An Analysis of Mutative σ-Self-Adaptation on Linear Fitness Functions

Nikolaus Hansen nikolaus.hansen@inf.ethz.ch
ETHZ Computational Laboratory (CoLab), Institute of Computational Science (ICoS),
Swiss Federal Institute of Technology, 8092 Zürich, Switzerland

Abstract

This paper investigates σ-self-adaptation for real valued evolutionary algorithms on linear fitness functions. We identify the step-size logarithm $\log\sigma$ as a key quantity to understand strategy behavior. Knowing the bias of mutation, recombination, and selection on $\log\sigma$ is sufficient to explain σ-dynamics and strategy behavior in many cases, even from previously reported results on non-linear and/or noisy fitness functions. On a linear fitness function, if intermediate multi-recombination is applied on the object parameters, the i-th best and the i-th worst individual have the same σ-distribution. Consequently, the correlation between fitness and step-size σ is zero. Assuming additionally that σ-changes due to mutation and recombination are unbiased, then σ-self-adaptation enlarges σ if and only if $\mu < \lambda/2$, given (μ, λ)-truncation selection. Experiments show the relevance of the given assumptions.

Keywords

Evolutionary algorithms, evolution strategy, self-adaptation, linear fitness function.

1 Introduction

Mutative strategy parameter control, often denoted as *self-adaptation* (SA) (Schwefel, 1995), is a widely used method to adjust strategy parameters in evolutionary algorithms (EAs). In particular in evolution strategies (ESs) and in evolutionary programming, self-adaptation is regarded as a key feature (Bäck and Schwefel, 1993). In the concept of mutative strategy parameter control, the strategy parameters undergo an evolutionary process of reproduction and selection, similar to the object parameters. Usually recombination and mutation operators are applied to the strategy parameters first. The newly generated strategy parameter values are typically used to determine the probability distribution for the object parameter mutation of the same individual. Consider for example the following generation of new search points. Given object parameter $\boldsymbol{x}^{(g)} \in \mathbb{R}^n$ and strategy parameter $\theta^{(g)} \in \mathbb{R}$ at generation g, we have for each descendant $k = 1, \ldots, \lambda$

$$\theta_k^{(g+1)} = \theta^{(g)} + Y_k(\tau) \quad (1)$$

$$\boldsymbol{x}_k^{(g+1)} = \boldsymbol{x}^{(g)} + \gamma\left(\theta_k^{(g+1)}\right)\mathcal{N}_k(\mathbf{0}, \mathbf{I}) \quad , \quad (2)$$

where $Y_k(\tau) \in \mathbb{R}$ denotes a random number with variance τ^2, the function $\gamma : \mathbb{R} \to \mathbb{R}$ is monotonic, and $\mathcal{N}(\mathbf{0}, \mathbf{I}) \in \mathbb{R}^n$ denotes a normally distributed random vector with zero mean and identity $\mathbf{I}$ as covariance matrix. The strategy parameter θ determines, via $\gamma(\theta)$, the scaling factor for the mutation of the object parameter $\boldsymbol{x}$, and therefore the expected step length.

The function γ is most often chosen to be the exponential function, as in evolution strategies (Rechenberg, 1994; Schwefel, 1995) and often in evolutionary programming (Yao et al., 1999). In meta-evolutionary programming, γ was chosen to be the square root, where Y is bounded to ensure $\theta > 0$, and τ becomes a function of θ (Bäck and Schwefel, 1993).[1] If not stated otherwise we assume γ to be the exponential function in the following.

To avoid an a priori bias on the strategy parameter change, the distribution of $Y(\tau)$ usually has zero mean and zero median. Mostly the normal distribution $Y(\tau) \sim \mathcal{N}(0, \tau^2)$ is used (Schwefel, 1995), or a symmetric, discrete two-point distribution, e.g. ± 0.26 (Rechenberg, 1994, p. 48). Given symmetry, zero mean, and variance τ^2 for Y, the specific distribution is, to our experience, of secondary relevance. Consequently, our discussion does not depend on the specific distribution chosen for mutating θ.

