mu,\lambda}} \sqrt{\frac{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) - 2e^{1,1}_{\mu,\lambda}}{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1) - (2e^{1,1}_{\mu,\lambda} + 1)}} \\ &\quad \times \frac{2d\mu c_{\mu/\mu,\lambda}}{\sqrt{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) - 2e^{1,1}_{\mu,\lambda}}} \end{aligned} \tag{E.62}$$

$$\Rightarrow \varsigma_{st} = \frac{\sigma_\epsilon}{\sqrt{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1) - (2e^{1,1}_{\mu,\lambda} + 1)}}. \tag{E.63}$$

Similarly, the non-normalized progress rate (5.61) is obtained. Plugging the normalized mutation strength and noise strength in (5.63) into the progress rate (5.50) leads to

$$\begin{aligned} \varphi^{st}_x(\sigma_\epsilon) &= \frac{R_{st}}{N} \varphi^{*\,st}_x(\varsigma^*_{st}, \sigma^*_{\epsilon\,st}) = \frac{R_{st}}{N} \frac{\varsigma^{*\,2}_{st}}{2d\mu} \\ &= \frac{\sigma_\epsilon}{2d\mu c_{\mu/\mu,\lambda}} \sqrt{\frac{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) - 2e^{1,1}_{\mu,\lambda}}{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1) - 2e^{1,1}_{\mu,\lambda} - 1}} \\ &\quad \times \frac{2d\mu c^2_{\mu/\mu,\lambda}}{d^2(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda}) - 2e^{1,1}_{\mu,\lambda}} \end{aligned}$$

which finally gives (5.61)

$$\begin{aligned}\varphi_x^{st}(\sigma_\epsilon) &= \sigma_\epsilon c_{\mu/\mu,\lambda}\sqrt{\frac{1}{d^2(4\mu c_{\mu/\mu,\lambda}^2 - 2e_{\mu,\lambda}^{1,1} - 1) - 2e_{\mu,\lambda}^{1,1} - 1}} \\ &\quad \times\sqrt{\frac{1}{d^2(4\mu c_{\mu/\mu,\lambda}^2 - 2e_{\mu,\lambda}^{1,1}) - 2e_{\mu,\lambda}^{1,1}}}. \end{aligned} \tag{E.64}$$

E.2.2 The Noisy Parabolic Ridge

This subsection describes how the main results in Subsection 5.2.2 are obtained. This consists mainly in obtaining the stationary distance to the axis, i.e., in deriving and solving the respective equation.

Derivation of the Third-Order Polynomial The starting point is the stationarity condition for the R-evolution (5.74)

$$\begin{aligned} & 0 = \varphi_R^\star(R, \sigma^\star, \sigma_\epsilon^\star) \\ \Rightarrow\quad & 0 = \frac{2dRc_{\mu/\mu,\lambda}}{\sqrt{(1+4d^2R^2)\sigma^{\star 2} + \sigma_\epsilon^{\star 2}}}\sigma^{\star 2} - \frac{\sigma^{\star 2}}{2R\mu} \\ \Rightarrow\quad & \sigma^\star = 0 \bigvee 16d^2R^4\mu^2 c_{\mu/\mu,\lambda}^2 - (1+4d^2R^2)\sigma^{\star 2} - \sigma_\epsilon^{\star 2} = 0 \\ \Rightarrow\quad & \sigma^\star = 0 \bigvee \sigma^{\star 2} = \frac{16d^2R^4\mu^2 c_{\mu/\mu,\lambda}^2}{1+4d^2R^2} - \frac{\sigma_\epsilon^{\star 2}}{1+4d^2R^2}. \end{aligned} \tag{E.65}$$

Since the $\langle\varsigma^\star\rangle$-evolution becomes also stationary, either $\sigma^\star = 0$ has to hold or (5.75)

$$\begin{aligned} & 0 = \psi(\sigma^\star, \sigma_\epsilon^\star) \\ & = \tau^2\Big(1/2 - \frac{2dRc_{\mu/\mu,\lambda}\sigma^{\star 2}}{R\sqrt{(1+4d^2R^2)\sigma^{\star 2} + \sigma_\epsilon^{\star 2}}} + e_{\mu/\mu,\lambda}^{1,1}\frac{(1+4d^2R^2)\sigma^{\star 2}}{(1+4d^2R^2)\sigma^{\star 2} + \sigma_\epsilon^{\star 2}}\Big) \end{aligned} \tag{E.66}$$

