Performance Analysis of Simultaneous Perturbation Stochastic Approximation on the Noisy Sphere Model

Steffen Finck[a,*], Hans-Georg Beyer[a]

[a] *FH Vorarlberg University of Applied Sciences*

Abstract

To theoretically compare the behavior of different algorithms compatible performance measures are necessary. Thus, in the first part an analysis approach, developed for Evolution Strategies, was applied to Simultaneous Perturbation Stochastic Approximation on the noisy sphere model. A considerable advantage of this approach is that convergence results for non-noisy and noisy optimization can be obtained simultaneously. Next to the convergence rates, optimal step sizes and convergence criteria for 3 different noise models were derived. These results were validated by simulation experiments. Afterwards, the results were used for a comparison with Evolution Strategies on the sphere model in combination with the 3 noise models. It was shown that both strategies perform similarly, with a slight advantage for SPSA if optimal settings were used and the noise strength is not too large.

Keywords: algorithm comparison, stochastic gradient approximation, evolution strategy, noisy optimization

1. Introduction

In recent years noisy optimization became an important research topic, especially due to increased use of simulation optimization and the advances in computer hardware development. Therefore, an interesting aspect concerns the question as to what kind of strategies one should use for such optimization problems. To answer this question, one needs to compare these strategies. One way is to do this on a purely empirical level, as it was done in the recent Black Box Optimization Benchmarking (BBOB) at the Genetic and Evolutionary Computation Conference (GECCO) in 2009 and 2010.[1] However, there is also a desire to compare strategies on a deeper and more theoretical level. Given the diverse research fields concerned with noisy optimization (e.g. Operations Research, Engineering Optimization, Evolutionary Computation, Robust Optimization), the strategies developed were mainly analyzed with methods tailored to their specific fields. This may cause obstacles in the comparison across fields, since the derived results are not compatible and do not allow for a direct comparison. A solution is to use a unified approach

*Corresponding address: FH Vorarlberg University of Applied Sciences, Hochschulstrasse 1, 6850 Dornbirn, Austria; Phone: +43 5572 7927122; Fax: +43 5572 7929510

Email addresses: Steffen.Finck@fhv.at (Steffen Finck), Hans-Georg.Beyer@fhv.at (Hans-Georg Beyer)

[1] More details about this workshop can be found at http://coco.gforge.inria.fr/doku.php?id=start.

which results in the same performance measures which then can be used as basis for a comparison.

Such a unified approach is presented in this work for the analysis of Simultaneous Perturbation Stochastic Approximation (SPSA) [1, 2]. The approach itself was developed for Evolution Strategies [3] and will here be applied to a different type of strategy for the first time. The aim is to derive equations for the dynamic behavior, convergence criteria and optimal strategy parameter settings. We will show that the approach also provides insight in the short term dynamics which are usually not captured with common analysis methods for SPSA. See Appendix A for an overview of the proofs obtained in [1]. The presented analysis method will be applied to a restricted class of test functions. That is, simple test functions are to be considered which allow for mathematical tractability which in turn allows to derive conclusions not (always) available from other approaches (e.g. optimal parameter settings). While this might be considered as a too less general approach, we like to point out that the same approach was successfully applied to other test functions, e.g. the ridge function [4] or ellipsoidal-type functions [5, 6]. However, such analyses present a demanding task which in turn means that progress in this field proceeds gradually. That is why we will consider the sphere model test function, however, in combination with three different noise models:

- noise-free
- constant noise
- state-dependent noise

These models can be analyzed using the same analysis approach, which is not possible for SPSA with the method presented in [2], where an additional treatment of the noise-free case was necessary [7, 8]. Later on, we will compare the results obtained with the respective results from literature.

After introducing SPSA in Section 2, a detailed description of the steps for the theoretical approach will be given in Section 3. A peculiarity of the approach used is that one has to consider the test function in the limit of infinite search space dimensionality. However, in Section 4 it will be shown that the derived results will provide good approximations for finite search space dimensionalities as well. This will be done by simulation experiments. Afterwards, in Section 5 a comparison with Evolution Strategies is performed. There, the equations derived will be used to obtain performance measures. In Section 6 a summary of the work is given and conclusions from the results derived are drawn.

2. The Basic SPSA algorithm

This section reviews the basic SPSA algorithm. This algorithm belongs to the class of stochastic approximation algorithms [9], performing basically an approximate gradient descent. The pseudo code of SPSA is given in Alg. 1. In lines 1–3 the initial solution vector $\mathbf{x}^{(1)} \in \mathbb{R}^N$ is set and the strategy parameters are chosen. In SPSA the following strategy parameters[2] are used:

- $\alpha \in [0, 1]$ - reduction rate for the gradient step size factor $a^{(t)}$

[2]There exist SPSA variants which use more than these basic parameter. For examples see [2] and `www.jhuapl.edu/SPSA`. The web site also provides many examples for practical problems solved with SPSA.

Algorithm 1 Simultaneous Perturbation Stochastic Approximation

1: initialize $\mathbf{x}^{(1)}$
2: initialize $a^{(0)}$ and $c^{(0)}$
3: choose α, γ, and A
4: **for** $t := 1$ to $t_{\max}$ **do**
5: choose perturbation vector $\mathbf{\Delta}^{(t)}$
6: $c^{(t)} = c^{(0)} t^{-\gamma}$
7: $f_+^{(t)} = f\left(\mathbf{x}^{(t)} + c^{(t)}\mathbf{\Delta}^{(t)}\right)$
8: $f_-^{(t)} = f\left(\mathbf{x}^{(t)} - c^{(t)}\mathbf{\Delta}^{(t)}\right)$
9: $\mathbf{g}^{(t)} = \frac{f_+^{(t)} - f_-^{(t)}}{2c^{(t)}} \mathbf{\Delta}^{(t)^{-1}}$ ▷ $\mathbf{\Delta}^{-1} := \left(\Delta_1^{-1},\ \Delta_2^{-1},\ \ldots,\ \Delta_N^{-1}\right)^{\mathrm{T}}$
10: $a^{(t)} = a^{(0)}(t + A)^{-\alpha}$
11: $\mathbf{x}^{(t+1)} = \mathbf{x}^{(t)} - a^{(t)}\mathbf{g}^{(t)}$
12: check termination_criterion
13: **end for**

- $\gamma \in [0, 1]$ - reduction rate for the gradient approximation step size factor $c^{(t)}$
- $a^{(0)} > 0$ - initial value of the gradient step size factor
- $c^{(0)} > 0$ - initial value of gradient approximation step size factor
- $A \geq 0$ - stability factor

The core of SPSA is represented by the loop within lines 4–13. Defining $t_{\max}$ as maximal number of iterations, the loop is repeated until $t_{\max}$ or any other termination criterion defined in line 12 is satisfied. At the start of the loop the perturbation vector $\mathbf{\Delta}^{(t)}$ is chosen from a given random distribution. This distribution must satisfy the following properties [2]:

1. symmetry,
2. zero mean and finite variance,
3. finite inverse moments.

The components of the perturbation vector must be independent and identically distributed (iid). A common choice is the symmetric ± 1 Bernoulli distribution. This distribution generates ± 1, each with a probability of $p = 0.5$. Surveys [10, 11] showed that this distribution is well suited for most test functions considered. Therefore, this work will only consider this distribution for $\mathbf{\Delta}^{(t)}$. Next, the current gradient approximation step size factor $c^{(t)}$ is determined (line 6). As recommended in [2], $c^{(0)}$ should be set approximately equal to the noise at the initial point and $\gamma = 0.101$ being the smallest admissible value fulfilling the assumptions of Spall's convergence proof [2]. Afterwards, the gradient is approximated in line 9 by means of the function values at the points $\mathbf{x}^{(t)} \pm c^{(t)}\mathbf{\Delta}^{(t)}$ (line 7 and line 8). Note, $\mathbf{\Delta}^{-1}$ is defined as

$$\mathbf{\Delta}^{-1} := \left(\Delta_1^{-1},\ \Delta_2^{-1},\ \ldots,\ \Delta_N^{-1}\right)^{\mathrm{T}} \tag{1}$$

where $\Delta_1, \ldots, \Delta_N$ are the components of $\mathbf{\Delta}$. It is a remarkable property of SPSA that it needs only two function evaluations to approximate the gradient. This is in contrast to other methods relying on, e.g, $N + 1$ or $2N$ function evaluations using forward and central difference approximation

schemes, respectively (e.g. Implicit Filtering [12]). The update of the current solution is done in line 11, where the approximated gradient is multiplied by the gradient step size factor $a^{(t)}$. This factor depends on $a^{(0)}$, t, A, and α (see line 10). The recommendations for these parameters are: $\alpha = 0.602$ and $A \approx 0.1t_{\max}$. With these values and the desired minimal change in the magnitude of the components of $\mathbf{x}^{(1)}$ in the first iterations one can determine $a^{(0)}$ [2]. As for γ, the setting for α is equal to the smallest admissible value fulfilling the assumptions of the convergence proof. Choosing the smallest values is beneficial for practical applications with strong noise. Note, these recommendations are based on empirical investigations on several test functions. The theoretical asymptotic optimal values were determined as $\alpha = 1$ and $\gamma = 1/6$ in [1]. The interested reader is also referred to [2, Chapter 7].

The basic algorithm can be enhanced by using some kind of gradient smoothing and applying thresholds for the updates. See [2] for a discussion of these options. Another improvement is the use of adaptive SPSA [13, 14]. Where the Hessian matrix is also approximated (by at least 2 more function evaluations per iteration) and then it is used for the update of the solution vector. In this work we are only concerned with the basic algorithm, although a slight modification will be introduced shortly.

3. Analysis of the dynamical behavior

For a comparison of different algorithms one can use a benchmark suite (e.g, [15, Chapter 6] which especially considers noisy optimization and the one used in the BBOB 2009 workshop, see footnote 1), which gives information about the performance of the algorithms over a range of test functions. But there is still a need (and desire) to compare strategies on a theoretical level. This gives more insight about the behaviors of the algorithms. A first step was presented in [16] where five different methods (Random Search, SPSA, Evolution Strategies (ES), Genetic Algorithms, and Simulated Annealing) were compared. The comparison was based on the respective theoretical convergence rates for an unimodal and separable objective function. The restriction to this function class was necessary, since for other function classes the theoretical results were not comparable.

The approach pursued here is slightly different. Rather than using different formulations for the convergence rate, a unique formulation for all algorithms is considered. The approach was developed in [3] for the analysis of ES. It was successfully applied to different variants of ES and different test functions (e.g., sphere model, ridge, and quadratic functions). In the current paper the approach will be applied to a non-ES algorithm for the first time. To this end, we restrict ourselves to the sphere model (which is unimodal and separable) in combination with three different noise models. In the following a detailed step-by-step description of the analysis approach will be given. First an one-iteration performance measure will be derived. The result obtained will then be used to derive convergence criteria, optimal gradient step sequences, and equations representing the overall dynamic behavior. Afterwards in Section 5, a comparison of the results obtained with respective results for ES will be presented.

3.1. Deriving the Fitness Gain - A One-step Performance Measure

First, let us start with the definition of the sphere model

$$f(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{x}, \tag{2}$$

where $\mathbf{x} \in \mathbb{R}^N$ and $f : \mathbb{R}^N \to \mathbb{R}$. Since we are interested in noisy optimization, the function definition is generalized to

$$f(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{x} + \sigma_\epsilon(\mathbf{x})\mathcal{N}(0,1) \tag{3}$$

where the first term in (3) represents the true (non-noisy) function value and the second term represents a scalar noise term. The noise term consists of the noise strength $\sigma_\epsilon(\mathbf{x})$ and a standard normally distributed random scalar $\mathcal{N}(0,1)$, a common model in noisy optimization. The investigation of other noise distributions is beyond the scope of this paper. The three different noise models under consideration are: $\sigma_\epsilon(\mathbf{x}) = 0$ (noise-free model), $\sigma_\epsilon(\mathbf{x}) = \text{const.}$ (constant noise model), and $\sigma_\epsilon(\mathbf{x}) = f_{\text{noise}}(\mathbf{x})$ (state-dependent noise model) where the noise strength depends on the location and vanishes at the optimum. In the constant noise model, the variance of the noise will be constant. For the state-dependent noise model it is assumed that σ_ϵ will only depend on the current $\mathbf{x}^{(t)}$, i.e. $\sigma_\epsilon(\mathbf{x}^{(t)}) = \sigma_\epsilon(\mathbf{x}^{(t)} \pm c^{(t)}\mathbf{\Delta})$. With this simplification the math involved is much more amenable as if σ_ϵ would depend on the actual evaluated point. The same noise model was used for the analysis of ES [15] which allows for a comparison of both strategies. However, as shown in [15], for $N \to \infty$, a frequently used assumption in the derivation process, the behavior of both models is the same. Last but not least, for all models considered no correlation between different evaluations of the noise term is assumed, i.e. $\mathcal{N}(0,1)$ is iid.

The analysis starts by considering the gradient approximation in SPSA

$$\mathbf{g}^{(t)} = \frac{f_+^{(t)} - f_-^{(t)}}{2c^{(t)}}\mathbf{\Delta}^{(t)^{-1}} \tag{4}$$

where $f_\pm^{(t)}$ represent the evaluation of (3) at the points $\mathbf{x}^{(t)} \pm \mathbf{\Delta}^{(t)}$. Due to the noise in (3) and the manner in which the gradient is estimated in (4), the resulting $\mathbf{g}^{(t)}$ has only limited accuracy. After all it is an *approximation*. To improve the accuracy, one can use an average of multiple gradient approximations. This is achieved by adding a loop into Alg. 1, which encloses lines 5–9. Thus, each approximation has different $\mathbf{\Delta}^{(t)}$, but the same $c^{(t)}$. To differentiate between the different gradient approximations a subscript w is added. Applying this idea, the gradient approximation changes from (4) to

$$\mathbf{g}^{(t)} = \frac{1}{W}\sum_{w=1}^{W}\mathbf{g}_w^{(t)} = \frac{1}{W}\sum_{w=1}^{W}\frac{f_{w+}^{(t)} - f_{w-}^{(t)}}{2c^{(t)}}\mathbf{\Delta}_w^{(t)^{-1}} \tag{5}$$

where W is the number of gradient approximations. The function evaluations at the test points can be written with (3) as

$$\begin{aligned} f_w\left(\mathbf{x}^{(t)} \pm c^{(t)}\mathbf{\Delta}_w^{(t)}\right) &= \left(\mathbf{x}^{(t)} \pm c^{(t)}\mathbf{\Delta}_w^{(t)}\right)^{\mathrm{T}}\left(\mathbf{x}^{(t)} \pm c^{(t)}\mathbf{\Delta}_w^{(t)}\right) + \sigma_\epsilon^{\pm}\left(\mathbf{x}^{(t)}\right)\mathcal{N}_w(0,1) \\ &= \mathbf{x}^{(t)^{\mathrm{T}}}\mathbf{x}^{(t)} \pm 2c^{(t)}\mathbf{x}^{(t)^{\mathrm{T}}}\mathbf{\Delta}_w^{(t)} + c^{(t)^2}\mathbf{\Delta}_w^{(t)^{\mathrm{T}}}\mathbf{\Delta}_w^{(t)} + \sigma_\epsilon^{\pm}\left(\mathbf{x}^{(t)}\right)\mathcal{N}_w(0,1). \end{aligned} \tag{6}$$

Thus, the fitness difference in (5) can be expressed as

$$f_{w+}^{(t)} - f_{w-}^{(t)} = 4c^{(t)}\mathbf{x}^{(t)^{\mathrm{T}}}\mathbf{\Delta}_w^{(t)} + \tilde{\sigma}_\epsilon^{(t)}\mathcal{N}_w(0,1) \tag{7}$$

where $\tilde{\sigma}_\epsilon^{(t)}$ represents the difference in the noise factors and depends on the chosen noise model. Substituting (7) into (5) yields

$$\mathbf{g}^{(t)} = \frac{1}{W}\sum_{w=1}^{W}\left(2\mathbf{x}^{(t)^{\mathrm{T}}}\mathbf{\Delta}_w^{(t)} + \frac{\tilde{\sigma}_\epsilon^{(t)}\mathcal{N}_w(0,1)}{2c^{(t)}}\right)\mathbf{\Delta}_w^{(t)^{-1}}. \tag{8}$$

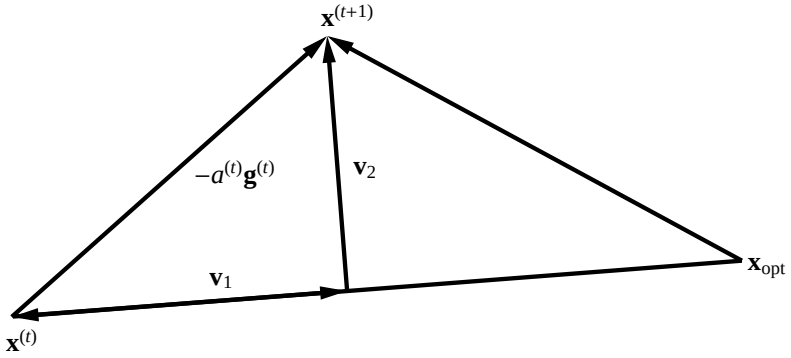

Figure 1: Decomposition of the update step $-a^{(t)}\mathbf{g}^{(t)}$ with the help of Pythagoras' theorem.