The selection of descendents is solely based on the fitness of the object parameter values $\boldsymbol{x}_k^{(g+1)}, k = 1, \ldots, \lambda$. Therefore, selection of strategy parameters, here $\theta_k^{(g+1)}$, is indirect and stochastic: it is based on the probabilistic connection between the strategy parameters (here the covariance matrix $\gamma(\theta_k)^2\mathbf{I}$) and a given realization of the object parameter vector. A better strategy parameter setting can, by chance, result in a worse object parameter setting and vice versa. Hence stochastic fluctuations of the strategy parameters will occur. In (1) only a single strategy parameter θ is in use and stochastic fluctuations are usually unproblematic. In a more general scheme we can have $\theta \in \mathbb{R}^n$. Then, $\gamma(\theta)$ results in a diagonal matrix and defines a different scaling for each variable. In this case, stochastic fluctuations of elements of θ are problematic whenever an element becomes small and changes of this element do not produce selection relevant differences in the object parameter $\boldsymbol{x}$ anymore.[2] Because the size of stochastic fluctuations can be scaled down by an increasing parent number, this happens, roughly speaking, only if the number of adapted strategy parameters exceed the parent population size.

This problem is related to the mutation strength τ of the mutation of the strategy parameters. Small mutation strengths, i.e., small changes of the strategy parameters, can result in virtually selection-irrelevant changes. Large mutation strengths result in large fluctuations of the strategy parameters. Both can lead to a failure of the strategy. Again, this problem becomes particularly relevant if a larger number of parameters is adapted. The problem can be approached by choosing an appropriately large population size, e.g. together with intermediate multi-recombination, or by derandomization (Ostermeier et al., 1994; Hansen and Ostermeier, 2001).

Another fundamental problem connected to self-adaptation regards the discrepancy between the improvement of a single individual versus the improvement of the whole population. On the one hand, selection accounts for a high fitness of the *single* individual. On the other hand, an optimal strategy parameter setting must take into account the fitness gain of the whole population. These objectives can be opposed: Consider the $(\mu/\mu_\mathrm{I}, \lambda)$-ES[3] on the sphere model, $f(\boldsymbol{x}) = \sum_{i=1}^n x_i^2$. The optimal step length is approximately proportional to μ (assuming $\lambda \ll n$) (Rechenberg, 1994; Beyer, 2001). The σ-self-adaptation (σ SA) accounts for selection of single individuals and leads to a nearly optimal step-size σ for $\mu = 1$. But, σ SA cannot account for parent

[1] Here $\theta_k^{(g)}$ instead of $\theta_k^{(g+1)}$ is used in (2). This can only be feasible if different $\theta_k^{(g)}$ are used in (2) and accordingly in (1).

[2] Schwefel (1995) observes the same effect, referred to as *overadaptation*, and states, in contrast to our hypothesis, that elements become small because selection is favorable to individuals with a single small element.

[3] The algorithm is outlined in Section 5 as $\mathrm{ES}_{\mathrm{I}\ldots}$.

number μ, and the selected step-sizes are roughly independent of μ.[4] Consequently, for large μ the step-sizes adapted by σ SA are (far) too small. The problem is less pronounced with dominant recombination of the object parameters and the problem can be alleviated using a biased step-size changing operator (see below).[5] The problem can be solved using competing populations rather than competing individuals or by cumulative path length control (Hansen and Ostermeier, 2001, Eqs. (16) and (17), where $\boldsymbol{B} = \boldsymbol{D} = \mathbf{I}$).

This paper addresses yet another problem concerning the link between selection rank (fitness) and strategy parameter setting. It seems plausible that a better object parameter setting (in terms of fitness, determining the selection rank) is connected with a better expected strategy parameter setting. This plausible conjecture turns out to be wrong in general. We show that a missing link between fitness and strategy parameter quality can lead to a failure even on a linear fitness function.