must be fulfilled. Inserting (5.74) into (5.75) leads to a third-order polynomial in R^2 (**??**)

$$\begin{aligned} 0 &= \frac{1}{2} - \frac{2dc_{\mu/\mu,\lambda}}{4d\mu c_{\mu/\mu,\lambda}R^2}\Big(\frac{16d^2R^4\mu^2c_{\mu/\mu,\lambda}^2 - \sigma_\epsilon^{\star 2}}{1+4d^2R^2}\Big) + \frac{e_{\mu,\lambda}^{1,1}}{16d^2R^4\mu^2c_{\mu/\mu,\lambda}^2}\Big(16d^2R^4\mu^2c_{\mu/\mu,\lambda}^2 - \sigma_\epsilon^{\star 2}\Big) \\ &= \frac{1}{2} - \frac{1}{2\mu R^2}\Big(\frac{16d^2R^4\mu^2c_{\mu/\mu,\lambda}^2 - \sigma_\epsilon^{\star 2}}{1+4d^2R^2}\Big) + \frac{e_{\mu,\lambda}^{1,1}}{16d^2R^4\mu^2c_{\mu/\mu,\lambda}^2}\Big(16d^2R^4\mu^2c_{\mu/\mu,\lambda}^2 - \sigma_\epsilon^{\star 2}\Big) \\ &= 8d^2\mu^2c_{\mu/\mu,\lambda}^2R^4 - 8d^2\mu c_{\mu/\mu,\lambda}^2R^2\Big(\frac{16d^2R^4\mu^2c_{\mu/\mu,\lambda}^2 - \sigma_\epsilon^{\star 2}}{1+4d^2R^2}\Big) \\ &\quad + e_{\mu,\lambda}^{1,1}\Big(16d^2R^4\mu^2c_{\mu/\mu,\lambda}^2 - \sigma_\epsilon^{\star 2}\Big) \\ &= 8d^2\mu^2c_{\mu/\mu,\lambda}^2R^4(1+4d^2R^2) - 8d^2\mu c_{\mu/\mu,\lambda}^2R^2(16d^2R^4\mu^2c_{\mu/\mu,\lambda}^2 - \sigma_\epsilon^{\star 2}) \\ &\quad + e_{\mu,\lambda}^{1,1}\Big(16d^2R^4\mu^2c_{\mu/\mu,\lambda}^2 - \sigma_\epsilon^{\star 2}\Big)(1+4d^2R^2) \\ &= 8d^2\mu^2c_{\mu/\mu,\lambda}^2R^4 + 32d^4\mu^2c_{\mu/\mu,\lambda}^2R^6 - 128d^4R^6\mu^3c_{\mu/\mu,\lambda}^4 + 8d^2\mu c_{\mu/\mu,\lambda}^2\sigma_\epsilon^{\star 2}R^2 \\ &\quad + 64d^4\mu^2c_{\mu/\mu,\lambda}^2e_{\mu,\lambda}^{1,1}R^6 + 16d^2\mu^2c_{\mu/\mu,\lambda}^2R^2 - 4d^2e_{\mu,\lambda}^{1,1}\sigma_\epsilon^{\star 2}R^2 - e_{\mu,\lambda}^{1,1}\sigma_\epsilon^{\star 2} \end{aligned}$$