An analysis of above equation shows that for $\sigma_\epsilon^{(t)} = 0$, the value of $c^{(t)}$ has no influence on the gradient approximation. This is typical for SPSA on quadratic functions. If $\sigma_\epsilon^{(t)} > 0$, increasing $c^{(t)}$ will reduce the noisy disturbance.

The next step is to decompose the gradient step, $-a^{(t)}\mathbf{g}^{(t)}$, into a vector $\mathbf{v}_1$ which points in the direction of the optimum $\mathbf{x}_{\text{opt}}$ and a vector with perpendicular direction $\mathbf{v}_2$. This enables one to determine the achieved gain in the iteration step and the influence of the algorithm parameters on this gain. The decomposition is outlined in Fig. 1. The optimum is marked with $\mathbf{x}_{\text{opt}}$, the solution at the start of the iteration with $\mathbf{x}^{(t)}$, and the solution at the end of the iteration with $\mathbf{x}^{(t+1)}$. The gradient step from $\mathbf{x}^{(t)}$ to $\mathbf{x}^{(t+1)}$ is marked with $-a^{(t)}\mathbf{g}^{(t)}$. From the definition of the noisy sphere (3) it is clear that $\mathbf{x}_{\text{opt}} = \mathbf{0}$.[3] Writing $R = \|\mathbf{x}^{(t)}\|$ and $r = \|\mathbf{x}^{(t+1)}\|$ and using Pythagoras' theorem one obtains

$$\|\mathbf{v}_1\|^2 + \|\mathbf{v}_2\|^2 = \|a^{(t)}\mathbf{g}^{(t)}\|^2 \tag{9}$$

$$(R - \|\mathbf{v}_1\|)^2 + \|\mathbf{v}_2\|^2 = r^2. \tag{10}$$

Solving (9) and (10) yields the so-called evolution equation

$$r^2 = R^2 - 2R\|\mathbf{v}_1\| + \|a^{(t)}\mathbf{g}^{(t)}\|^2 \tag{11}$$

which describes the change in the distance to the optimum after a single iteration step. The unknown in (11) is the norm of vector $\mathbf{v}_1$, hence deriving an expression for $\|\mathbf{v}_1\|$ is the next step.

By means of the scalar product one obtains

$$\mathbf{v}_1 = -\frac{\mathbf{x}^{(t)^{\mathrm{T}}} a^{(t)}\mathbf{g}^{(t)}}{R^2}\mathbf{x}^{(t)}. \tag{12}$$

The minus in front of the fraction is due to $\mathbf{x}^{(t)}$ and $\mathbf{v}_1$ having anti-parallel directions. Recalling that $\|\mathbf{x}^{(t)}\| = R$, the norm of $\mathbf{v}_1$ yields

$$\|\mathbf{v}_1\| = \frac{|a^{(t)}\mathbf{x}^{(t)^{\mathrm{T}}}\mathbf{g}^{(t)}|}{R}. \tag{13}$$

[3] The obtained results will still hold if an additional translation is applied to (3).

Using (8) the scalar product in (13) can be written as

$$a^{(t)}\mathbf{x}^{(t)^{\mathrm{T}}}\mathbf{g}^{(t)} = a^{(t)}\mathbf{x}^{(t)^{\mathrm{T}}}\left(\frac{1}{W}\sum_{w=1}^{W}\left(2\mathbf{x}^{(t)^{\mathrm{T}}}\boldsymbol{\Delta}_w^{(t)} + \frac{\tilde{\sigma}_\epsilon^{(t)}\mathcal{N}_w(0,1)}{2c^{(t)}}\right)\boldsymbol{\Delta}_w^{(t)^{-1}}\right). \tag{14}$$

From now on, we will use that $\boldsymbol{\Delta}^{(t)}$ obeys a symmetric ± 1 Bernoulli distribution. Hence, the components of $\boldsymbol{\Delta}^{(t)}$ are ± 1 and according to (1) $\boldsymbol{\Delta}^{(t)^{-1}} = \boldsymbol{\Delta}^{(t)}$ is valid. Rewriting (14) yields

$$a^{(t)}\mathbf{x}^{(t)^{\mathrm{T}}}\mathbf{g}^{(t)} = \frac{a^{(t)}}{W}\sum_{w=1}^{W}\left(2\left(\mathbf{x}^{(t)^{\mathrm{T}}}\boldsymbol{\Delta}_w^{(t)}\right)^2 + \frac{\tilde{\sigma}_\epsilon^{(t)}\mathcal{N}_w(0,1)}{2c^{(t)}}\mathbf{x}^{(t)^{\mathrm{T}}}\boldsymbol{\Delta}_w^{(t)}\right). \tag{15}$$

So far the quantity defined by the right-hand side (rhs) of (15) is a random variable. A main idea of the analysis approach is to use expected values and to neglect the fluctuation, similar to an ordinary differential equation approach. This will yield asymptotic correct equations for $N \to \infty$. As a consequence, this requires validating the obtained results for finite N by simulation experiments as it will be done in the next section. The expectation of $\left(\mathbf{x}^{(t)^{\mathrm{T}}}\boldsymbol{\Delta}_w^{(t)}\right)^2$ is

$$\mathrm{E}\left[(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Delta})^2\,|\mathbf{x}\right] = \mathrm{E}\left[\left(\sum_{i=1}^{N}x_i\Delta_i\right)^2|x_i\right] = \sum_{i=1}^{N}x_i^2\mathrm{E}\left[\Delta_i^2\right] + \sum_{i=1}^{N}\sum_{j\neq i}x_ix_j\mathrm{E}\left[\Delta_i\Delta_j\right]. \tag{16}$$

Note, the iteration index t and gradient approximation index w were omitted for brevity. Since $\boldsymbol{\Delta}$ has i.i.d. components, $\Delta_i = \pm 1$, and $\mathrm{E}[\Delta_i] = 0$, the relations

$$\mathrm{E}\left[\Delta_i\Delta_j\right] = \mathrm{E}\left[\Delta_i\right]\mathrm{E}\left[\Delta_j\right] = 0 \quad \text{and} \quad \Delta_i^2 = 1 \tag{17}$$

are valid. Using (17), (16) can be written as

$$\mathrm{E}\left[(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Delta})^2\,|\mathbf{x}\right] = \sum_{i=1}^{N}x_i^2 = R^2. \tag{18}$$

Now substituting (18) into (15) and taking the expectation yields

$$a^{(t)}\mathrm{E}\left[\mathbf{x}^{(t)^{\mathrm{T}}}\mathbf{g}^{(t)}|\mathbf{x}\right] = \frac{a^{(t)}}{W}\sum_{w=1}^{W}\left(2R^2 + \frac{\tilde{\sigma}_\epsilon^{(t)}\mathcal{N}_w(0,1)}{2c^{(t)}}\mathbf{x}^{(t)^{\mathrm{T}}}\mathrm{E}\left[\boldsymbol{\Delta}_w^{(t)}\right]\right). \tag{19}$$

Recalling that $\mathrm{E}[\Delta_i] = 0$, the last term in (19) vanishes. Thus, the resulting expectation for the norm of $\mathbf{v}_1$ is

$$\mathrm{E}\left[\|\mathbf{v}_1\||R\right] = \frac{a^{(t)}}{R}\mathrm{E}\left[|\mathbf{x}^{(t)^{\mathrm{T}}}\mathbf{g}^{(t)}||R\right] = \frac{a^{(t)}}{W}\sum_{w=1}^{W}2R = 2a^{(t)}R. \tag{20}$$

For further analysis we would like to have the evolution equation (11) only dependent on R, the strategy parameters ($a^{(t)}$, $c^{(t)}$, W), and the function parameters N and σ_ϵ. Thus, the term $\|\mathbf{g}^{(t)}\|^2$ in (11) needs to be expressed with those parameters. With (8) and recalling $\boldsymbol{\Delta}^{-1} = \boldsymbol{\Delta}$ one obtains

$$\|\mathbf{g}^{(t)}\|^2 = \left\|\frac{1}{W}\sum_{w=1}^{W}\left(2\mathbf{x}^{(t)^{\mathrm{T}}}\boldsymbol{\Delta}_w^{(t)} + \frac{\tilde{\sigma}_\epsilon^{(t)}\mathcal{N}_w(0,1)}{2c^{(t)}}\right)\boldsymbol{\Delta}_w^{(t)}\right\|^2. \tag{21}$$

This is a random variable and again we are interested in its expectation. The derivation of $\mathrm{E}\left[\|\mathbf{g}^{(t)}\|^2\right]$ is rather technically involved and is given in detail in Appendix B. The result obtained is

$$\mathrm{E}\left[\|\mathbf{g}^{(t)}\|^2|R\right] = \frac{N}{W}\left(4R^2 + \frac{\tilde{\sigma}_\epsilon^{(t)^2}}{4c^{(t)^2}}\right) + 4R^2\left(1 - \frac{1}{W}\right). \tag{22}$$

Substituting (22) and (20) into (11) yields

$$\mathrm{E}\left[r^2|R\right] = R^2 - 4a^{(t)}R^2 + \frac{a^{(t)^2}N}{W}\left(4R^2 + \frac{\tilde{\sigma}_\epsilon^{(t)^2}}{4c^{(t)^2}}\right) + 4a^{(t)^2}R^2\left(1 - \frac{1}{W}\right). \tag{23}$$

With (23) it is possible to determine the expected gain by a single iteration step. Since (23) depends on R^2 and r^2, the non-noisy function values at $\mathbf{x}^{(t)}$ and $\mathbf{x}^{(t+1)}$, this performance measure will be called the *quality gain*. Defining the quality gain as

$$q^{(t)} = \mathrm{E}\left[R^2 - r^2|R\right], \tag{24}$$

one obtains with (23)

$$\boxed{q^{(t)} = 4a^{(t)}R^2\left(1 - \frac{a^{(t)}}{W}(N + W - 1)\right) - \frac{a^{(t)^2}N\tilde{\sigma}_\epsilon^{(t)^2}}{4Wc^{(t)^2}}}\ . \tag{25}$$

From above expression one obtains the necessary condition for convergence in expectation as $q^{(t)} > 0 \ \forall\ t > T_0$ where $T_0 \geq 0$ is constant. In the following we use (25) to derive convergence criteria and optimal gradient step sizes $a^{(t)}$ for the three noise models.

3.2. Convergence Criteria and Optimal Gradient Step Sizes

First, the noise-free model, $\sigma_\epsilon = 0$, will be considered. In this case the quality gain reads

$$q^{(t)} = 4a^{(t)}R^2\left(1 - \frac{a^{(t)}}{W}(N + W - 1)\right). \tag{26}$$

Convergence to the optimizer in expectation will be achieved if

$$4a^{(t)}R^2\left(1 - \frac{a^{(t)}}{W}(N + W - 1)\right) > 0 \tag{27}$$

necessarily holds. Given that $a^{(t)}$ and R^2 are positive scalars one obtains

$$a^{(t)} < \frac{W}{N + W - 1}. \tag{28}$$

Further, one can derive an optimal step size $a^{(t)}$ from (26) yielding the maximal change towards the optimizer. Requiring $\mathrm{d}q^{(t)}/\mathrm{d}a^{(t)} = 0$ yields

$$4R^2 - a^{(t)}\left(\frac{8R^2(N + W - 1)}{W}\right) = 0 \tag{29}$$

which can be solved for $a^{(t)}$, obtaining

$$a_{\mathrm{nf}}^{(t)} = \frac{W}{2(N+W-1)}. \tag{30}$$

The denotation $a_{\mathrm{nf}}^{(t)}$ stands for optimal $a^{(t)}$ in the noise-free case. As one can see, it does not depend on t or R, thus, it is constant ($\alpha = 0$, cf. line 10 in Alg. 1) throughout the optimization process. Note, this result is specific for the noise-free sphere model and can not be applied to other test function classes. Still, it allows for an insight in the algorithm's behavior and will be later used for the comparison with Evolution Strategies.

Next, the constant noise model, $\sigma_\epsilon = \text{const.}$, will be considered. In this case the substitution $\tilde{\sigma}_\epsilon^{(t)} = \sqrt{(\sigma_\epsilon^+)^2 + (\sigma_\epsilon^-)^2} = \sqrt{2}\sigma_\epsilon$ will be applied to the quality gain (25). Hence, the requirement for convergence reads

$$4a^{(t)}R^2\left(1 - \frac{a^{(t)}}{W}(N+W-1)\right) - \frac{a^{(t)^2}N\sigma_\epsilon^2}{2\lambda c^{(t)^2}} > 0. \tag{31}$$

The convergence criterion w.r.t. $a^{(t)}$ yields

$$a^{(t)} < \frac{W}{(N+W-1) + \frac{N}{8R^2}\left(\frac{\sigma_\epsilon}{c^{(t)}}\right)^2}. \tag{32}$$

Comparing (32) with the noise-free criterion (28), one can see that the upper limit of $a^{(t)}$ is smaller for the constant noise model. Furthermore, $a^{(t)}$ now depends on the current location R and iteration t. Moreover, if $R \to 0$ (convergence towards $\mathbf{x}_{\mathrm{opt}}$) $a^{(t)}$ must decrease. Additionally, one can derive convergence criteria w.r.t. R and σ_ϵ. These are

$$R^2 > \frac{a^{(t)}N}{8\left(W - a^{(t)}(N+W-1)\right)}\left(\frac{\sigma_\epsilon}{c^{(t)}}\right)^2, \tag{33}$$

$$\sigma_\epsilon < Rc^{(t)}\sqrt{\frac{8\left(W - a^{(t)}(N+W-1)\right)}{a^{(t)}N}}. \tag{34}$$

The first criterion (33) states that for a given set of $a^{(t)}$, $c^{(t)}$, W, and σ_ϵ, SPSA will converge until the distance to the optimizer is equal to the term on the rhs of (33). Thus the optimum will not be reached (for that given set), however, decreasing $a^{(t)}$ or increasing $c^{(t)}$ or W will further reduce the distance. Since, the rhs of (33) will appear frequently throughout the text we define

$$f_{\min}(a^{(t)}, c^{(t)}) = \frac{a^{(t)}N}{8\left(W - a^{(t)}(N+W-1)\right)}\left(\frac{\sigma_\epsilon}{c^{(t)}}\right)^2, \tag{35}$$

recalling that $f(\mathbf{x}^{(t)}) = R^2$. Note, so far $f(\mathbf{x})$ was used for the observable (noisy) function value, however, $f_{\min}$ represents a true (non-noisy) function value. The second criterion (34) gives an insight on how large the noise strength can be while SPSA is still able to converge. Note, the criteria (32)–(34) are not independent and all parameters must satisfy criterion (31).