We will consider the following, simple optimization problem. Maximize the (affine) linear fitness function $f_{\text{linear}} : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{x} \mapsto f_0 + \langle \boldsymbol{v}, \boldsymbol{x} \rangle = f_0 + \sum_{i=1}^n v_i x_i$, where the constants $f_0 \in \mathbb{R}$ and $\boldsymbol{v} \in \mathbb{R}^n$, and $\boldsymbol{v} \neq \mathbf{0}$. In a simple case, $f_0 = 0$ and $\boldsymbol{v}$ is equal to the first unit vector and $f_{\text{linear}}(\boldsymbol{x}) = x_1$, as used in the simulations. The results in this paper hold for any $f_0 \in \mathbb{R}$ and $\boldsymbol{v} \neq \mathbf{0}$. Furthermore, we will consider the commonly used σ-self-adaptation (σ SA) of one global step-size. To derive the equations of σ SA from (1) and (2) we set $\gamma = \exp$ and $\sigma = \gamma(\theta) \in \mathbb{R}_+$. The well-known σ SA method then reads

$$\begin{aligned} \sigma_k^{(g+1)} &= \sigma^{(g)} \exp\left(Y_k(\tau)\right) && (3) \\ \boldsymbol{x}_k^{(g+1)} &= \boldsymbol{x}^{(g)} + \sigma_k^{(g+1)} \mathcal{N}_k(\mathbf{0}, \mathbf{I}) \enspace . && (4) \end{aligned}$$

Remark that (3) and (4) are only a specific way to rewrite (1) and (2).

The σ SA of one global step-size is a commonly used method for the adaptation of the overall distribution variance of the mutation distribution. What is the interest in f_{linear} when considering σ SA of one global step-size? When step-size σ (and hence population diversity) decreases to zero, from the algorithms viewpoint any smooth fitness function f becomes (affine) linear, that is $f \to f_{\text{linear}}$ for certain f_0 and $\boldsymbol{v}$, when $\sigma \to 0$. Therefore, a linear fitness environment must be regarded as the strongest realistic indication for enlarging the step-size. Hence, we have a necessary (minimal) demand on any control mechanism for the global step-size: to enlarge the step-size on f_{linear}. This paper investigates when and why σ SA increases the step-size on f_{linear} and therefore meets the minimal demand.

The next section describes the initial motivation for this paper. In Section 3 the general evolutionary algorithm is outlined and the assumptions for the theoretical results are described. Section 4 presents the theoretical results. Section 5 describes the specific algorithms used for the experiments. In Section 6 experiments confirm the role of the assumptions. Section 7 gives a summary and a conclusion.

[4] Actually, the second best individual has an even smaller expected step-size than the best individual (Hansen, 1998).

[5] Even though the target step-size of σ SA is smaller than the optimal step-size, the dynamics of the optimal step-size when approaching the optimum can have a remarkable influence. If the realized change rate of σ, controlled by parameter τ, is small enough, the changing optimal step-size approaches the realized step-size and can even become smaller (as becomes obvious for $\tau = 0$). This effect can explain the results reported by Grünz and Beyer (1999) but, for a reasonable parameter setting of τ, will be observed only on functions where the optimum can be approached fast.