$$
\begin{aligned}
&= -R^6\Big(128d^4\mu^3c^4_{\mu/\mu,\lambda} - 64d^4\mu^2c^2_{\mu/\mu,\lambda}e^{1,1}_{\mu,\lambda} - 32d^4\mu^2c^2_{\mu/\mu,\lambda}\Big)\\
&\quad +R^4\Big(8d^2\mu^2c^2_{\mu/\mu,\lambda} + 16d^2\mu^2c^2_{\mu/\mu,\lambda}e^{1,1}_{\mu,\lambda}\Big) + R^2\Big(8d^2\mu c^2_{\mu/\mu,\lambda}\sigma^{\star 2}_\epsilon - 4d^2e^{1,1}_{\mu,\lambda}\sigma^{\star 2}_\epsilon\Big) - e^{1,1}_{\mu,\lambda}\sigma^{\star 2}_\epsilon\\
&= R^6 32d^4\mu^2c^2_{\mu/\mu,\lambda}\Big(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1\Big) - R^4 8d^2\mu^2c^2_{\mu/\mu,\lambda}\Big(2e^{1,1}_{\mu,\lambda} + 1\Big)\\
&\quad -R^2 4\sigma^{\star 2}_\epsilon d^2\Big(2\mu c^2_{\mu/\mu,\lambda} - 1\Big) + e^{1,1}_{\mu,\lambda}\sigma^{\star 2}_\epsilon\\
&= R^6 - R^4\frac{8d^2\mu^2c^2_{\mu/\mu,\lambda}}{32d^4\mu^2c^2_{\mu/\mu,\lambda}}\Big(\frac{2e^{1,1}_{\mu/\mu,\lambda}+1}{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1}\Big)\\
&\quad -R^2\frac{4\sigma^{\star 2}_\epsilon d^2}{32d^4\mu^2c^2_{\mu/\mu,\lambda}}\Big(\frac{2\mu c^2_{\mu/\mu,\lambda} - e^{1,1}_{\mu,\lambda}}{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1}\Big) + \frac{e^{1,1}_{\mu,\lambda}\sigma^{\star 2}_\epsilon}{32d^4\mu^2c^2_{\mu/\mu,\lambda}\Big(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1\Big)}\\
&= R^6 - R^4\frac{1}{4d^2}\Big(\frac{2e^{1,1}_{\mu,\lambda}+1}{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1}\Big)\\
&\quad -R^2\frac{\sigma^{\star 2}_\epsilon}{8d^2\mu^2c^2_{\mu/\mu,\lambda}}\Big(\frac{2\mu c^2_{\mu/\mu,\lambda} - e^{1,1}_{\mu,\lambda}}{4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1}\Big) + \frac{e^{1,1}_{\mu,\lambda}\sigma^{\star 2}_\epsilon}{32d^4\mu^2c^2_{\mu/\mu,\lambda}\Big(4\mu c^2_{\mu/\mu,\lambda} - 2e^{1,1}_{\mu,\lambda} - 1\Big)}. \quad \text{(E.67)}
\end{aligned}
$$

The cubic polynomial (E.67) in R^2 leads to analytical solutions. Let us first consider the general cubic equation.

Solutions of the Cubic Equation $x^3 - ax^2 - bx + c = 0$ The solutions can be given as follows (see, e.g., [34])

$$
\begin{aligned}
x_1 &= \frac{a}{3}\\
&\quad -\frac{\sqrt[3]{2}(-a^2-3b)}{3\sqrt[3]{2a^3+9ab-27c+3\sqrt{3}\sqrt{27c^2-18abc-4a^3c-4b^3-a^2b^2}}}\\
&\quad +\frac{1}{3\sqrt[3]{2}}\\
&\quad \times\sqrt[3]{2a^3+9ab-27c+3\sqrt{3}\sqrt{27c^2-18abc-4a^3c-4b^3-a^2b^2}}\\
x_2 &= \frac{a}{3} + \frac{1+i\sqrt{3}}{3\sqrt[3]{4}}\\
&\quad \times\frac{-a^2-3b}{\sqrt[3]{2a^3+9ab-27c+3\sqrt{3}\sqrt{27c^2-18abc-4a^3c-4b^3-a^2b^2}}}\\
&\quad -\frac{1-i\sqrt{3}}{6\sqrt[3]{2}}\\
&\quad \times\sqrt[3]{2a^3+9ab-27c+3\sqrt{3}\sqrt{27c^2-18abc-4a^3c-4b^3-a^2b^2}}\\
x_3 &= \frac{a}{3} + \frac{1-i\sqrt{3}}{3\sqrt[3]{4}}\\
&\quad \times\frac{-a^2-3b}{\sqrt[3]{2a^3+9ab-27c+3\sqrt{3}\sqrt{27c^2-18abc-4a^3c-4b^3-a^2b^2}}}
\end{aligned}
$$

$$-\frac{1+i\sqrt{3}}{6\sqrt[3]{2}} \times \sqrt[3]{2a^3+9ab-27c+3\sqrt{3}\sqrt{27c^2-18abc-4a^3c-4b^3-a^2b^2}}. \quad \text{(E.68)}$$