Similar to the noise-free model, one can derive an optimal $a^{(t)}$. Performing the same steps as before yields

$$a_{\mathrm{cn}}^{(t)} = \frac{4WR^2}{\left(8R^2(N+W-1) + N\frac{\sigma_\epsilon^2}{c^{(t)^2}}\right)} \tag{36}$$

where $a_{\mathrm{cn}}^{(t)}$ is the optimal $a^{(t)}$ for the constant noise model. Comparing $a_{\mathrm{cn}}^{(t)}$ with $a_{\mathrm{nf}}^{(t)}$ reveals that $a_{\mathrm{cn}}^{(t)}$ depends on R and t. If $R^2 \gg \sigma_\epsilon^2/c^{(t)^2} \Longrightarrow a_{\mathrm{cn}}^{(t)} \approx a_{\mathrm{nf}}^{(t)}$, while for $R \to 0 \Longrightarrow a_{\mathrm{cn}}^{(t)} \to 0$.

Finally, let us give some comments on $c^{(t)}$ for the constant noise model, which also apply to a certain extent to the state-dependent noise model considered next. The common sequence for $c^{(t)}$ is given in line 6 of Alg. 1. It is a decreasing sequence with the constant $c^{(0)}$ being chosen approximately equal to the observed standard deviation of several function evaluations at the initial point $x^{(1)}$. It appears from the quality gain (25), that $c^{(t)}$ only influences the noise term and choosing $c^{(t)}$ large and constant is advantageous[4]. On the other hand, if the observed standard deviation at $x^{(1)}$ is sufficiently large, this will yield a choice of $c^{(t)}$ (especially for the state-dependent noise-model where the noise strength increases with the fitness) causes a reduced accuracy close to the optimizer due to numerical problems. Decreasing $c^{(t)}$ would increase the noise factor $\sigma_\epsilon/c^{(t)}$ and thus decreasing $q^{(t)}$ (25). From previous analyses of SPSA [1] it is known, however, that the bias of the gradient approximation for a general test function is $O\left(c^{(t)^2}\right)$ and hence a decreasing $c^{(t)}$-sequence is beneficial in the general case.

In the state-dependent noise model the noise strength σ_ϵ will depend on the underlying true function value. Such a relationship is for example observed in physical measurements where the observed errors are relative to the value of the measurement. Using σ_ϵ^* = const. and definition

$$\sigma_\epsilon^* = \sigma_\epsilon \frac{N}{2R^2} \tag{37}$$

yields $\sigma_\epsilon \propto R^2$. As stated in the introduction we assume $\sigma_\epsilon(\mathbf{x}) = \sigma_\epsilon(\mathbf{x} \pm c^{(t)}\mathbf{\Delta})$ for $N \to \infty$ for the state-dependent noise model.

Now substituting $\tilde{\sigma}_\epsilon^{(t)}$ in (25) with

$$\tilde{\sigma}_\epsilon^{(t)} = \sqrt{2}\sigma_\epsilon^* \frac{2R^2}{N} \tag{38}$$

yields the necessary condition for convergence

$$4a^{(t)}R^2\left(1 - \frac{a^{(t)}}{W}(N + W - 1)\right) - \frac{2a^{(t)^2}\sigma_\epsilon^{*2}R^4}{NWc^{(t)^2}} > 0. \tag{39}$$

As before, convergence criteria for $a^{(t)}$, R, and σ_ϵ^* will be determined next. Convergence w.r.t. the step size factor $a^{(t)}$ is achieved if

$$a^{(t)} < \frac{2NW}{2N(N + W - 1) + R^2\left(\frac{\sigma_\epsilon^*}{c^{(t)}}\right)^2} \tag{40}$$

holds. Similar to the constant noise model, $a^{(t)}$ depends on the current location. Assuming $c^{(t)}$ to be constant, the upper limit for $a^{(t)}$ increases towards the rhs of (28) if $R \to 0$. The criteria w.r.t. R and σ_ϵ^* read

$$R^2 < \frac{2N(W - a^{(t)}(N + W - 1))}{a^{(t)}\left(\frac{\sigma_\epsilon^*}{c^{(t)}}\right)^2}, \tag{41}$$

$$\sigma_\epsilon^* < \frac{c^{(t)}}{R}\sqrt{\frac{2N(W - a^{(t)}(N + W - 1))}{a^{(t)}}}. \tag{42}$$

[4]Increasing $c^{(t)}$ is only beneficial if $\mathbf{x}^{(t)} \pm c^{(t)}\mathbf{\Delta}^{(t)}$ remains inside the feasible domain. However, such problems will not be considered here.

Note the difference in the sign between (33) and (41). For the state-dependent noise model SPSA converges only if the initial distance to $\mathbf{x}_{\text{opt}}$ is smaller than the expression on the rhs of (41). A respective conclusion concerning the maximal admissible σ_ϵ^* can be drawn from (42). From (42) one can conclude that choosing $c^{(t)} \propto R$ is an alternate valid choice for this factor. Apart from that one is referred to the discussion on the choice of $c^{(t)}$ at the end the constant noise model analysis on page 10. As before, valid parameter sets must still satisfy (40).

Finally, an optimal setting for $a^{(t)}$ is determined by performing the same steps as before. One obtains

$$a_{\text{sn}}^{(t)} = \frac{W}{2(N+W-1) + \frac{R^2}{N}\left(\frac{\sigma_\epsilon^*}{c^{(t)}}\right)^2} \tag{43}$$

where $a_{\text{sn}}^{(t)}$ is the optimal $a^{(t)}$ for the state-dependent noise model. Comparing $a_{\text{sn}}^{(t)}$ with $a_{\text{nf}}^{(t)}$ shows that both are approximately the same for $R \to 0$. If R or $\sigma_\epsilon^*/c^{(t)}$ is large $a_{\text{sn}}^{(t)}$ tends towards 0.

3.3. Determining the Dynamics with the ODE Approach

To analyze the dynamic behavior for successive iterations, one could iterate (25) or use (25) as basis for a differential equation which describes the dynamics. In the following the latter method is considered. One starts by assuming

$$\frac{\mathrm{d}f}{\mathrm{d}t} \approx -q \tag{44}$$

where f represents the non-noisy function value. The restriction for this assumption is that the higher order derivatives of f w.r.t. t are small. For the step sequences $a^{(t)}$ and $c^{(t)}$ the expressions from Alg. 1

$$a^{(t)} = a^{(0)}(t+A)^{-\alpha}, \tag{45}$$

$$c^{(t)} = c^{(0)}t^{-\gamma} \tag{46}$$

will be used. This additionally allows to determine the influence of the reduction rates α and γ on the dynamics. Moreover, since we only consider the dynamics in the non-noisy fitness space, we will replace R^2 with f in the respective equations. Starting with the noise-free case, one obtains

$$f' + \left(4a^{(0)}(t+A)^{-\alpha} - \frac{4a^{(0)^2}(N+W-1)}{W}(t+A)^{-2\alpha}\right) f = 0 \tag{47}$$

where $f' = \mathrm{d}f/\mathrm{d}t$. Since we are mostly interested in the long-term behavior we assume $t + A \approx t$ and for $\alpha > 0$ that $t^{-\alpha} >> t^{-2\alpha}$ holds. Further note that $t \geq 1$ holds. See Appendix C for the detailed solution steps.

Equation (47) is a homogeneous differential equation, stated as an initial value problem. Using

$$f_{\text{start}} := f(\mathbf{x}^{(1)}) \tag{48}$$

and the solution ansatz $f = c\exp(-Z(t))$, where c is a constant and $Z(t)$ is the integral over the respective term inside the brackets in (47). With above assumptions the following solutions are obtained

$$f(\mathbf{x}^{(t)}) = \begin{cases} f_{\text{start}}\exp\left(\overline{q}(1-t)\right), & \text{for } \alpha = 0, \\ f_{\text{start}}\exp\left(\dfrac{4a^{(0)}}{1-\alpha}\left(1-t^{1-\alpha}\right)\right), & \text{for } 0 < \alpha < 1, \\ f_{\text{start}}t^{-4a^{(0)}}, & \text{for } \alpha = 1. \end{cases} \tag{49}$$

The term

$$\bar{q} = \frac{4a^{(0)}}{W}\left(W - a^{(0)}(N + W - 1)\right) \tag{50}$$

is reminiscent of the noise-free quality gain (25) with constant step size factor and normalized by R^2. The respective asymptotic behavior ($t \to \infty$) which yields

$$f(\mathbf{x}^{(t)}) \sim \begin{cases} \exp\left(-t^{1-\alpha}\right), & \text{for } 0 \leq \alpha < 1, \\ t^{-4a^{(0)}}, & \text{for } \alpha = 1. \end{cases} \tag{51}$$

From (51) one can deduce that the fastest convergence rate is obtained for $\alpha = 0$, i.e constant $a^{(t)}$. For $\alpha < 1$ one observes a log-linear convergence behavior, while for $\alpha = 1$ sublinear convergence is attained. The result obtained in [7] reads

$$\lim_{t\to\infty} \frac{1}{t} \log \|\mathbf{x}^{(t)}\| = \beta \; a.s. \tag{52}$$

where $\beta < 0$ is a small constant depending on $a^{(t)}$. Further the proof assumes $c^{(t)}$ and $a^{(t)}$ to be constant, f three-times continuously differentiable w.r.t. $\mathbf{x}$ with bounded derivatives up to order three in any bounded domain, and $\mathbf{x}_{\text{opt}} = 0$ to be unique. For more details and an extension of the proof see [7, 8]. For $\alpha = 0$ (51) and (52) both predict a log-linear convergence behavior. In [14] a convergence for noise-free quadratic function is presented for SPSA with additional Hessian matrix adaptation. There the fastest convergence rate for the expected error of the trace in the Hessian matrix is $\sim \exp\left(-t^{1/2}\right)$ which is constrained by the parameters for the Hessian approximation.

In the constant noise case, the differential equation reads

$$f' + \left(4a^{(0)}t^{-\alpha} - \frac{4a^{(0)^2}(N + W - 1)}{W}t^{-2\alpha}\right) f = \frac{a^{(0)^2} N\sigma_\epsilon^2}{2Wc^{(0)^2}} t^{2(\gamma-\alpha)}, \tag{53}$$

where $t + A \approx t$ was used. To solve this inhomogeneous differential equation a particular solution will be added to the solution of the homogeneous equation (47). Using variation of constants, the following integral is obtained[5]

$$\int c'(t)\mathrm{d}t = \frac{a^{(0)^2} N\sigma_\epsilon^2}{2Wc^{(0)^2}} \int t^{2(\gamma-\alpha)} \exp\left(\frac{4a^{(0)}}{1-\alpha}t^{1-\alpha} - \frac{4a^{(0)^2}(N + W - 1)}{W(1-2\alpha)}t^{1-2\alpha}\right)\mathrm{d}t. \tag{54}$$

Unfortunately, this integral has a closed-form solution only for some special cases. One case of interest is $\alpha = 0$ and $\gamma = 0$, which represents constant step sizes factors $a^{(t)}$ and $c^{(t)}$. Using the initial condition (48) and adding the respective homogeneous solution (49) to the obtained particular solution yields

$$f(\mathbf{x}^{(t)})_{\alpha=0,\gamma=0} = f_{\min}(a^{(0)}, c^{(0)}) + \left(f_{\text{start}} - f_{\min}(a^{(0)}, c^{(0)})\right) \exp\left(\bar{q}(1-t)\right). \tag{55}$$

where $f_{\min}(a^{(0)}, c^{(0)})$ is defined by (35). The asymptotic behavior ($t \to \infty$) of (55) reads

$$\lim_{t\to\infty} f(\mathbf{x}^{(t)}) = f_{\min}(a^{(0)}, c^{(0)}) = \frac{a^{(0)} N}{8(W - a^{(0)}(N + W - 1))} \left(\frac{\sigma_\epsilon}{c^{(0)}}\right)^2. \tag{56}$$

[5] For brevity some intermediate steps are not shown. See Appendix C for detailed solution steps.

Equation (56) shows that $\mathbf{x}_{\text{opt}}$ can not be reached if SPSA with constant gradient step size factor $a^{(t)}$ is used.

Next, the case with $\alpha = 1$ is investigated, which represents SPSA with a fast decreasing gradient step size. Performing the same steps as above yields

$$f(\mathbf{x}^{(t)})_{\alpha=1} = \frac{a^{(0)^2} N \sigma_\epsilon^2}{2Wc^{(0)^2}\left(2\gamma - 1 + 4a^{(0)}\right)} \left(t^{2\gamma-1} - t^{-4a^{(0)}}\right) + f_{\text{start}} t^{-4a^{(0)}}, \tag{57}$$

where $t^{-\alpha} \gg t^{-2\alpha}$ was used. Due to $a^{(0)} \propto 1/N$ (see (32)) the asymptotic convergence rate for $N \to \infty$ can be written as

$$f(\mathbf{x}^{(t)}) \sim t^{2\gamma-1} \quad \text{for} \quad t \to \infty. \tag{58}$$

The result obtained by Spall [2] reads

$$t^{\frac{\beta}{2}}(\mathbf{x}^{(t)} - \mathbf{x}_{\text{opt}}) \xrightarrow{dist.} \mathcal{N}(\mu, \mathbf{\Sigma}) \quad \text{as } t \to \infty, \tag{59}$$

under the conditions given in Appendix A. Further, μ and $\mathbf{\Sigma}$ are mean vector and covariance matrix of the attained normal distribution and $\beta = \alpha - 2\gamma$, which in the considered case equates to $\beta = 1 - 2\gamma$. Since one of the requirements for the proof is

$$3\gamma - \frac{\alpha}{2} \geq 0 \tag{60}$$

the maximal β is $\beta = \frac{2}{3}$ with $\gamma = \frac{1}{6}$. Details of the proof can be found in [1, 2]. Noting that (58) is stated in terms of $f(\mathbf{x}^{(t)})$ and (59) in terms of $\mathbf{x}$ both state the same convergence rate.