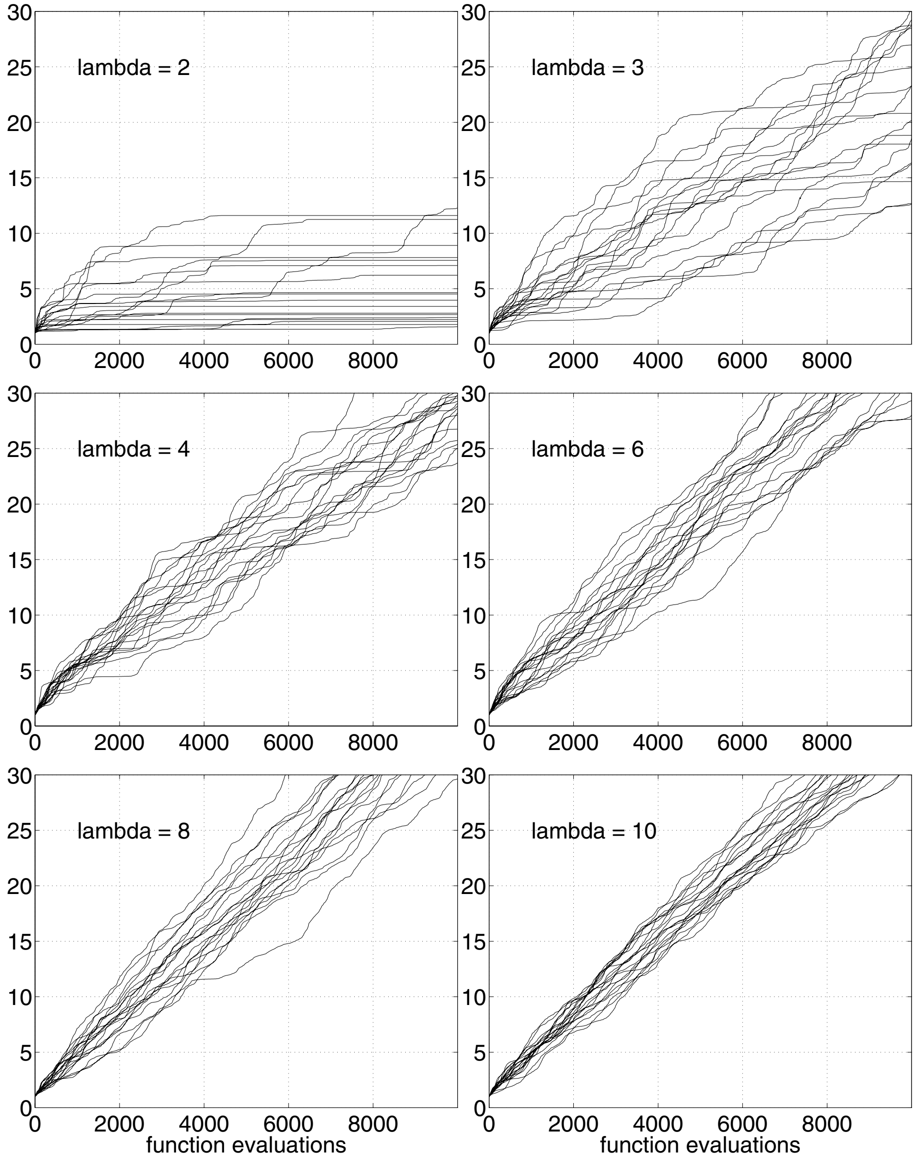

Figure 1: $\lg(f_{\text{linear}})$ versus number of function evaluations. $(1,\lambda)$-σ SA-ES for $\lambda = 2; 3; 4; 6; 8; 10$, problem dimension $n = 10$, 19 runs per figure.

2 Motivation: The $(1,2)$-σ SA-ES Fails

We start from a somewhat surprising observation from experiments on f_{linear}, originally brought up in Ostermeier (1997, p. 9ff). Fig. 1 shows log-fitness curves of the $(1,\lambda)$-σ SA-ES, where $\lambda = 2; 3; 4; 6; 8; 10$. (The algorithm is outlined in Section 5, where $\langle\sigma\rangle_k^{(g)} = \sigma_{1:\lambda}^{(g)}$ in (16) and $\langle\boldsymbol{x}\rangle_k^{(g)} = \boldsymbol{x}_{1:\lambda}^{(g)}$ in (17), where $1:\lambda$ denotes the index of the best individual). With constant step-size, i.e., without σ-self-adaptation, theoretical results (Schwefel, 1995; Hansen et al., 1995) as well as experiments (Herdy, 1993) reveal that $\lambda = 2$ *and* $\lambda = 3$ are the optimal choice on f_{linear}. The σ SA-ES cannot realize this result.

The performance of the $(1,4)$-σ SA-ES is about 1.5 times faster than the performance of the $(1,3)$-σ SA-ES and even the σ SA-ES with $\lambda = 6; 8; 10$ is clearly faster than with $\lambda = 4$. This result is particularly surprising, because the parameter $\tau = 1/\sqrt{2n}$ is chosen identical in all cases. Because τ determines the possible step-size increase *per generation* the set up is favorable to smaller λ. Even more remarkable is the degradation of the $(1,2)$-σ SA-ES. For $\lambda = 2$, the log-fitness gain over time is clearly sub-linear.

The reason for the performance degradation of the $(1,2)$-σ SA-ES can be observed in Fig. 2. The log-σ plots from the same runs reveal a qualitative difference between the $(1,2)$-σ SA-ES on the one hand, and the strategies where $\lambda > 2$ on the other hand. The former shows an unbiased random walk of $\log(\sigma)$, while for the latter $\log(\sigma)$ increases linearly with time.