Considering (E.67) and (E.68), the coefficients read

$$\begin{aligned} a &= \frac{1}{4d^2}\Big(\frac{2e^{1,1}_{\mu/\mu,\lambda}+1}{4\mu c^2_{\mu/\mu,\lambda}-2e^{1,1}_{\mu,\lambda}-1}\Big) \\ b &= \frac{\sigma^{\star 2}_\epsilon}{8d^2\mu^2 c^2_{\mu/\mu,\lambda}}\Big(\frac{2\mu c^2_{\mu/\mu,\lambda}-e^{1,1}_{\mu,\lambda}}{4\mu c^2_{\mu/\mu,\lambda}-2e^{1,1}_{\mu,\lambda}-1}\Big) \\ c &= \frac{e^{1,1}_{\mu,\lambda}\sigma^{\star 2}_\epsilon}{32d^4\mu^2 c^2_{\mu/\mu,\lambda}(4\mu c^2_{\mu/\mu,\lambda}-2e^{1,1}_{\mu,\lambda}-1)}. \end{aligned}$$

Although analytical solutions can be provided, the results are quite clumsy. Therefore, the solutions are not given explicitely.

Bibliography

[1] L. Altenberg. The evolution of evolvability in genetic programming. In K. Kinnear, editor, *Advances in Genetic Programming*, pages 47–74. MIT Press, Cambridge, MA, 1994.

[2] P. J. Angeline. Adaptive and self-adaptive evolutionary computations. In M. Palaniswami and Y. Attikiouzel, editors, *Computational Intelligence: A Dynamic Systems Perspective*, pages 152–163. IEEE Press, 1995.

[3] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja. *A First Course in Order Statistics*. Wiley, New York, 1992.

[4] D. V. Arnold. *Noisy Optimization with Evolution Strategies*. Kluwer Academic Publishers, Dordrecht, 2002.

[5] D. V. Arnold. Cumulative step length adaptation on ridge functions. In T.P. Runarsson et al., editors, *Parallel Problem Solving from Nature PPSN IX*, pages 11–20. Springer Verlag Heidelberg, 2006.

[6] D. V. Arnold and H.-G. Beyer. Performance analysis of evolution strategies with multi-recombination in high-dimensional $\mathbb{R}^n$-search spaces disturbed by noise. *Theoretical Computer Science*, 289:629–647, 2002.

[7] D. V. Arnold and H.-G. Beyer. On the benefits of populations for noisy optimization. *Evolutionary Computation*, 11(2):111–127, 2003.

[8] D. V. Arnold and H.-G. Beyer. Evolution strategies with cumulative step length adaptation on the noisy parabolic ridge. Technical Report CS-2006-02, Dalhousie University, Faculty of Computer Science, January 2006.

[9] D. V. Arnold and H.-G. Beyer. Evolution strategies with cumulative step length adaptation on the noisy parabolic ridge. *Natural Computing*, 2007. To appear.

[10] D. V. Arnold and A. MacLeod. Hierarchically organised evolution strategies on the parabolic ridge. Technical Report CS-2006-03, Dalhousie University, Faculty of Computer Science, January 2006.

[11] D. V. Arnold and A. MacLeod. Step length adaptation on ridge functions. Technical Report CS-2006-08, Dalhousie University, Faculty of Computer Science, August 2006.

[12] A. Auger. Convergence results for the $(1, \lambda)$-SA-ES using the theory of ϕ-irreducible Markov chains. *Theoretical Computer Science*, 334:35–69, 2005.

[13] T. Bäck. The interaction of mutation rate, selection, and self-adaptation within a genetic algorithm. In R. Männer and B. Manderick, editors, *Parallel Problem Solving from Nature, 2*, pages 85–94. North Holland, Amsterdam, 1992.

[14] T. Bäck. Self-adaptation in genetic algorithms. In F. J. Varela and P. Bourgine, editors, *Toward a Practice of Autonomous Systems: proceedings of the first European conference on Artificial Life*, pages 263–271. MIT Press, 1992.

[15] T. Bäck. Optimal mutation rates in genetic search. In S. Forrest, editor, *Proceedings of the Fifth International Conference on Genetic Algorithms*, pages 2–8, San Mateo (CA), 1993. Morgan Kaufmann.

[16] T. Bäck. Self-adaptation. In T. Bäck, D. Fogel, and Z. Michalewicz, editors, *Handbook of Evolutionary Computation*, pages C7.1:1–C7.1:15. Oxford University Press, New York, 1997.

[17] T. Bäck and M. Schütz. Intelligent mutation rate control in canonical genetic algorithms. In *ISMIS*, pages 158–167, 1996.

[18] J. D. Bagley. *The Behavior of Adaptive Systems Which Employ Genetic and Correlation Algorithms*. PhD thesis, University of Michigan, 1967.