Finally, the state-dependent noise model will be considered. Using the quality gain formulation (39) where the normalized noise strength σ_ϵ^* (recall σ_ϵ^* is constant during the optimization process) is used, the resulting differential equation reads

$$f' + 4a^{(0)}t^{-\alpha}\left(1 - \frac{a^{(0)}t^{-\alpha}}{W}(N + W - 1)\right) f - \frac{2a^{(0)^2}\sigma_\epsilon^{*2}}{NWc^{(0)^2}} t^{2(\gamma-\alpha)} f^2 = 0. \tag{61}$$

This differential equation is a first-order non-linear differential equation. However, (61) is a Bernoulli differential equation which can be transformed into a linear differential equation. Using the substitutions

$$u = f^{-1} \text{ and } u' = -f^{-2} f', \tag{62}$$

one obtains

$$u' - 4a^{(0)}t^{-\alpha}\left(1 - \frac{a^{(0)}t^{-\alpha}}{W}(N + W - 1)\right) u = -\frac{2a^{(0)^2}\sigma_\epsilon^{*2}}{NWc^{(0)^2}} t^{2(\gamma-\alpha)}. \tag{63}$$

This equation is of the same type as the inhomogeneous differential equation for the constant noise case (53). Hence, the same solution steps can be performed and the same restrictions (closed-form solution only for special cases) apply. As done for the constant noise model, the setting $\alpha = 0$ and $\gamma = 0$ is considered first. Performing the appropriate steps yields

$$f(\mathbf{x}^{(t)})_{\alpha=0,\gamma=0} = \frac{2f_{\text{start}} N c^{(0)^2}\left(a^{(0)}(N + W - 1) - W\right)}{-f_{\text{start}} a^{(0)} \sigma_\epsilon^{*2} + \left(f_{\text{start}} a^{(0)} \sigma_\epsilon^{*2} + \frac{2}{Nc^{(0)^2}}\left(a^{(0)}(N + W - 1) - W\right)\right) \exp\left(-\bar{q}(1 - t)\right)}. \tag{64}$$

The asymptotic behavior of (64) is

$$f(\mathbf{x}^{(t)}) \sim \exp(-t) \quad \text{for } t \to \infty. \tag{65}$$

This is the same asymptotic rate as for the noise-free scenario with $\alpha = 0$. In Spall's proof, no differentiation between the constant noise and state-dependent noise was made, hence the same result (59) applies. The second case under consideration is the one with $\alpha = 1$. The solution for the dynamics in this case reads

$$f(\mathbf{x}^{(t)})_{\alpha=1,A=0} = \frac{NWc^{(0)^2} f_{\text{start}} \left(2\gamma - 4a^{(0)} - 1\right)}{-2a^{(0)^2}\sigma_\epsilon^{*2} f_{\text{start}} \left(t^{2\gamma-1} - t^{4a^{(0)}}\right) + NWc^{(0)^2} \left(2\gamma - 1 - 4a^{(0)}\right) t^{4a^{(0)}}}. \tag{66}$$

The asymptotic analysis yields

$$f(\mathbf{x}^{(t)}) \sim t^{-4a^{(0)}} \quad \text{for } t \to \infty, \tag{67}$$

i.e, the same rate as for $\alpha = 1$ and the noise-free model. Note, the exponent $2\gamma - 1$ is negative.

3.4. Summary

This section presented the detailed steps of a theoretical analysis approach developed for Evolution Strategies and its application to SPSA. The function under consideration was the sphere model in combination with three different noise models. First the quality gain, a performance measure for the one-iteration gain for the non-noisy function values, was derived. Using the derived equations, convergence criteria and optimal gradient step sizes were determined. Afterwards, an ordinary differential equation approach, based on the quality gain equations, was used to derive the overall dynamics. The results obtained were then compared with previous results from literature. A core assumption of the presented approach is the neglect of the stochastic fluctuations. Therefore, the derived equations are asymptotically correct for $N \to \infty$. To validate the equations for finite N, simulations will be performed and compared with the equations. This is the topic of the next section.

4. Experimental Analysis

In this section the results derived from the previous section will be compared with simulation experiments. The aim is to show the quality of the theoretical equations for finite N. Additionally, parameter studies will be performed to gain insight on the influence of the strategy parameter. These studies will yields insight in the general relation between the parameter and the performance of SPSA. First, the experimental settings will be described. The basic settings for the noise-free and constant noise model analysis were:

- The components of the start point were chosen from the $\mathcal{N}(100, 25)$ normal distribution for each sample anew.
- 10 samples were performed for each setting.
- The maximal number of function evaluations was set to $FEs_{\max} = 10^4 N$.
- The run was terminated when $f_{\text{target}} = f(\mathbf{x}_{\text{opt}}) + 10^{-20}$ was reached.

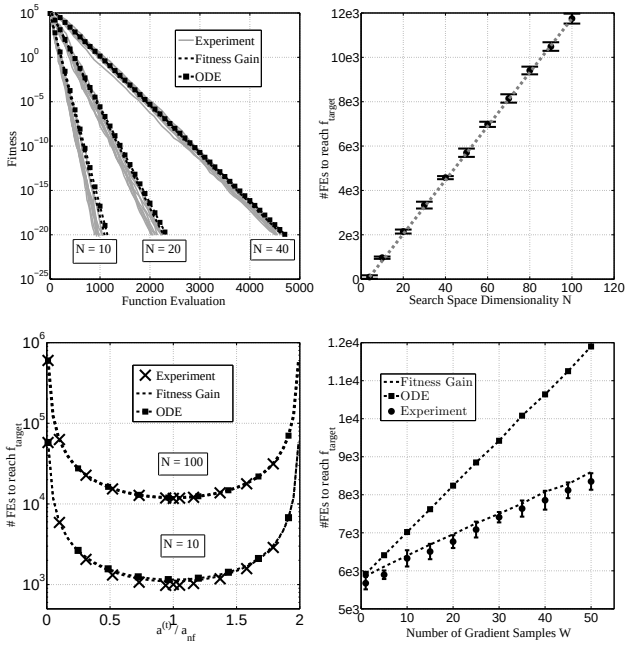

Figure 2: Results of the simulation experiments for the noise-free case. *Top left*: Sample runs and theoretical predictions for different search space dimensionalities. *Top right*: Influence of the search space dimensionality N on the number of function evaluations necessary to reach f_{target}. *Bottom left*: Influence of $a^{(t)}$ to reach f_{target} in terms of necessary function evaluations. All sequences of $a^{(t)}$ considered are constant, i.e. , $a^{(t)} = a^{(0)}\ \forall t$. *Bottom right*: Number of function evaluations to reach f_{target} for different number of gradient samples W for $N = 50$.

- The default strategy parameters were: $c^{(0)} = 1$, $\gamma = 0$, $\alpha = 0$, $A = 0$, $W = 1$, and $a^{(t)} = \frac{\lambda}{2(N+\lambda-1)} = a_{\text{nf}}^{(t)}$.
- The default value for the noise strength was $\sigma_\epsilon = 1$.

For the state-dependent noise, a slightly different setup has been used.

The analysis is performed for the noise-free model first. In the top left-hand plot of Fig. 2 the dynamic behavior of 10 sample runs for $N = 10, 20, 40$ is shown. One can clearly see the predicted log-linear convergence behavior of SPSA. Also the theoretical predictions based on the iteration of the quality gain (25) and the solution to the homogeneous ordinary differential equation (47) are shown. The theory predicts in both cases a slightly worse performance w.r.t. the number of function evaluations necessary. In the top right-hand plot of Fig. 2 the influence of the search space dimensionality N on the dynamic behavior is shown. From the curve it appears that there is a linear relation between N and the number of function evaluations for a given value of f_{target}. In the bottom left-hand plot of Fig. 2 the influence of $a^{(t)}$ on the dynamics is shown. All sequences considered of $a^{(t)}$ are constant w.r.t. t. Note the scaling of the horizontal axis for $a^{(t)}$ by $1/a_{\text{nf}}^{(t)}$. As one can see, the actual choice of $a^{(t)}$ is rather uncritical for the performance, as long as $a^{(t)}$ is in the range $0.5a_{\text{nf}}^{(t)} \ldots 1.5a_{\text{nf}}^{(t)}$. For non-constant sequences of $a^{(t)}$, one can conclude that performance will be poor if $a^{(t)}$ will be outside this range. As to the influence of gradient samples per iteration W, shown in the bottom right-hand plot in Fig. 2, increasing the W always

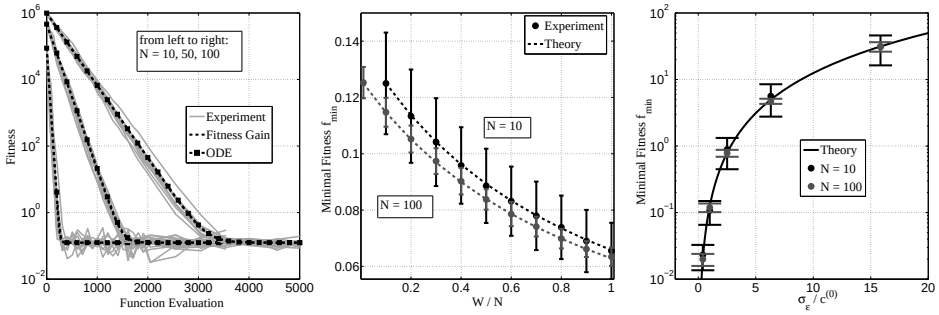

Figure 3: Results of the simulation experiments for SPSA on the sphere model with constant noise ($\sigma_\epsilon = 1$). All results shown were obtained with $a^{(t)} = a^{(t)}_{\text{nf}}$, $c^{(t)} = \sigma_\epsilon$, and $W = 1$ if not stated otherwise. *Left*: Dynamics of sample runs and mean value dynamics for $N = 10, 50, 100$ (from left to right). *Center*: The minimal fitness $f_{\min}$ as function of W/N. The theoretical predictions are based on Equation (35). *Right*: Influence of the noise as $\sigma_\epsilon / c^{(0)}$ on $f_{\min}$.

increases the number of function evaluations to reach f_{target}. Given that W can be interpreted as a form of resampling, the results discourage the use of it for the noise-free sphere model. More interesting is the fact, that the approximation quality of (25) is not reduced by increasing W, while the results of the ODE approach (49) deviate strongly for large W. The reason is that the gradient step is increasing with W (30) and thus the granularity can not be accurately represented by the ODE approach. However, the ODE approach still can be used as an approximation for the lower bound of the performance.

Next, the noise model with $\sigma_\epsilon = \text{const.}$ is considered. First, in the left-hand plot of Fig. 3 the dynamic behavior of SPSA with step size factor sequence $a^{(t)}_{\text{nf}}$ and $\sigma_\epsilon = 1$ is shown. Initially the same behavior as for the noise-free case is observed (see top left-hand plot of Fig. 2), until the noisy influence is not negligible anymore and SPSA finally stagnates. As for the noise-free model, theory and simulation results agree very well and the predicted dynamics appear closer to the observed mean value dynamics than for the noise-free case. The (mean) fitness value were stagnation occurs is defined by $f_{\min}$ (35). In center plot of Fig. 3 the influence of the number of gradient samples per iteration W is shown for search space dimensionalities $N = 10, 100$. Increasing W yields decreasing values of $f_{\min}$, albeit at the cost of more function evaluations per iteration step. The influence of the noise strength on $f_{\min}$ is shown in the bottom right-hand plot of Fig. 3. Instead of using the noise strength σ_ϵ as main parameter, $\sigma_\epsilon / c^{(0)}$ is used. This reflects the situation where one does not exactly know the value of σ_ϵ and thus must estimate $c^{(0)}$ (which should be chosen equal to σ_ϵ according to [2]). For the sphere it makes no differences if either σ_ϵ is increased or $c^{(0)}$ is decreased. Again, the results of the simulation experiments and the theoretical prediction by (35) agree well.

So far only constant sequences of $a^{(t)}$ were considered. To improve the performance w.r.t. $f_{\min}$, SPSA with a decreasing factor $a^{(t)}$ is analyzed next. The theoretical results (57) and (58) predict that SPSA should converge to $\mathbf{x}_{\text{opt}}$, as $t \to \infty$. In the right-hand plot of Fig. 4 the dynamic behavior for different values of α is shown. One can observe that $\alpha > 0$ results in a continuously decreasing non-noisy function value, but one also observes a simultaneous decrease in the convergence rate. Since all the curves were obtained with $W = 1$, one can conclude that using $\alpha > 0$ has a more pronounced effect on decreasing $f_{\min}$ than increasing W. Additionally, the dynamic behavior for $a^{(t)}_{\text{cn}}$ (36) is shown. It outperforms all other variants in terms of convergence

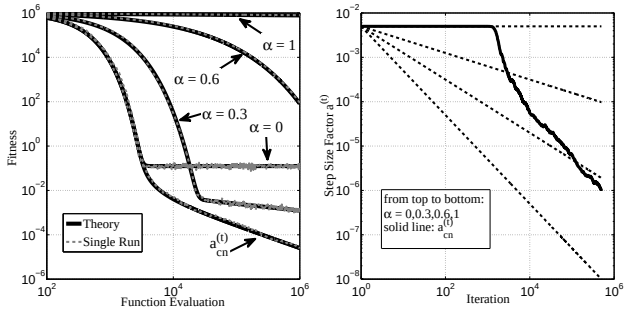

Figure 4: Influence of α for the sphere model with constant noise and search space dimensionality $N = 100$.. *Left*: Sample run dynamics and mean value dynamics for different sequence of $a^{(t)}$. *Right*: Sequence of $a^{(t)}$ for $\alpha = 0, 0.3, 0.6, 1$ and $a^{(t)} = a_{\text{cn}}^{(t)}$.

rate and obtains the lowest f_{min} within the given budget of function evaluations. The reason for this is shown in the left-hand plot of Fig. 4 where the history of $a_{\text{cn}}^{(t)}$ is shown. As long as the influence of the noise is negligible the strategy uses $a_{\text{nf}}^{(t)}$. As soon as the noise has a noticeable influence $a^{(t)}$ will be decreased. For comparison the curves for $a^{(t)}$ with different values of α are shown. This result suggest that an improvement of the performance can be achieved if SPSA is operated with non-constant α-values. The development of such a sequence is beyond the scope of the presented work and also should be further based on the performance on different objective functions.

Finally, the state-dependent noise model will be investigated. Contrary to the previous analyses, looking at the dynamic behavior reveals no new information, cf. left-hand plot of Fig. 6. If SPSA is able to reach $\mathbf{x}_{\text{opt}}$, the dynamic curves are similar to the noise-free case behavior in Fig. 2. Further, the quality of the agreement between the theoretical predictions and the simulations is the same. On the other hand, if SPSA diverges one only observes the diverging behavior without gaining any insight. However, of particular interest is the question as to when SPSA does diverge (e.g. for which parameter setting). Since our theoretical analysis is based on a mean value approach, it only can predict either diverging or converging behavior. To gain more insight we define the success probability

$$p_{\text{succ}} = \frac{\#\text{ samples where } f_{\text{target}} \text{ was reached}}{\#\text{ all samples}}. \tag{68}$$

This allows to track settings where some samples reach f_{target} and some do not. This requires a change in the experimental settings to account for this behavior. The new experimental setup is:

- For each set of parameter 100 samples were performed.
- The termination criteria were $f_{\text{target}} = 10^{-20}$ or a maximal number of function evaluations of $10^6 N$.
- The default parameters were: $W = 1$, $\sigma_\epsilon^* = 1$, $a^{(t)} = a_{\text{nf}}^{(t)}$, $f_{\text{start}} = 3N^2$, $\gamma = 0$, and $c^{(t)} = \sigma_\epsilon^*$.