Why can we observe a random walk of $\log(\sigma)$? One might argue that the additional parameter σ to be adapted needs a larger population size. For a slight degradation this argument might be acceptable. But it is unsatisfactory for the observed σ-dynamics for $\lambda = 2$. Also, a large number of object parameters does not generally demand a large population size. For $\mu = 1$ optimal λ-values are usually small, even for large search space dimensions. On the sphere model and on f_{linear}, even the $(1,2)$-ES works well for search space dimensions of at least up to 1000 if, e.g., cumulative path length control (Hansen and Ostermeier, 1996) is applied to adapt the global step-size.

It was already concluded in Ostermeier (1997, p. 11) that the $(1,2)$-σ SA-ES keeps $\log(\sigma)$ constant on a linear function. We can explain this by simple symmetry considerations. In a $(1,2)$-σ SA-ES three different selection situations can occur: 1) one offspring is better than the parent while the other is worse. In this situation selection takes place regardless of σ. 2) both offspring are better than the parent. Selection will favor the larger σ in this situation. 3) both offspring are worse than the parent. Selection will favor the smaller σ now. From the symmetry of the offspring distribution and of f_{linear} follows that in situation 3) exactly those σ are selected which are disregarded in situation 2). Subsuming 1)-3) implies that the *selection* on f_{linear} does not change the distribution of σ in the $(1,2)$-σ SA-ES! Therefore, with (3) we can derive

$$\begin{aligned}
\mathrm{E}\left[\log \sigma^{(g+1)} \,\middle|\, \sigma^{(g)}\right] &= \mathrm{E}\left[\log \sigma_k^{(g+1)} \,\middle|\, \sigma^{(g)}\right] \\
&= \mathrm{E}\left[\log\left(\sigma^{(g)} \exp\left(Y_k(\tau)\right)\right) \,\middle|\, \sigma^{(g)}\right] \\
&= \log\left(\sigma^{(g)}\right) + \mathrm{E}\left[\log\left(\exp(Y_k(\tau))\right)\right] \\
&= \log\left(\sigma^{(g)}\right) + \mathrm{E}[Y_k(\tau)] \qquad (5)
\end{aligned}$$

for the $(1,2)$-σ SA-ES on f_{linear}. Because Y has zero mean, for $\lambda = 2$ we must expect an *unbiased* random walk of $\log(\sigma)$ *on a linear fitness function*. This random walk of $\log \sigma$ can be observed in Fig. 2, upper left, where the normal distribution $\mathcal{N}(0, \tau^2)$ for Y in (3) is used.

Beyer and Deb (2001) argue that the "constant σ claim" for a linear function is wrong because Ostermeier argues w.r.t. the *probability of increase* of σ rather than w.r.t. the *expectation value* of σ. Based on Beyer (1996), they show that the $(1,2)$-σ SA-ES *exponentially increases* σ and σ^2 on f_{linear} (Beyer and Deb, 2001). In fact, *both* assertions are correct: The "constant σ claim" by Ostermeier (1997) holds for the expectation of $\log(\sigma)$, as shown above, while the expectation of σ and σ^2 increase exponentially: from $\mathrm{E}\left[\log \sigma^{(g+1)} \middle| \sigma^{(g)}\right] = \log\left(\sigma^{(g)}\right)$, according to (5), we multiply with $\alpha \neq 0$ and derive