[19] H.-G. Beyer. Estimating the steady-state of CSA-ES on ridge functions. The Theory of Evolutionary Algorithms, Dagstuhl Seminar, Wadern, Germany, February 2004.

[20] H.-G. Beyer. Toward a theory of evolution strategies: On the benefit of sex – the $(\mu/\mu, \lambda)$-theory. *Evolutionary Computation*, 3(1):81–111, 1995.

[21] H.-G. Beyer. Toward a theory of evolution strategies: Self-adaptation. *Evolutionary Computation*, 3(3):311–347, 1996.

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

[23] H.-G. Beyer. *The Theory of Evolution Strategies*. Natural Computing Series. Springer, Heidelberg, 2001.

[24] H.-G. Beyer and D. V. Arnold. Fitness noise and localization errors of the optimum in general quadratic fitness models. In W. Banzhaf, J. Daida, A.E. Eiben, M.H. Garzon, V. Honavar, M. Jakiela, and R.E. Smith, editors, *GECCO-99: Proceedings of the Genetic and Evolutionary Computation Conference*, pages 817–824, San Francisco, CA, 1999. Morgan Kaufmann.

[25] H.-G. Beyer and D. V. Arnold. The steady state behavior of $(\mu/\mu_I, \lambda)$-ES on ellipsoidal fitness models disturbed by noise. In E. Cantú-Paz et al., editors, *GECCO-2003: Proceedings of the Genetic and Evolutionary Computation Conference*, pages 525–536, Berlin, Germany, 2003. Springer.

[26] H.-G. Beyer and K. Deb. On self-adaptive features in real-parameter evolutionary algorithms. *IEEE Transactions on Evolutionary Computation*, 5(3):250–270, 2001.

[27] H.-G. Beyer and S. Meyer-Nieberg. Self-adaptation of evolution strategies under noisy fitness evaluations. *Genetic Programming and Evolvable Machines*, 7(4):295–328, 2006.

[28] H.-G. Beyer and S. Meyer-Nieberg. Self-adaptation on the ridge function class: First results for the sharp ridge. In T.P. Runarsson et al., editors, *Parallel Problem Solving from Nature PPSN IX*, pages 71–80, Heidelberg, 2006. Springer Verlag.

[29] H.-G. Beyer and H.-P. Schwefel. Evolution strategies: A comprehensive introduction. *Natural Computing*, 1(1):3–52, 2002.

[30] H.-G. Beyer, H.-P. Schwefel, and I. Wegener. How to analyse evolutionary algorithms. *Theoretical Computer Science*, 287:101–130, 2002.

[31] A. Bienvenüe and O. François. Global convergence for evolution strategies in spherical problems: Some simple proofs and difficulties. *Theoretical Computer Science*, 308:269–289, 2003.

[32] S. Blinnikov and R. Moessner. Expansions for nearly Gaussian distributions. *Astron. Astrophys. Suppl. Ser.*, 130, 1998.

[33] M. Braun. *Differential Equations and their Applications*. Springer, Berlin, 1998.

[34] I. N. Bronshtein and K. A. Semendyayev. *Handbook of Mathematics*. Verlag Harri Deutsch, Frankfurt, 1985.

[35] L. Davis. Adapting operator probabilities in genetic algorithms. In J. D. Schaffer, editor, *Proc. 3rd Int'l Conf. on Genetic Algorithms*, pages 61–69, San Mateo, CA, 1989. Morgan Kaufmann.

[36] M. W. Davis. The natural formation of gaussian mutation strategies in evolutionary programming. In *roceedings of the Third Annual Conference on Evolutionary Programming*, San Diego, CA, 1994. Evolutionary Programming Society.

[37] K. Deb and R. B. Agrawal. Simulated binary crossover for continuous search space. *Complex Systems*, 9:115–148, 1995.

[38] K. Deb and H.-G. Beyer. Self-adaptation in real-parameter genetic algorithms with simulated binary crossover. In W. Banzhaf, J. Daida, A.E. Eiben, M.H. Garzon, V. Honavar, M. Jakiela, and R.E. Smith, editors, *GECCO-99: Proceedings of the Genetic and Evolutionary Computation Conference*, pages 172–179, San Francisco, CA, 1999. Morgan Kaufmann.

[39] K. Deb and H.-G. Beyer. Self-adaptive genetic algorithms with simulated binary crossover. Series CI 61/99, SFB 531, University of Dortmund, March 1999.