The choice of $a^{(t)}$ was made with the intention to show how the state-dependent noise model influences the behavior of SPSA. On the other hand, the choice of $c^{(t)}$ is somewhat artificial and will be discussed later.

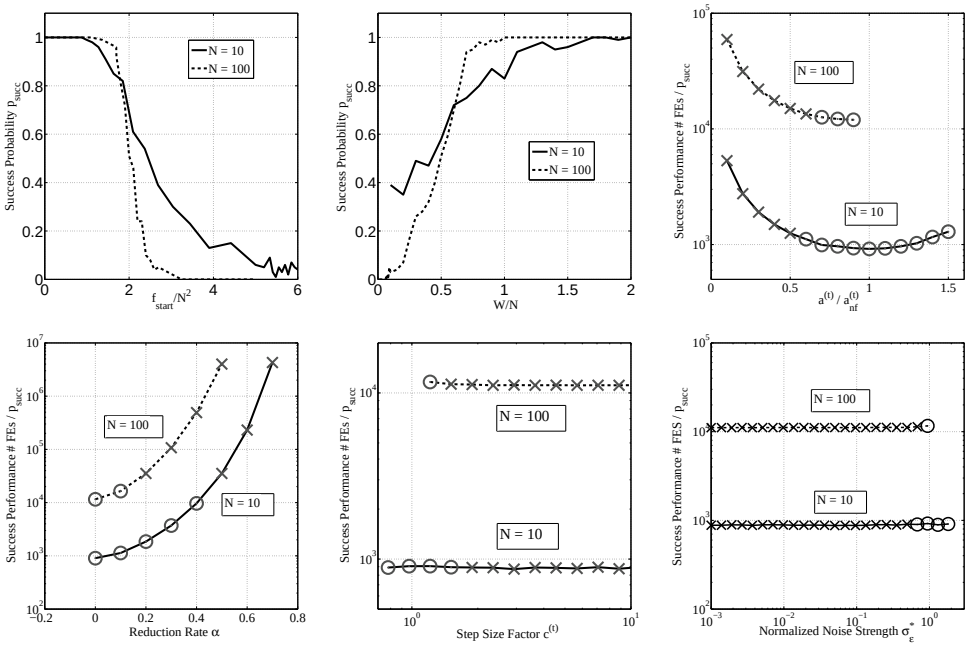

Figure 5: Experimental analysis of SPSA on the sphere with state-dependent noise. See text for the standard parameter settings. *Top left:* Success probability p_{succ} to reach $f_{\text{target}} = 10^{-20}$ for different initial states. *Top middle:* Success probability p_{succ} to reach $f_{\text{target}} = 10^{-20}$ as function of the number of gradient samples W. *Top right:* Success performance $\#FEs/p_{\text{succ}}$ for different constant $a^{(t)}$ sequences. Circles indicate $p_{\text{succ}} < 1$ and crosses indicate $p_{\text{succ}} = 1$. *Bottom left:* Success performance $\#FEs/p_{\text{succ}}$ as function of the reduction rate α for $a^{(t)}$-sequences defined by Eq. (45) with $a^{(0)} = a^{(t)}_{\text{nf}}$. *Bottom middle:* Success performance $\#FEs/p_{\text{succ}}$ for different constant $c^{(0)}$-sequences. *Bottom right:* Success performance $\#FEs/p_{\text{succ}}$ for different normalized noise strengths σ^*_ϵ.

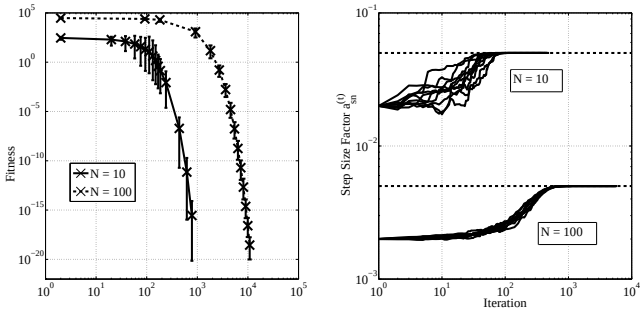

Figure 6: *Left*: Dynamical behavior of SPSA with optimal gradient step size factor $a^{(t)}_{\text{sn}}$ on the sphere model with state-dependent noise $\sigma^*_\epsilon = 1$ for different search space dimensionalities. *Right*: History of $a^{(t)}$ by using $a^{(t)}_{\text{sn}}$ (solid lines) and $a^{(t)}_{\text{nf}}$ as defined by Equation (30) (dashed line). Note the latter is constant and agrees with a sequence with setting $\alpha = 0$. For other choices of α see right-hand plot of Fig. 4.

From the convergence criterion of R^2 for the state-dependent noise model (41), one can infer that f_{start} must be smaller than a certain value in order to achieve convergence. This is investigated firstly and the results are shown in the top left-hand plot of Fig. 5. Using above parameter settings one can derive from (41)

$$f_{\text{start}} < 2N^2 \tag{69}$$

as necessary convergence condition. For $N = 10$ there are runs with $p_{\text{succ}} > 0$ for $f_{\text{start}} > 2N^2$, however, increasing N reveals a sharp drop in p_{succ} in the vicinity of $f_{\text{start}} = 2N^2$. One can speculate, that for $N \to \infty$ a jump in p_{succ} at $f_{\text{start}}/N^2 = 2$ from 1 to 0 will appear. The middle plot in the top row of Fig. 5 shows the influence of the number of gradient samples per iteration W. Increasing W increases p_{succ}, albeit with a simultaneous increase in the number of function evaluations. In the top right-hand plot of Fig. 5 the influence of using different constant gradient step sizes is shown. The results are shown in terms of $\#FEs/p_{\text{succ}}$, where $\#FEs$ is the mean of the numbers of function evaluations to reach $f_{\text{target}} = 10^{-20}$. This measure was introduced in [17] and represents an estimation of the success performance, i.e., the number of function evaluations necessary to reach a given target value. It accounts for sample run being unsuccessful, meaning the target function value was not achieved. In the remaining plots of Fig. 5 circles indicate runs where $p_{\text{succ}} < 1$ and crosses indicate runs with $p_{\text{succ}} = 1$. One can observe, that a small $a^{(0)}$ is necessary to reach $p_{\text{succ}} = 1$, which goes hand in hand with a slow convergence rate. The best convergence rate is reached close to $a^{(t)}_{\text{nf}}$, however, with $p_{\text{succ}} < 1$. Using decreasing gradient step sizes as defined by (45) with $\alpha > 0$ can improve the success rate as shown in the bottom left-hand plot of Fig. 5. However, the performance in terms of function evaluations is considerably reduced. The values for α not shown in the plots indicate runs where f_{target} was not reached within the budget of function evaluations for all samples. Finally, the influences of σ^*_ϵ and $c^{(t)}$ are shown in the bottom middle and bottom right plots of Fig. 5. While the measure $\#FEs/p_{\text{succ}}$ remains constant for all values of σ^*_ϵ or $c^{(t)}$, one observes a drop in the success probability for large σ^*_ϵ and small $c^{(t)}$ respectively. The curves suggest choosing $c^{(t)}$ large is beneficial since it only influences the noise term. However, as stated before this might results in a reduced accuracy of the gradient estimation close to the optimizer due to $\mathbf{x}^{(t)} \pm c^{(t)}\mathbf{\Delta} \approx \pm c^{(t)}\mathbf{\Delta}$. Therefore, using a decreasing sequence is advisable. Such a sequence could be proportional to R which can be approximated from the function value or to σ_ϵ which can be obtained by measuring the standard deviation of several function evaluations during the run. Of course, the sequence (46) commonly used is also valid, however, it does not use any information obtained during the run of SPSA.

Using $a^{(t)}_{\text{sn}}$ (43) improves the performance considerably as shown in Fig. 6. The success probability is always $p_{\text{succ}} = 1$ and the number of function evaluations is close to the best values for the constant gradient step sizes (where $p_{\text{succ}} < 1$), cf. top right-hand plot of Fig. 5. In the right-hand plot of Fig. 6 the history of $a^{(t)}_{\text{sn}}$ is plotted. One can observe that $a^{(t)}$ increases as predicted by the theoretical analysis (43). This is contrary to the requirement $\lim_{k\to\infty} a^{(t)} \to 0$ which is used in the analysis of SPSA by Spall et al. The explanation is, that at the initial point (away from the optimum) the noise is large and therefore small $a^{(t)}$ are necessary (same as for the constant noise model). Converging toward the optimum, the noise decreases until it is negligible. Hence, $a^{(t)}$ should be converging towards $a^{(t)}_{\text{nf}}$. Again, this behavior suggests that SPSA can be improved by some (adaptive) rule for α which uses information obtained during the run rather than being pre-determined. Such rule must be able to decrease and increase α depending on the underlying model.

5. Comparison with Evolution Strategies

In this section SPSA will be compared with Evolution Strategies (ESs) [18]. At first, the concept of the ES will be introduced shortly. Later, comparisons for each noise model will be performed based on performance criteria derived from the previous theoretical analysis.

ESs are nature-inspired strategies for optimization, which use a simplified model of Darwin's evolution paradigm. For an introduction into ES the reader is referred to [19]. Starting from an initial solution $\mathbf{x}$, *mutation* is used to generate a population of λ offspring. In the ES variants considered here, the probability distribution for the mutation obeys a normal distribution with variance σ^2, where σ is the so-called mutation strength . The offspring are evaluated and *selection* is performed, where the μ offspring with the best function value(s) (smallest in the case of minimization) are selected. The μ ($\mu < \lambda$) selected offspring, also referred to as parents, are then used for *recombination* to create the new solution, which for $\mu > 1$ equals the centroid of the selected offspring. Given the variety of ES variants we will consider two basic variants in this section only. First, for the noise-free model the (1+1)-ES is used. This strategy generates 1 offspring in each generation. The selection process compares the function value of the offspring with the function value of the current search point. In the case that the offspring function value is better, the offspring will be the new solution, else the parental point is kept. A pseudo code of this variant is shown in Alg. 2. For the constant noise and the state-dependent noise model, the $(\mu/\mu_I, \lambda)$-ES is used. This variant generates λ offspring from which the μ best will be selected (the parental solution is always discarded). By averaging these μ offspring, the new parental centroid will be created. The respective pseudo code is shown in Alg. 3.

Algorithm 2 The (1+1)-Evolution Strategy

1: initialize $\mathbf{x}^{(1)}$ and mutation strength σ
2: $g := 1$
3: **repeat**
4: $\quad \mathbf{y} = \mathbf{x}_g + \sigma\mathcal{N}(\mathbf{0}, \mathbf{I})$
5: $\quad$ **if** $f(\mathbf{y}) < f(\mathbf{x}_g)$ **then**
6: $\quad\quad \mathbf{x}_{g+1} \leftarrow \mathbf{y}$
7: $\quad$ **else**
8: $\quad\quad \mathbf{x}_{g+1} \leftarrow \mathbf{x}_g$
9: $\quad$ **end if**
10: $\quad \sigma \leftarrow AF(\sigma)$ ▷ Adaptation of σ
11: $\quad g \leftarrow g + 1$
12: **until** any termination criterion is fulfilled

In most variants of ES, an additional adaptation procedure for the mutation strength σ is needed. There exist different variants for this procedure, ranging from the $1/5th$ rule [18] over self-adaptation procedures [20] to derandomized adaptation procedures [21, 22][6]. However, in this work we will not consider the influence of the adaptation procedure.

5.1. The Noise-Free Sphere

In this section a comparison of SPSA and ES on the noise-free sphere is performed. In detail, we will compare SPSA with constant gradient step $a_{\text{nf}}^{(t)}$ with the (1+1)-ES. From [3] it is known,

[6] For a more comprehensive overview see also [23, 24, 19, 25]

Algorithm 3 The $(\mu/\mu_I, \lambda)$-Evolution Strategy

initialize $\mathbf{x}$ and σ
set strategy parameter μ and λ ▷ usually $\mu \approx \frac{\lambda}{4} \ldots \frac{\lambda}{2}$
repeat
  for $l = 1$ to λ **do** ▷ create offspring
    $\mathbf{y}_l = \mathbf{x}_g + \sigma\mathcal{N}(\mathbf{0}, \mathbf{I})$
    $f_l = f(\mathbf{y}_l)$
  end for
  $\tilde{\mathbf{f}} \leftarrow \text{sort}(f_1, \ldots, f_l)$ ▷ selection
  $\mathbf{x}_{g+1} \leftarrow \frac{1}{\mu}\sum_{i=1}^{\mu} \mathbf{y}_{i;\lambda}$
  $\sigma \leftarrow AF(\sigma)$ ▷ Adaptation of σ
  $g \leftarrow g + 1$
until any termination criterion is fulfilled

that the (1+1)-ES is – apart from the (λ_{opt})-ES [26] not considered here – the best performing ES on the noise-free sphere. The comparison will be based on the quality gain. Defining the normalized quality gain as

$$q^* = q\frac{N}{2R^2}, \tag{70}$$

one obtains with (25) and $W = 1$

$$q^*_{\text{SPSA}} = 2a^{(t)}N\left(1 - a^{(t)}N\right). \tag{71}$$

Substituting the optimal gradient step $a^{(t)}_{\text{nf}}$ (30) yields

$$q^*_{\text{SPSA,opt}} = \frac{1}{2}. \tag{72}$$

For ES, there exist two common performance measures, the quality gain and the progress rate. The latter measures the progress in the objective vector space. It was shown in [27] that both measures coincide for $N \to \infty$.[7] The equation for the (1+1)-ES on the noise-free sphere is

$$q^*_{\text{ES}} = \frac{\sigma^*}{\sqrt{2\pi}}\exp\left(-\frac{1}{8}\sigma^{*2}\right) - \frac{\sigma^{*2}}{2}\left(1 - \Phi\left(\frac{\sigma^*}{2}\right)\right), \tag{73}$$

where

$$\sigma^* = \sigma\frac{N}{R} \quad \text{and} \tag{74}$$

$$\Phi(x) := \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{1}{2}t^2\right)\mathrm{d}t \tag{75}$$

is the cumulative distribution function of the standard normal distribution. The maximal progress for the (1+1)-ES occurs at $\sigma^* \approx 1.224$, cf. [3]. Finally, let us define efficiency as

$$\nu = \begin{cases} \dfrac{q^*_{\text{ES}}}{\lambda}, & \text{for ES,} \\ \dfrac{q^*_{\text{SPSA}}}{2W}, & \text{for SPSA,} \end{cases} \tag{76}$$

[7] The same can be shown for SPSA, however, it is omitted for brevity.