[40] K. Deb and H.-G. Beyer. Self-adaptive genetic algorithms with simulated binary crossover. *Evolutionary Computation*, 9(2):197–221, 2001.

[41] K. Deb and M. Goyal. A robust optimization procedure for mechanical component design based on genetic adaptive search. *Transactions of the ASME: Journal of Mechanical Desiogn*, 120(2):162–164, 1998.

[42] A. E. Eiben, R. Hinterding, and Z. Michalewicz. Parameter control in evolutionary algorithms. *IEEE Transactions on Evolutionary Computation*, 3(2):124–141, 1999.

[43] A. E. Eiben and J. E. Smith. *Introduction to Evolutionary Computing*. Natural Computing Series. Springer, Berlin, 2003.

[44] L. J. Eshelman and J. D. Schaffer. Real-coded genetic algorithms and interval schemata. In L. D. Whitley, editor, *Foundations of Genetic Algorithms, 2*, pages 187–202. Morgan Kaufmann, San Mateo, CA, 1993.

[45] D. B. Fogel. *Evolving Artificial Intelligence*. PhD thesis, University of California, San Diego, 1992.

[46] D. B. Fogel, L. J. Fogel, and J. W. Atma. Meta-evolutionary programming. In R.R. Chen, editor, *Proc. of 25th Asilomar Conference on Signals, Systems & Computers*, pages 540–545, Pacific Grove, CA, 1991.

[47] L. J. Fogel. Autonomous automata. *Industrial Research*, 4:14–19, 1962.

[48] L. J. Fogel, A. J. Owens, and M. J. Walsh. *Artificial Intelligence through Simulated Evolution*. Wiley, New York, 1966.

[49] M. Glickman and K. Sycara. Reasons for premature convergence of self-adapting mutation rates. In *Proc. of the 2000 Congress on Evolutionary Computation*, pages 62–69, Piscataway, NJ, 2000. IEEE Service Center.

[50] D.E. Goldberg. *Genetic Algorithms in Search, Optimization, and Machine Learning*. Addison Wesley, Reading, MA, 1989.

[51] L. Grünz and H.-G. Beyer. Some observations on the interaction of recombination and self-adaptation in evolution strategies. In P.J. Angeline, editor, *Proceedings of the CEC'99 Conference*, pages 639–645, Piscataway, NJ, 1999. IEEE.

[52] N. Hansen. An analysis of mutative σ-self-adaptation on linear fitness functions. *Evolutionary Computation*, 14(3):255–275, 2006.

[53] N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[54] W. E. Hart and J. M. DeLaurentis. Convergence of a discretized self-adapative evolutionary algorithm on multi-dimensional problems. submitted.

[55] W. E. Hart, J. M. DeLaurentis, and L. A. Ferguson. On the convergence of an implicitly self-adaptive evolutionary algorithm on one-dimensional unimodal problems. *IEEE Transactions on Evolutionary Computation*, 2003. To appear.

[56] M. Herdy. Reproductive isolation as strategy parameter in hierachically organized evolution strategies. In R. Männer and B. Manderick, editors, *Parallel Problem Solving From Nature, PPSN II,*, Berlin, 1992. Springer-Verlag.

[57] M. Herdy. Reproductive isolation as strategy parameter in hierarchically organized evolution strategies. In R. Männer and B. Manderick, editors, *Parallel Problem Solving from Nature, 2*, pages 207–217. Elsevier, Amsterdam, 1992.

[58] J. H. Holland. Outline for a logical theory of adaptive systems. *JACM*, 9:297–314, 1962.

[59] J. H. Holland. *Adaptation in Natural and Artificial Systems*. The University of Michigan Press, Ann Arbor, 1975.

[60] C. Igel and M. Toussaint. Neutrality and self-adaptation. *Natural Computing: an international journal*, 2(2):117–132, 2003.

[61] J. Jägersküpper. Analysis of a simple evolutionary algorithm for mimimization in euclidean spaces. In J. Baeten et al., editors, *Proceedings of the 30th International Colloquium on Automata, Languages, and Programming (ICALP 2003)*, volume 2719 of *Lecture Notes in Computer Science*, pages 1068–1079. Springer, 2003.

[62] J. Jägersküpper. *Probabilistic Analysis of Evolution Strategies Using Isotropic Mutations*. Ph.d. thesis, Dortmund University, Dortmund, 2006.

[63] B. A. Julstrom. Adaptive operator probabilities in a genetic algorithm that applies three operators. In *SAC*, pages 233–238, 1997.