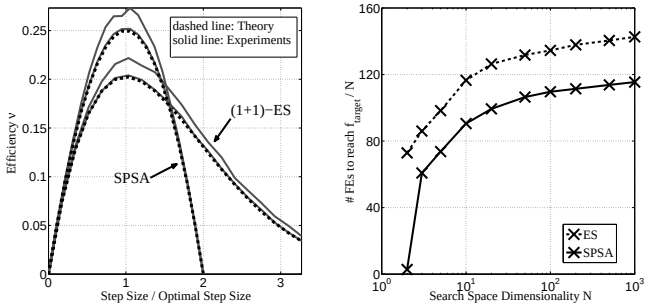

Figure 7: Efficiency comparison of SPSA and (1+1)-ES. *Left*: Efficiency over different step sizes. Shown are the theoretical results (dashed lines) and simulation results for $N = 10$ (top solid line) and $N = 100$ (bottom solid line). The ES is shown as black line, while the results for SPSA are shown as grey line. *Right*: Ratio of necessary function evaluations to search space dimensionality N for reaching $f_{\text{target}} = 10^{-20}$ for different search space dimensionalities N.

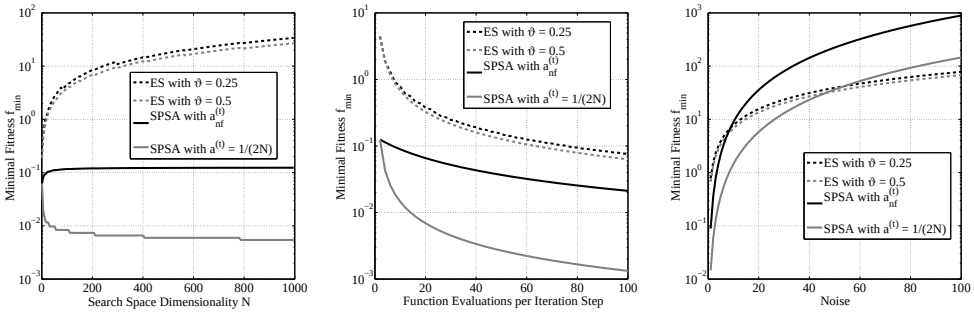

Figure 8: Comparison of $(\mu/\mu_I, \lambda)$-ES with SPSA with constant gradient step.

i.e., for the (1+1)-ES yielding $\nu = q^*_{\text{ES}}$. The reason for the difference in the definition is, the difference in the number of function evaluations per iteration step. In the left-hand plot of Fig. 7 the efficiency for both strategies is compared, based on the theoretical equations and for simulations with $N = 10$ and $N = 100$. While SPSA reaches slightly higher efficiency values, ES has a broader range for the step size to attain convergence. For $N = 100$ both strategies are close to the theoretical value, while for $N = 10$ the theoretical predictions underestimate the efficiency. In the right-hand plot of Fig. 7, the number of function evaluations (FEs) to reach f_{target} for different search space dimensionalities is shown. Again, SPSA performs better than the (1+1)-ES, especially for low N.

5.2. The Sphere with Constant Noise

For the constant noise model SPSA can reach the optimum ($t \to \infty$) if the optimal step sequence $a^{(t)}_{\text{cn}}$ (36) or (45) with $\alpha \leq 1$ is used. For ES, on the other hand, it will always have a approximation error (*residual location error*) the expected value of which is given by [28]

$$f_{\text{min}} = \frac{\sigma_\epsilon N}{4\mu c_{\mu/\mu_I,\lambda}}, \tag{77}$$

where $c_{\mu/\mu_I,\lambda}$ is the so-called progress coefficient [3]

$$c_{\mu/\mu_I,\lambda} = \frac{\lambda-\mu}{2\pi}\binom{\lambda}{\mu}\int_{-\infty}^{\infty}\exp\left(-t^2\right)\Phi(t)^{\lambda-\mu-1}\left(1-\Phi(t)\right)^{\mu-1}\mathrm{d}t \tag{78}$$

with $\Phi(t)$ defined by (75). A comparison of ES with SPSA is performed under assumption that $a^{(t)} = \text{const}$. The residual location error for SPSA with constant gradient step was determined in (35) and reads

$$f_{\min} = \frac{a^{(0)}N}{8(W - a^{(0)}(N+W-1))}\left(\frac{\sigma_\epsilon}{c^{(0)}}\right)^2. \tag{79}$$

Thus, both strategies will attain a residual location error and we are interested in the influence of certain parameters on the residual location error. The results are shown in Fig. 8. The left-hand plot shows the influence of N. Two ES variants, one with $\vartheta = \mu/\lambda = 0.25$ and one with $\vartheta = 0.5$ are shown. The value of λ is calculated by

$$\lambda = 4 + \lfloor 3\log(N)\rfloor. \tag{80}$$

Both ES variants display a similar behavior and scale linearly with N. For SPSA, a variant using $a_{\mathrm{nf}}^{(t)}$ and a variant with $a^{(t)} = 1/(2N)$ are shown. Both use $W = \lfloor\lambda/2\rfloor$ gradient samples per iteration. Thus, all strategies use the same budget of function evaluations. For the latter choice of $a^{(t)}$ the gradient step size is independent of W which allows for smaller residual location errors since $a^{(t)}$ remains small. In the first case, $a^{(t)}$ increases with W, however the attained residual location error remains almost constant. The middle plot shows the influence of W and λ. One can clearly observe that SPSA reaches smaller residual location errors, especially the variant with $a^{(t)} = 1/(2N)$. However, this variant will need much more function evaluations to reach the vicinity of the steady state since the convergence rate is not optimal during the phase where the influence of the noise is negligible. Finally, in the right plot the influence of the noise strength is displayed. For ES the noisy strength equals σ_ϵ, while for SPSA it equals $\sigma_\epsilon/c^{(t)}$ (see discussion in Section 4). The stronger increase for the SPSA variants is due to the quadratic appearance of the noise strength in (79), while it is only linear for ES (77). Additionally, the most significant difference between the two strategies is that the residual location error depends on the step size for SPSA. Decreasing the gradient step size decreases the (expected) minimal distance to $\mathbf{x}_{\mathrm{opt}}$, hence $a^{(t)} \to 0 \Rightarrow f_{\min} \to 0$. For ES, the minimal distance does also depend on the step size, however, if $\sigma^* \to 0$ (77) is obtained. Overall, one can conclude that SPSA is able to attain smaller residual location errors than ES except for large noise strengths.

5.3. The Sphere with State-Dependent Noise

For this noise model, the noise at the initial state is critical. From (42) we already know for SPSA that the initial distance must be smaller than a certain value to attain convergence. For ES, one can conclude from the constant noise model, that the initial noise strength must be connected with a residual location error which is smaller than the initial distance to $\mathbf{x}_{\mathrm{opt}}$. For the following comparison, we assume that both strategies are able to converge. Then, our interest lies in how *efficiently* the strategies approach the optimum. Using the efficiency definition (76) and (39) one obtains

$$\nu_{\mathrm{SPSA}} = \frac{a^{(t)}N}{W}\left(1 - \frac{a^{(t)}}{W}(N+W-1)\right) - \frac{{a^{(t)}}^2R^2}{2W^2}\left(\frac{\sigma_\epsilon^*}{c^{(t)}}\right)^2. \tag{81}$$

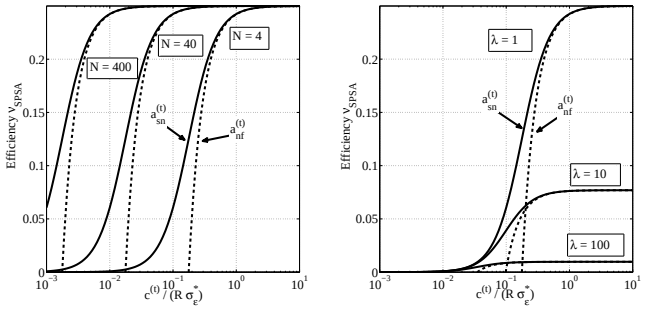

Figure 9: Efficiency of SPSA for the sphere model with state-dependent noise. The dashed curves represents theoretical results with $a^{(t)} = a^{(t)}_{\mathrm{nf}}$, while the solid curves represent $a^{(t)} = a^{(t)}_{\mathrm{sn}}$. *Left*: Influence of the search space dimensionality N on the efficiency ν_{SPSA}. *Right*: Influence of the number of gradient approximations per iteration W on the efficiency ν_{SPSA}.

From the theoretical analysis we know that the choice of $c^{(t)}$ is critical. Given that the algorithm itself has to deal with σ_ϵ instead of σ^*_ϵ,[8] one can conclude from (81) that if $c^{(t)} \approx R_0\sigma^*_\epsilon$ is chosen, convergence can be achieved for all possible initial states R_0. In Fig. 9 the efficiency ν_{SPSA} is shown for different gradient step sequences. The solid lines represent the results of (81) with $a^{(t)} = a^{(t)}_{\mathrm{sn}}$, while the dashed ones are obtained with $a^{(t)} = a^{(t)}_{\mathrm{nf}}$. In both plots one can observe that using $a^{(t)}_{\mathrm{nf}}$ can yield $\nu_{\mathrm{SPSA}} < 0$ for a given set of $c^{(t)}, R$, and σ^*_ϵ, while using $a^{(t)}_{\mathrm{sn}}$ the efficiency is always greater than zero. On the other hand, substituting the respective gradient step sequences (30) or (43) into (81) and taking the limit $N \to \infty$ yields $\nu_{\mathrm{SPSA}} = 0.25$ for both, cf. Fig. 7. From the right-hand plot of Fig. 9 one can see that $W = 1$ is the best choice and that for $W > 1$ ν_{SPSA} can not reach the noise-free value of $\nu_{\mathrm{SPSA}} = 0.25$.

The $(\mu/\mu_I, \lambda)$-ES was thoroughly analyzed for the sphere with state-dependent noise in [27]. Since, we don't want to reproduce this work, we will just state some of the interesting facts. First, the sphere in the limit of infinite search space dimensionality was considered and the efficiency was derived yielding

$$\nu_{\mathrm{ES}} = \frac{\sigma^* c_{\mu/\mu_I,\lambda}}{\lambda\sqrt{1 + \left(\frac{\sigma^*_\epsilon}{\sigma^*}\right)^2}} - \frac{\sigma^{*2}}{2\mu\lambda} \tag{82}$$

with the assumption $N \to \infty$. From (82), one can derive the following convergence criterion

$$\sigma^*_\epsilon < 2\mu c_{\mu/\mu_I,\lambda}. \tag{83}$$

This shows that increasing μ – and therefore λ for constant $\vartheta = \mu/\lambda$ – ES should be able to converge for any σ^*_ϵ. This behavior is also shown in the right-hand plot of Fig. 10, where the maximal efficiency for different λ with $\mu \approx \lambda/3$ is shown. The efficiency itself depends on σ^* as shown in the left-hand plot. Comparing (82) with its noise-free version (see for example [15]), one sees that ES reaches the noise-free efficiency for $\sigma^*_\epsilon/\sigma^* \to 0$. The derived maximal efficiency is 0.202. However, a more detailed analysis [15] showed that for finite search space

[8] The term σ^*_ϵ is an artificial term which is useful for the analysis, however, it never appears in the actual implementation.

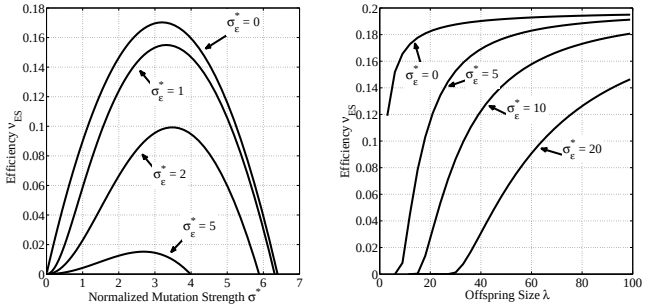

Figure 10: Efficiency of $(\mu/\mu_I, \lambda)$-ES for the sphere model with state-dependent noise. *Left*: Influence of the normalized mutation strength σ^* on the efficiency ν_{ES} for $(3/3_I, 10)$-ES. *Right*: Maximal efficiency ν_{ES} in dependence of the parental population size μ with $\lambda = 100$.

dimensionalities the efficiency is reduced by decreasing N. It was shown that the maximal efficiency is reached for intermediate values of λ. Further increasing λ reduces the efficiency, an effect which can not be predicted by the asymptotic ($N \to \infty$) Eq. (82). Overall, the same result as for the other noise models can be stated, namely that SPSA slightly outperforms ES if both strategies operate with parameter settings close to optimality.

6. Summary and Conclusion

In this work Simultaneous Perturbation Stochastic Approximation was analyzed with the help of the theoretical approach developed for Evolution Strategies. The advantage of this approach is that it can be applied to noisy and noise-free optimization at the same time. It allows to (approximately) determine the short term dynamic behavior ($t \ll \infty$). Furthermore, the influence of the strategy parameters on the dynamic behavior of the strategy can be evaluated, which provides valuable information for practitioners in the field. A drawback of the approach is that the results derived are only valid for the class of functions considered and no guarantee for generalization can be given. On the other hand, the results might be (partially) reused as done for the analysis of ESs on certain ellipsoidal functions [29]. Another simplification is that an infinite search space dimensionality must be considered. However, simulation results showed that the equations derived are good approximations for finite search space dimensionalities. The function under consideration in this work was the sphere model and it was shown that the approach was able to

a) derive theoretical approximations for the (one-step and overall) dynamics,

b) obtain convergence criteria and optimal parameter settings.

Especially the derived optimal gradient step sizes showed that an improvement for SPSA can be made by using gradient step sizes the values of which are close to the optimal ones. However, to derive an *adaptive* gradient step size rule, more test functions need to be considered. First steps in this direction have been already made with the adaptive SPSA [13, 14], which uses additional function evaluations to approximate the Hessian matrix. Additionally, as for the sphere model the step size factor $c^{(t)}$ plays only a minor role, however, it is expected that this will not be the case for other types of test functions.

The insights obtained from the theoretical analysis were used in the 2nd part for a comparison with simple Evolution Strategies. Here we have taken advantage of using a unified theoretical approach, since the performance measures used are compatible. For all three noise models (noise-free, constant noise, state-dependent noise) SPSA performed better than the ES variants considered. In the constant noise case restriction had to be applied, given that SPSA could reach the optimum and ES could not. However, SPSA does reach the optimum for $t \to \infty$, an information not very useful for practical considerations. Hence, only the attained residual location error was compared neglecting any effects from step size adaptation procedures. For the state-dependent noise model, SPSA will diverge if the initial distance to the optimum is too large. One can influence the critical distance by use of resampling the gradient approximation, decreasing the initial gradient step size (which will reduce the convergence rate) or increasing the gradient approximation step size $c^{(t)}$ (which could be problematic if a bounded search space domain is considered). A peculiarity of SPSA is that for the constant noise model the residual location error depends on the gradient step size. This dependency is the reason why SPSA can reach the optimum ($t \to \infty$), however, on the other hand it reduces the convergence rate. Thus, for this noise model a decreasing $a^{(t)}$ sequence is beneficial if the noise can not be neglected in the function evaluation process.

The results obtained are promising. They should encourage the use of the presented approach to other test functions and optimization strategies. This will allow for a more detailed and comprehensive comparison of different strategies providing the option to also design improved algorithms for noisy optimization. The results should also be extended in the future. One improvement would be the analysis of ellipsoidal functions and possibly incorporating the adaptive version of SPSA [14]. Another question is how the results (and the analysis) can be transferred to less restrictive noise models, e.g., the noise being a iid sequence with zero mean and finite second or higher order moments.

7. Acknowledgments

Support by the Austrian Science Fund (FWF) under grant P19069-N18 is gratefully acknowledged.