[64] H. Kita. A comparison study of self-adaptation in evolution strategies and real-coded genetic algorithms. *Evolutionary Computation*, 9(2):223–241, 2001.

[65] H. Kita and M. Yamamura. A functional specialization hypothesis for designing genetic algorithms. In *Proc. IEEE International Conference on Systems, Man, and Cybernetics '99*, pages 579–584, Piscataway, New Jersey, 1999. IEEE Press.

[66] J. K. Kolossa. *Series Approximation Methods in Statistics*. Springer, New York, 2006.

[67] F. Kursawe. *Grundlegende empirische Untersuchungen der Parameter von Evolutionsstrategien — Metastrategien*. P.h.d. thesis, Department of Computer Science, Universität Dortmund, 1999.

[68] C.-Y. Lee and X. Yao. Evolutionary programming using mutations based on the levy probability distribution. *Evolutionary Computation, IEEE Transactions on*, 8(1):1– 13, Feb. 2004.

[69] K.-H. Liang, X. Yao, C. N. Newton, and D. Hoffman. An experimental investigation of self-adaptation in evolutionary programming. In V. W. Porto, N. Saravanan, Waagen D. E., and A. E. Eiben, editors, *Evolutionary Programming*, volume 1447 of *Lecture Notes in Computer Science*, pages 291–300. Springer, 1998.

[70] K.-H. Liang, X. Yao, and C. S. Newton. Adapting self-adaptive parameters in evolutionary algorithms. *Artificial Intelligence*, 15(3):171 – 180, November 2001.

[71] D. G. Luenberger. *Introduction to Dynamic Systems*. Wiley, Chichester, 1979.

[72] M. Lunacek and D. Whitley. Searching for balance: Understanding self-adaptation on ridge functions. In T.P. Runarsson et al., editors, *Parallel Problem Solving from Nature PPSN IX*, pages 82–91. Springer Verlag Heidelberg, 2006.

[73] R. E. Mercer and J. R. Sampson. Adaptive search using a reproductive metaplan. *Kybernetes*, 7:215–228, 1978.

[74] S. Meyer-Nieberg and H.-G. Beyer. On the analysis of self-adaptive recombination strategies: First results. In B. McKay et al., editors, *Proc. 2005 Congress on Evolutionary Computation (CEC'05), Edinburgh, UK*, pages 2341–2348, Piscataway NJ, 2005. IEEE Press.

[75] S. Meyer-Nieberg and H.-G. Beyer. Mutative self-adaptation on the sharp and parabolic ridge. In Ch. Stephens et al., editors, *Foundations of Genetic Algorithms IX (FOGA 2007)*, volume LNCS 4436, pages 70–96, Heidelberg, 2007. Springer Verlag.

[76] S. Meyer-Nieberg and H.-G. Beyer. Self-adaptation in evolutionary algorithms. In F. Lobo, C. Lima, and Z. Michalewicz, editors, *Parameter Setting in Evolutionary Algorithms*, pages 47–76. Springer Verlag, Heidelberg, 2007.

[77] S. P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Springer, 1993.

[78] A. Ostermeier, A. Gawelczyk, and N. Hansen. A derandomized approach to self-adaptation of evolution strategies. *Evolutionary Computation*, 2(4):369–380, 1995.

[79] A. I. Oyman. *Convergence Behavior of Evolution Strategies on Ridge Functions*. Ph.d. thesis, University of Dortmund, Department of Computer Science, 1999.

[80] H. Pruscha. *Vorlesungen über Mathematische Statistik*. B.G. Teubner, 2000.

[81] I. Rechenberg. *Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Frommann-Holzboog Verlag, Stuttgart, 1973.

[82] I. Rechenberg. Evolutionsstrategien. In B. Schneider and U. Ranft, editors, *Simulationsmethoden in der Medizin und Biologie*, pages 83–114. Springer-Verlag, Berlin, 1978.

[83] J. Reed, R. Toombs, and N. A. Barricelli. Simulation of biological evolution and machine learning. i. selection of self-reproducing numeric patterns by data processing machines, effects of hereditary control, mutation type and crossing. *Journal of Theoretical Biology*, 17:319–342, 1967.

[84] R.S. Rosenberg. *Simulation of genetic populations with biochemical properties*. Ph.d. thesis, Univ. Michigan, Ann Arbor, MI, 1967.

[85] G. Rudolph. Self-adaptive mutations may lead to premature convergence. *IEEE Transactions on Evolutionary Computation*, 5(4):410–414, 2001.

[86] J. D. Schaffer and A. Morishima. An adaptive crossover distribution mechanism for genetic algorithms. In J.J. Grefenstette, editor, *Genetic Algorithms and their Applications: Proc. of the Second Int'l Conference on Genetic Algorithms*, pages 36–40, 1987.

[87] H.-P. Schwefel. Adaptive Mechanismen in der biologischen Evolution und ihr Einfluß auf die Evolutionsgeschwindigkeit. Technical report, Technical University of Berlin, 1974. Abschlußbericht zum DFG-Vorhaben Re 215/2.

[88] H.-P. Schwefel. *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie*. Interdisciplinary systems research; 26. Birkhäuser, Basel, 1977.

[89] H.-P. Schwefel. *Numerical Optimization of Computer Models*. Wiley, Chichester, 1981.

[90] M. A. Semenov. Convergence velocity of evolutionary algorithms with self-adaptation. In *GECCO 2002: Proceeding of the Genetic And Evolutionary Computation Conference*, pages 210–213, 2002.

[91] M. A. Semenov and D. A. Terkel. Analysis of convergence of an evolutionary algorithm with self-adaptation using a stochastic Lyapunov function. *Evolutionary Computation*, 11(4):363–379, 2003.

[92] J. E. Smith. *Self-Adaptation in Evolutionary Algorithms*. PhD thesis, University of the West of England, Bristol, 1998.

[93] J. E. Smith. Modelling gas with self adaptive mutation rates. In L. Spector et al., editors, *GECCO 2001: Proceedings of the Genetic and Evolutionary Computation Conference*, pages 599–606, San Francisco, California, USA, 7-11 July 2001. Morgan Kaufmann.

[94] J. E. Smith. Parameter perturbation mechanisms in binary coded gas with self-adaptive mutation. In K. DeJong, R. Poli, and J. Rowe, editors, *Foundations of Genetic Algorithms 7*, pages 329–346. Morgan Kauffman, 2004.

[95] J. E. Smith and T. C. Fogarty. Recombination strategy adaptation via evolution of gene linkage. In *Proc. of the 1996 IEEE International Conference on Evolutionary Computation*, pages 826–831. IEEE Publishers, 1996.

[96] J. E. Smith and T. C. Fogarty. Self-adaptation of mutation rates in a steady state genetic algorithm. In *Proceedings of 1996 IEEE Int'l Conf. on Evolutionary Computation (ICEC '96)*, pages 318–323. IEEE Press, NY, 1996.

[97] J. E. Smith and T. C. Fogarty. Operator and parameter adaptation in genetic algorithms. *Soft Computing*, 1(2):81–87, June 1997.

[98] W. Spears. Adapting crossover in evolutionary algorithms. In *Proceedings of the Evolutionary Programming Conference*, pages 367–384, 1995.

[99] W. Spears. *Evolutionary Algorithms: The Role of Mutation and Recombination*. Springer-Verlag, Heidelberg, 2000.

[100] C. Stone and J. E. Smith. Strategy parameter variety in self-adaptation of mutation rates. In W. B. Langdon et al., editors, *GECCO 2002: Proceeding of the Genetic And Evolutionary Computation Conference*, pages 586–593. Morgan Kaufmann, 2002.

[101] H.-M. Voigt, H. Mühlenbein, and D. Cvetković. Fuzzy recombination for the breeder genetic algorithm. In L. J. Eshelman, editor, *Proc. 6th Int'l Conf. on Genetic Algorithms*, pages 104–111, San Francisco, CA, 1995. Morgan Kaufmann Publishers, Inc.

[102] R. Weinberg. *Computer Simulation of a Living Cell*. PhD thesis, University of Michigan, 1970.

[103] S. Wiggins. *Introduction to Applied Nonlinear Dynamical Systems and Chaos*. Springer, New York, 1990.

[104] J. M. Won and J. S. Lee. Premature convergence avoidance of self-adaptive evolution strategy,. In *The 5th International Conference on Simulated Evolution And Learning*, Busan, Korea, Oct 2004.

[105] X. Yao and Y. Liu. Fast evolutionary programming. In L. J. Fogel, P. J. Angeline, and T. Bäck, editors, *Proceedings of the Fifth Annual Conference on Evolutionary Programming*, pages 451–460. The MIT Press, Cambridge, MA, 1996.