References

[1] J. C. Spall, Multivariate Stochastic Approximation Using a Simultaneous Perturbation Gradient Approximation, IEEE Transactions on Automatic Control 37 (1992) 332–341.
[2] J. C. Spall, Introduction to Stochastic Search and Optimization, John Wiley & Sons, Hoboken, NJ, 2003.
[3] H.-G. Beyer, The Theory of Evolution Strategies, Natural Computing Series, Springer, Heidelberg, 2001.
[4] D. Arnold, A. MacLeod, Step Length Adaptation on Ridge Functions, Evolutionary Computation 16 (2008) 151–184.
[5] H.-G. Beyer, D. V. Arnold, The Steady State Behavior of $(\mu/\mu_I, \lambda)$-ES on Ellipsoidal Fitness Models Disturbed by Noise, in: E. e. a. Cantú-Paz (Ed.), GECCO-2003: Proceedings of the Genetic and Evolutionary Computation Conference, Springer, Berlin, Germany, 2003, pp. 525–536.
[6] D. V. Arnold, H.-G. Beyer, A. Melkozerov, On the Behaviour of Weighted Multi-Recombination Evolution Strategies Optimising Noisy Cigar Functions, in: G. Raidl et al. (Ed.), GECCO-2009: Proceedings of the Genetic and Evolutionary Computation Conference, ACM, New York, 2009, pp. 483–490.
[7] L. Gerencsér, Z. Vágó, SPSA in Noise Free optimization, in: Proceedings of the American Control Conference, IEEE, 2000, pp. 3284–3288.
[8] L. Gerencsér, Z. Vágó, The Mathematics of Noise-Free SPSA, in: Proceedings of the 40th IEEE Conference on Decision and Control, IEEE, 2001, pp. 4400–4405.

[9] J. Kiefer, J. Wolfowitz, Stochastic Estimation of the Maximum of a Regression Function, Annals of Mathematical Statistics 23 (1952) 462–466.
[10] P. Sadegh, J. C. Spall, Optimal Random Perturbations for Stochastic Approximation with a Simultaneous Perturbation Gradient Approximation, IEEE Transactions on Automatic Control 43 (1998) 1480–1484.
[11] D. W. Hutchinson, On an Efficient Distribution of Perturbation for Simulation Optimization using Simultaneous Perturbation Stochastic Approximation, in: M. H. Hamza (Ed.), Proceedings of the IASTED AMS 2002, ACTA Press, 2002, pp. 440–444.
[12] P. Gilmore, C. Kelley, An Implicit Filtering Algorithm for Optimization of Functions with Many Local Minima, SIAM Journal on Optimization 5 (1995) 269–285.
[13] J. C. Spall, Adaptive Stochastic Approximation by the Simultaneous Perturbation Method, IEEE Transactions on Automatic Control 45 (2000) 1839–1853.
[14] J. C. Spall, Feedback and Weighting Mechanisms for Improving Jacobian Estimates in the Adaptive Simultaneous Perturbation Algorithm, IEEE Transactions on Automatic Control 54 (2009) 1216–1229.
[15] D. V. Arnold, Local Performance of Evolution Strategies in the Presence of Noise, Ph.D. Thesis, University of Dortmund, Dortmund, 2001.
[16] J. C. Spall, S. D. Hill, D. R. Stark, Theoretical Comparison of Evolutionary Computation and Other Optimization Approaches, in: P. Angeline (Ed.), Proceedings of the CEC'99 Conference, IEEE, Piscataway, NJ, 1999, pp. 1398–1405.
[17] P. N. Suganthan, N. Hansen, J. J. Liang, K. Deb, Y. P. Chen, A. Auger, S. Tiwari, Problem definitions and evaluation criteria for the CEC 2005 Special Session on Real Parameter Optimization, Technical Report, Nanyang Technological University, 2005.
[18] I. Rechenberg, Evolutionsstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution, Frommann-Holzboog Verlag, Stuttgart, 1973.
[19] H.-G. Beyer, H.-P. Schwefel, Evolution Strategies: A Comprehensive Introduction, Natural Computing 1 (2002) 3–52.
[20] S. Meyer-Nieberg, H.-G. Beyer, Self-Adaptation in Evolutionary Algorithms, in: F. Lobo, C. Lima, Z. Michalewicz (Eds.), Parameter Setting in Evolutionary Algorithms, Springer, Berlin, 2007, pp. 47–75.
[21] N. Hansen, A. Ostermeier, A. Gawelczyk, Step-size adaption based on non-local use of selection information, in: Y. D. et al. (Ed.), Parallel Problem Solving from Nature - PPSN III, Springer Verlag, 1994, pp. 189–198.
[22] N. Hansen, A. Ostermeier, Completely Derandomized Self-Adaptation in Evolution Strategies, Evolutionary Computation 9 (2001) 159–195.
[23] I. Rechenberg, The Evolution Strategy. A Mathematical Model of Darwinian Evolution, in: E. Frehland (Ed.), Synergetics - From Microscopic to Macroscopic Order, Springer-Verlag, Berlin, 1984, pp. 122–132.
[24] I. Rechenberg, Evolutionsstrategie '94, Frommann-Holzboog Verlag, Stuttgart, 1994.
[25] N. Hansen, The CMA evolution strategy: a comparing review, in: J. Lozano, P. Larranaga, I. Inza, E. Bengoetxea (Eds.), Towards a new evolutionary computation. Advances on estimation of distribution algorithms, Springer, 2006, pp. 75–102.
[26] D. V. Arnold, Optimal Weighted Recombination, in: A. H. Wright et al. (Ed.), Foundations of Genetic Algorithms 8, Springer Verlag, 2005, pp. 215–237.
[27] D. V. Arnold, H.-G. Beyer, Local Performace of the $(\mu/\mu_I, \lambda)$-ES in a Noisy Environment, in: W. Martin, W. Spears (Eds.), Foundations of Genetic Algorithms, 6, Morgan Kaufmann, San Francisco, CA, 2001, pp. 127–141.
[28] H.-G. Beyer, D. V. Arnold, S. Meyer-Nieberg, A New Approach for Predicting the Final Outcome of Evolution Strategy Optimization under Noise, Genetic Programming and Evolvable Machines 6 (2005) 7–24.
[29] H.-G. Beyer, S. Finck, Performance of the $(\mu/\mu_I, \lambda)$-σSA-ES on a Class of PDQFs, IEEE Transactions on Evolutionary Computation 14 (2010) 400–418.
[30] H. F. Chen, T. E. Duncan, B. Pasik-Duncan, A Kiefer-Wolfowitz Algorithm with Randomized Differences, IEEE Transactions on Automatic Control 44 (1999) 442–453.
[31] H. Chen, Stochastic approximation and Its Applications, Kluwer Academic Publishers, 2002.
[32] I. J. Wang, E. K. P. Chong, A Deterministic Analysis of Stochastic Analysis with Randomized Directions, IEEE Transactions on Automatic Control 43 (1998) 1745–1749.
[33] M. Abramowitz, I. A. Stegun, Pocketbook of Mathematical Functions, Verlag Harri Deutsch, Thun, 1984.

Appendix A. Conditions and Theorems for the Convergence of SPSA

This appendix states the conditions and convergence theorems for SPSA form Spall's proof [1, 2]:

C1 **(Gain Sequences)** $a^{(t)}$ and $c^{(t)} > 0$, $a^{(t)}$ and $c^{(t)} \to 0$, $\sum_{k=1}^{\infty} a^{(t)} = \infty$, and $\sum_{k=1}^{\infty} \frac{a^{(t)^2}}{c^{(t)^2}} < \infty$.

C2 **(Relationship to ODE)** Let $\mathbf{f}(\mathbf{x})$ be continuous on $\mathbb{R}^N$. With $\mathbf{Z}(t) \in \mathbb{R}^N$ representing a time-varying function (t denoting time), suppose that the differential equation given by $\frac{\mathrm{d}\mathbf{Z}(t)}{\mathrm{d}t} = -\mathbf{f}(\mathbf{Z}(t))$ has an asymptotically stable equilibrium point at $\mathbf{x}^*$.

C3 **(Iterate boundedness)** $\sup_{k\geq 0} \|\mathbf{x}^{(t)}\| < \infty$ and $\mathbf{x}^{(t)}$ lies in a closed and bounded subset of the "domain of attraction" for the differential equation of C2 infinitely often.

C4 **(Measurement noise; relationship between the measurement noise and $\mathbf{\Delta}^{(t)}$)** For all t, $\mathrm{E}\left[(\epsilon^+ - \epsilon^-)|\mathcal{I}_k, \mathbf{\Delta}^{(t)}\right] = 0$ and the ratio of measurement to perturbation is such that $\mathrm{E}\left[\left(\frac{G(\mathbf{x}^{(t)} \pm c^{(t)}\mathbf{\Delta}^{(t)})}{\Delta_{ki}}\right)^2\right]$ is uniformly bounded over t and i.
($\mathcal{I}_k = \{\mathbf{x}^{(1)}, \mathbf{x}_2, \ldots, \mathbf{x}^{(t)}; \mathbf{\Delta}_1, \mathbf{\Delta}_2, \ldots, \mathbf{\Delta}_{k-1}\}$)

C5 **(Smoothness of F)** F is three-times continuously differentiable and bounded on $\mathbb{R}^N$.

C6 **(Statistical properties of the perturbation $\mathbf{\Delta}$)** All Δ_{ki} are independent for all k, i, identically distributed for all i at each t, symmetrically distributed around zero and uniformly bounded in magnitude for all k, i.

C7 The continuity and equicontinuity assumptions about $\mathrm{E}\left[(\epsilon^* - \epsilon^-)^2|\mathcal{I}_k\right]$ from [1, Prop. 2] hold.

C8 $\mathbf{H}(\mathbf{x}^*)$ is positive definite where $\mathbf{H}(\mathbf{x})$ is the Hessian matrix of $\hat{f}(\mathbf{x})$. Further, let λ_i, denote the ith eigenvalue of $a^{(0)}\mathbf{H}(\mathbf{x}^*)$, where $a^{(0)}$ is from the $a^{(t)}$-sequence. If $\alpha = 1$, then $\beta < 2\min_i(\lambda_i)$.

C9 $\mathrm{E}\left[\Delta_{ki}^2\right] \to \rho$, $\mathrm{E}\left[\Delta_{ki}^{-2}\right] \to \rho'$, and $\mathrm{E}\left[(\epsilon^* - \epsilon^-)^2|\mathcal{I}_k\right] \to \rho''$ for strictly positive constants ρ, ρ', and ρ'' (almost sure (a.s.) in the latter case) as $t \to \infty$.

Theorem 1 [2, p. 186]: *Suppose that the conditions C1 – C6 hold. Further, suppose that $\mathbf{x}^*$ is a unique minimum in the search domain. Then, for the SPSA algorithm, $\mathbf{x}^{(t)} \to \mathbf{x}^*$ a.s. as $t \to \infty$.*

Theorem 2 [2, p. 186]: *Suppose that the gains have the standard form $a^{(t)} = \frac{a^{(0)}}{(t+A)^\alpha}$ and $c^{(t)} = \frac{c^{(0)}}{k^\gamma}$, $k = 1, 2, \ldots$, with $a^{(0)}, c^{(0)}, \alpha$, and γ strictly positive, $A \geq 0$, $\beta = \alpha - 2\gamma > 0$, and $3\gamma - \frac{\alpha}{2} \geq 0$. Further, suppose that conditions C1 – C6 from Theorem 1 and conditions C7 – C9 hold. Then, for the SPSA algorithm,*

$$k^{\frac{\beta}{2}}(\mathbf{x}^{(t)} - \mathbf{x}^*) \xrightarrow{dist.} \mathcal{N}(\mu, \mathbf{\Sigma}) \quad \text{as } k \to \infty, \tag{A.1}$$

where μ and $\mathbf{\Sigma}$ are a mean vector and the covariance matrix.

Note, above proof does not apply to noise-free SPSA. A proof for this case was presented by Gerencsér and Vágó in [7, 8].

In addition to this proof there exist other variants [30, 31] which try to relax some of the above mentioned requirements. As a result one obtains almost sure convergence, however, without conditions C2-C4, relaxed conditions C1 and C5, and a weakened condition C6. For example in [31] the so-called Trajectory-Subsequence method is used for the analysis of the algorithm, which seems able to handle noise-free SPSA. Additionally, a deterministic approach is given in [32].

Appendix B. Deriving $\|g\|^2$

In the following the steps of deriving (22) from (21) are described. The square of the gradient norm with λ gradient approximations can be written as

$$\begin{aligned}\|\mathbf{g}\|^2 =& \|\frac{1}{\lambda}\sum_{l=1}^{\lambda}\left(2\mathbf{x}^{\mathrm{T}}\boldsymbol{\Delta}_l + \frac{\tilde{\sigma}_\epsilon \mathcal{N}_l(0,1)}{2c}\right)\boldsymbol{\Delta}_l\|^2 \\ =& \frac{1}{\lambda^2}\sum_{l=1}^{\lambda}\sum_{m=1}^{\lambda}\left(2\mathbf{x}^{\mathrm{T}}\boldsymbol{\Delta}_l + \frac{\tilde{\sigma}_\epsilon \mathcal{N}_l(0,1)}{2c}\right)\left(2\mathbf{x}^{\mathrm{T}}\boldsymbol{\Delta}_m + \frac{\tilde{\sigma}_\epsilon \mathcal{N}_m(0,1)}{2c}\right)\boldsymbol{\Delta}_l^{\mathrm{T}}\boldsymbol{\Delta}_m.\end{aligned} \tag{B.1}$$

Note, the iteration counter t is not shown for brevity. Equation B.1 can be expanded to

$$\begin{aligned}\|\mathbf{g}\|^2 =& \frac{1}{\lambda^2}\sum_{l=1}^{\lambda}\left(4(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Delta}_l)^2 + 2(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Delta}_l)\frac{\tilde{\sigma}_\epsilon \mathcal{N}_l(0,1)}{2c} + \frac{\tilde{\sigma}_\epsilon^2 \mathcal{N}_l(0,1)^2}{4c^2}\right)\boldsymbol{\Delta}_l^{\mathrm{T}}\boldsymbol{\Delta}_l \\ &+ \frac{1}{\lambda^2}\sum_{l=1}^{\lambda}\sum_{m\neq l}\left[4(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Delta}_l)(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Delta}_m)\boldsymbol{\Delta}_l^{\mathrm{T}}\boldsymbol{\Delta}_m + \frac{\tilde{\sigma}_\epsilon \mathcal{N}_m(0,1)}{c}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Delta}_l)\boldsymbol{\Delta}_l^{\mathrm{T}}\boldsymbol{\Delta}_m\right. \\ &\left.+ \frac{\tilde{\sigma}_\epsilon \mathcal{N}_l(0,1)}{c}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Delta}_m)\boldsymbol{\Delta}_l^{\mathrm{T}}\boldsymbol{\Delta}_m + \frac{\tilde{\sigma}_\epsilon^2 \mathcal{N}_l(0,1)\mathcal{N}_m(0,1)}{4c^2}\boldsymbol{\Delta}_l^{\mathrm{T}}\boldsymbol{\Delta}_m\right].\end{aligned} \tag{B.2}$$

Defining

$$\mathrm{S}_1 := \frac{1}{\lambda^2}\sum_{l=1}^{\lambda}\left(4(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Delta}_l)^2 + 2(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Delta}_l)\frac{\tilde{\sigma}_\epsilon \mathcal{N}_l(0,1)}{2c} + \frac{\tilde{\sigma}_\epsilon^2 \mathcal{N}_l(0,1)^2}{4c^2}\right)\boldsymbol{\Delta}_l^{\mathrm{T}}\boldsymbol{\Delta}_l, \tag{B.3}$$

the expectation of S_1 for a given point $\mathbf{x}$ with $\|\mathbf{x}\| = R$ yields

$$\mathrm{E}\left[\mathrm{S}_1|\mathbf{x}\right] = \frac{1}{\lambda^2}\sum_{l=1}^{\lambda} 4\mathrm{E}\left[(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Delta}_l)^2\boldsymbol{\Delta}_l^{\mathrm{T}}\boldsymbol{\Delta}_l|\mathbf{x}\right] + 2\mathrm{E}\left[(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Delta}_l)\frac{\tilde{\sigma}_\epsilon \mathcal{N}_l(0,1)}{2c}\boldsymbol{\Delta}_l^{\mathrm{T}}\boldsymbol{\Delta}_l|\mathbf{x}\right] + \mathrm{E}\left[\frac{\tilde{\sigma}_\epsilon^2 \mathcal{N}_l(0,1)^2}{4c^2}\boldsymbol{\Delta}_l^{\mathrm{T}}\boldsymbol{\Delta}_l|\mathbf{x}\right]. \tag{B.4}$$

The first expectation in (B.4) can be written as

$$\mathrm{E}\left[(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Delta}_l)^2\boldsymbol{\Delta}_l^{\mathrm{T}}\boldsymbol{\Delta}_l|\mathbf{x}\right] = \mathrm{E}\left[\sum_{n=1}^{N}\left(\sum_{i=1}^{N} x_i\Delta_{l_i}\right)^2\Delta_{l_n}^2|\mathbf{x}\right] = \mathrm{E}\left[\sum_{n=1}^{N}\sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j \Delta_{l_i}\Delta_{l_j}\Delta_{l_n}^2|\mathbf{x}\right]. \tag{B.5}$$

Using

$$\mathrm{E}\left[\Delta_{l_i}\Delta_{l_j}\right] = \begin{cases}0 & i \neq j \\ 1 & i = j\end{cases} \tag{B.6}$$

and

$$\Delta_{l_n}^2 = 1, \tag{B.7}$$

one obtains

$$\mathrm{E}\left[(\mathbf{x}^\mathrm{T}\boldsymbol{\Delta}_l)^2\boldsymbol{\Delta}_l^\mathrm{T}\boldsymbol{\Delta}_l|\mathbf{x}\right] = \sum_{k=1}^{N}\sum_{i=1}^{N} x_i^2 = R^2N. \tag{B.8}$$

The second expectation in (B.4) can be written as

$$\mathrm{E}\left[(\mathbf{x}^\mathrm{T}\boldsymbol{\Delta}_l)\frac{\tilde{\sigma}_\epsilon\mathcal{N}_l(0,1)}{2c}\boldsymbol{\Delta}_l^\mathrm{T}\boldsymbol{\Delta}_l|\mathbf{x}\right] = \mathrm{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{N} x_i\Delta_{l_i}\Delta_{l_j}^2 z|\mathbf{x}\right], \tag{B.9}$$

with $z = \dfrac{\tilde{\sigma}_\epsilon\mathcal{N}_l(0,1)}{2c}$. Note, in the following all expectations are w.r.t. $\boldsymbol{\Delta}$-terms. The expectation of $\mathcal{N}(0,1)$ will be handled separately. Using $\mathrm{E}\left[\Delta_{l_i}\Delta_{l_j}^2\right] = 0$ for all i, j, one obtains

$$\mathrm{E}\left[(\mathbf{x}^\mathrm{T}\boldsymbol{\Delta}_l)\frac{\tilde{\sigma}_\epsilon\mathcal{N}_l(0,1)}{2c}\boldsymbol{\Delta}_l^\mathrm{T}\boldsymbol{\Delta}_l|\mathbf{x}\right] = 0. \tag{B.10}$$

The last expectation term in (B.4) can be written as

$$\mathrm{E}\left[\frac{\tilde{\sigma}_\epsilon^2\mathcal{N}_l(0,1)^2}{4c^2}\boldsymbol{\Delta}_l^\mathrm{T}\boldsymbol{\Delta}_l|\mathbf{x}\right] = \frac{N\tilde{\sigma}_\epsilon^2}{4c^2}, \tag{B.11}$$

since $\mathrm{E}\left[\boldsymbol{\Delta}_l^\mathrm{T}\boldsymbol{\Delta}_l\right] = N$ (see (B.6)). Now (B.8), (B.10), and (B.11) can be substituted into (B.4). This yields the following expectation

$$\mathrm{E}\left[\mathrm{S}_1|\mathbf{x}\right] = \frac{1}{\lambda^2}\sum_{l=1}^{\lambda}\left(4R^2N + \frac{N\tilde{\sigma}_\epsilon^2}{4c^2}\right). \tag{B.12}$$

The sum of the squares of the Gaussian distributed random variables yields a χ^2-distribution. Thus, the expected value (B.12) can be written as

$$\mathrm{E}\left[\mathrm{S}_1|R\right] = \frac{1}{\lambda^2}\left(\lambda 4R^2N + \frac{N\tilde{\sigma}_\epsilon^2\lambda}{4c^2}\right) = \frac{N}{\lambda}\left(4R^2 + \frac{\tilde{\sigma}_\epsilon^2}{4c^2}\right) \tag{B.13}$$

Defining

$$\begin{aligned}\mathrm{S}_2 :=& \frac{1}{\lambda^2}\sum_{l=1}^{\lambda}\sum_{m\neq l} 4(\mathbf{x}^\mathrm{T}\boldsymbol{\Delta}_l)(\mathbf{x}^\mathrm{T}\boldsymbol{\Delta}_m)\boldsymbol{\Delta}_l^\mathrm{T}\boldsymbol{\Delta}_m + \frac{\tilde{\sigma}_\epsilon\mathcal{N}_m(0,1)}{c}(\mathbf{x}^\mathrm{T}\boldsymbol{\Delta}_l)\boldsymbol{\Delta}_l^\mathrm{T}\boldsymbol{\Delta}_m \\ &+ \frac{\tilde{\sigma}_\epsilon\mathcal{N}_l(0,1)}{c}(\mathbf{x}^\mathrm{T}\boldsymbol{\Delta}_m)\boldsymbol{\Delta}_l^\mathrm{T}\boldsymbol{\Delta}_m + \frac{\tilde{\sigma}_\epsilon^2\mathcal{N}_l(0,1)\mathcal{N}_m(0,1)}{4c^2}\boldsymbol{\Delta}_l^\mathrm{T}\boldsymbol{\Delta}_m,\end{aligned} \tag{B.14}$$

the expectations of the different terms in (B.14) will be determined next. At first the expectation of $4(\mathbf{x}^\mathrm{T}\boldsymbol{\Delta}_l)(\mathbf{x}^\mathrm{T}\boldsymbol{\Delta}_m)\boldsymbol{\Delta}_l^\mathrm{T}\boldsymbol{\Delta}_m$ will be obtained. Starting with

$$\mathrm{E}\left[4(\mathbf{x}^\mathrm{T}\boldsymbol{\Delta}_l)(\mathbf{x}^\mathrm{T}\boldsymbol{\Delta}_m)\boldsymbol{\Delta}_l^\mathrm{T}\boldsymbol{\Delta}_m|\mathbf{x}\right] = 4\mathrm{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{n=1}^{N} x_i\Delta_{l_i}x_j\Delta_{m_j}\Delta_{l_n}\Delta_{m_n}|\mathbf{x}\right], \tag{B.15}$$

one can group the terms according to their indices. This yields

$$\mathrm{E}\left[4(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Delta}_l)(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Delta}_m)\boldsymbol{\Delta}_l^{\mathrm{T}}\boldsymbol{\Delta}_m|\mathbf{x}\right] = 4\mathrm{E}\left[\sum_{i=1}^{N}\sum_{j=1}^{N}\sum_{n=1}^{N} x_i x_j \Delta_{m_j}\Delta_{m_n}\Delta_{l_i}\Delta_{l_n}|\mathbf{x}\right]. \tag{B.16}$$

Using (B.6), one sees that for $i = n$ the expectation for $\boldsymbol{\Delta}_l$ does not vanish and that for $j = n$ the expectation for $\boldsymbol{\Delta}_m$ does not vanish. Hence, only for $i = n = j$ the expectation is not zero. Thus, one obtains

$$\mathrm{E}\left[4(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Delta}_l)(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Delta}_m)\boldsymbol{\Delta}_l^{\mathrm{T}}\boldsymbol{\Delta}_m|\mathbf{x}\right] = 4\mathrm{E}\left[\sum_{i=1}^{N} x_i^2\Delta_{m_i}^2\Delta_{l_i}^2|\mathbf{x}\right] = 4R^2, \tag{B.17}$$

by using the fact $\Delta^2 = 1$. The next expectation is

$$\mathrm{E}\left[\frac{\tilde{\sigma}_\epsilon \mathcal{N}_m(0,1)}{c}(\mathbf{x}^{\mathrm{T}}\boldsymbol{\Delta}_l)\boldsymbol{\Delta}_l^{\mathrm{T}}\boldsymbol{\Delta}_m|\mathbf{x}\right] = 0, \tag{B.18}$$

The expectation vanishes since $\mathrm{E}\left[\Delta_{m_j}\right] = 0$ for all j and due $\mathcal{N}_m(0,1)$ and $\boldsymbol{\Delta}$ being uncorrelated. A closer inspection of the remaining terms in (B.14) reveals that they all contain either a single $\boldsymbol{\Delta}_l$ or $\boldsymbol{\Delta}_m$ term. Thus, the same reasoning as above can be used to show that the expectations of these terms will vanish. Thus, the expectation of S_2 is obtained by substituting (B.17) into the expectation of (B.14)

$$\mathrm{E}\left[\mathrm{S}_2|R\right] = \frac{1}{\lambda^2}\sum_{l=1}^{\lambda}\sum_{m\neq l} 4R^2 = 4R^2\left(1 - \frac{1}{\lambda}\right). \tag{B.19}$$

Now putting everything together by substituting (B.12) and (B.19) into the expectation of (B.2) yields

$$\mathrm{E}\left[||\mathbf{g}||^2|R\right] = \frac{N}{\lambda}\left(4R^2 + \frac{\tilde{\sigma}_\epsilon^2}{4c^2}\right) + 4R^2\left(1 - \frac{1}{\lambda}\right). \tag{B.20}$$

Appendix C. Solving the inhomogeneous differential equation

The solution of inhomogeneous differential equation appearing in the constant noise case and the state-dependent noise case (after applying the substitution $u = f^{-1}$) is described below. The differential equation has the form

$$f' + f(z_1 t^{-\alpha} + z_2 t^{-2\alpha}) = z_3 k^{2\gamma - 2\alpha}, \tag{C.1}$$

where $f' = \frac{\mathrm{d}f}{\mathrm{d}kt}$ and the z_i are constants depending on the strategy and function parameters. Further $t \gg A$ is assumed and for $\alpha > 0$ $t^{-\alpha} \gg t^{-2\alpha}$ will be assumed. First, the homogeneous solution f_{h} will be obtained. Afterwards an ansatz is used to derive a particular solution for inhomogeneous equation f_{ih}. Finally, both solutions will be added to obtain the general solution for (C.1). The homogeneous equations are

$$f_{\mathrm{h}}' + f_{\mathrm{h}}(z_1 t + z_2 t) = 0 \quad \text{for } \alpha = 0 \text{ and} \tag{C.2}$$

$$f_{\mathrm{h}}' + f_{\mathrm{h}}(z_1 t^{-\alpha}) = 0 \quad \text{for } \alpha > 0, \tag{C.3}$$

which can be solved by using the ansatz

$$f_{\mathrm{h}} = c\exp\left(-Z(T)\right). \tag{C.4}$$

The exponent $Z(T)$ is given by

$$Z(t) = \begin{cases} \int_{T=1}^{t} z_1 T + z_2 T \mathrm{d}T, \text{ for } \alpha = 0 \\ \int_{T=1}^{t} z_1 T^{-\alpha} \mathrm{d}T, \text{ for } \alpha > 0. \end{cases} \tag{C.5}$$

The solution of (C.5) yields

$$Z(t) = \begin{cases} (z_1 + z_2)(t-1), \text{ for } \alpha = 0, \\ z_1 \ln(t), \text{ for } \alpha = 1, \\ \frac{z_1}{1-\alpha}\left(t^{1-\alpha} - 1\right), \text{ for } \alpha \neq 0, 1. \end{cases} \tag{C.6}$$

Substituting the respective equation in (C.4), the homogenous solution is obtained

$$f_{\mathrm{h}} = \begin{cases} c\exp\left((z_1 + z_2)(1-t)\right), \text{ for } \alpha = 0, \\ ct^{-z_1}, \text{ for } \alpha = 1, \\ c\exp\left(\frac{z_1}{1-\alpha}\left(1 - t^{1-\alpha}\right)\right), \text{ for } \alpha \neq 0, 1. \end{cases} \tag{C.7}$$

The constant c is obtained by solving $f_{\mathrm{h}}(t=1) = f_{\mathrm{start}}$.

For constant noise and state-dependent noise the ansatz for the particular solution is

$$f_{\mathrm{ih}} = c(t) f_{\mathrm{h}}, \tag{C.8}$$

where f_{h} is given by (C.7), however, without constant c. Substitution of (C.8) into (C.1) yields an integral equation for $c(t)'$ of type

$$c(t) = z_3 \int_{T=1}^{t} T^{2(\gamma-\alpha)} \exp\left(f(T)\right) \mathrm{d}T, \tag{C.9}$$

where $f(T)$ is a function depending on the homogeneous solution. The closed-form solution of above integral exist only for some special cases of α and γ. For some other cases the solution involves the generalized incomplete gamma function [33]. For the settings $\alpha = 0$, $\gamma = 0$ and $\alpha = 1$ solutions can be obtained which yield a interpretable solution. As example the solution for $\alpha = 1$ is shown in the following.

Using the ansatz

$$f_{\mathrm{ih}} = c(t) t^{-z_1} \tag{C.10}$$

yields

$$\begin{aligned} c(t) &= z_3 \int_{T=1}^{t} T^{2(\gamma-1)+z_1} \mathrm{d}T \\ &= \frac{z_3}{2\gamma - 1 + z_1}\left(t^{2\gamma-1+z_1} - 1\right). \end{aligned} \tag{C.11}$$

Thus, the particular solution is

$$f_{\mathrm{ih},\alpha=1} = \frac{z_3}{2\gamma - 1 + z_1}\left(t^{2\gamma-1} - t^{-z_1}\right). \tag{C.12}$$

Then the general solution to (C.1) with $\alpha = 1$ is

$$f_{\alpha=1} = \frac{z_3}{2\gamma - 1 + z_1}\left(t^{2\gamma-1} - t^{-z_1}\right) + ct^{-z_1}. \tag{C.13}$$

As before, the constant c is determined by solving $f_{\alpha=1}(t=1) = f_{\mathrm{start}}$.