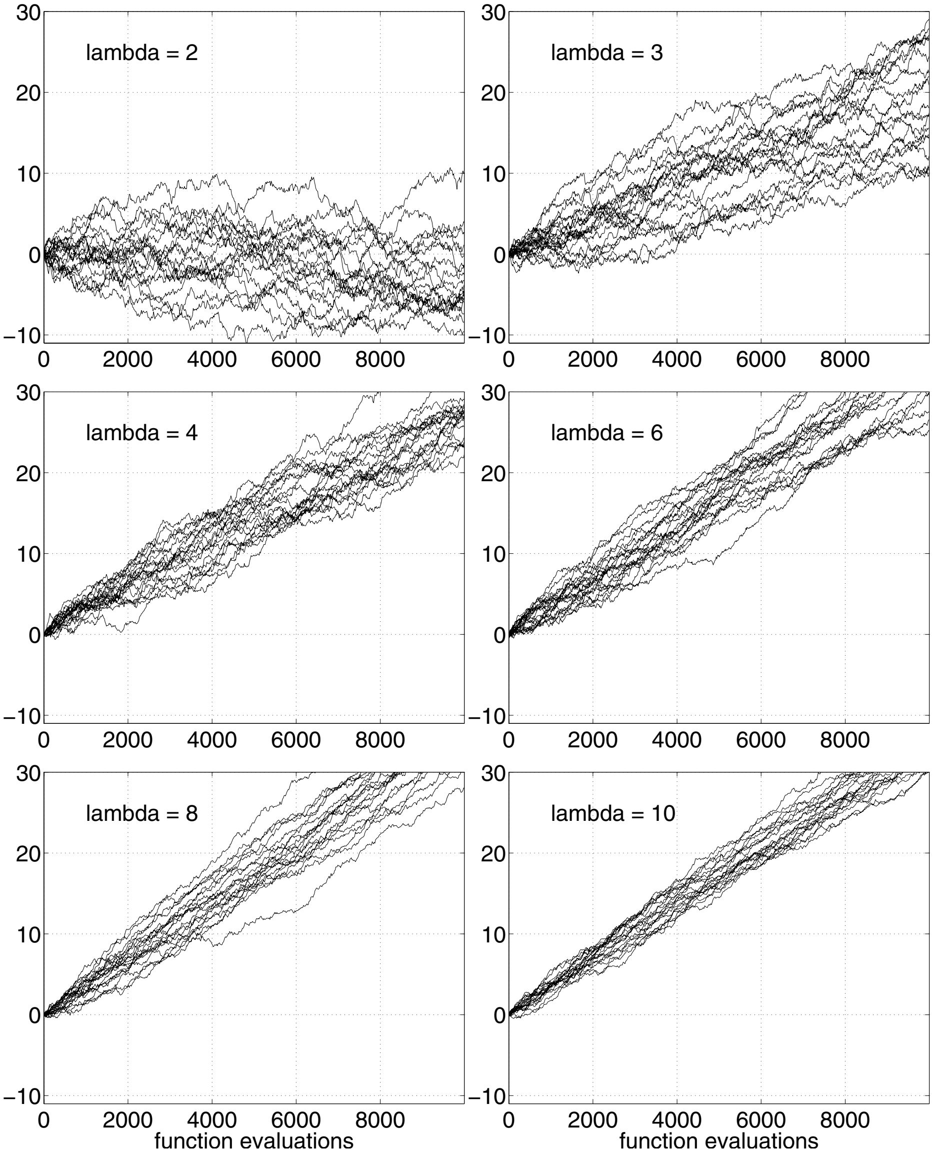

Figure 2: $\lg(\sigma)$ versus number of function evaluations. $(1,\lambda)$-σ SA-ES for $\lambda = 2; 3; 4; 6; 8; 10$, problem dimension $n = 10$, 19 runs per figure. Shown are the same runs as in Fig. 1.

with Jensen's inequality for all $\alpha \neq 0$ that

$$\log\Big(\mathrm{E}\big[\,\big(\sigma^{(g+1)}\big)^{\alpha}\,\big]\Big) > \mathrm{E}\big[\log\Big(\big(\sigma^{(g+1)}\big)^{\alpha}\Big)\,\big] = \log\Big(\big(\sigma^{(g)}\big)^{\alpha}\Big)$$

and therefore $\mathrm{E}\big[\,\big(\sigma^{(g+1)}\big)^{\alpha}\,\big] > \big(\sigma^{(g)}\big)^{\alpha}$.[6]

[6] Thanks to Anne Auger who pointed out this implication to me.

We summarize our observations in

Proposition 1 *On f_{linear} the σ-distribution in the $(1,2)$-σ SA-ES is identical before and after selection, that is, for $k = 1, 2$*

$$\sigma^{(g+1)} = \sigma_{1:2}^{(g+1)} \sim \sigma_{2:2}^{(g+1)} \sim \sigma_k^{(g+1)} = \sigma^{(g)} \exp(Y_k) \ . \tag{6}$$

Therefore, if $\mathrm{E}\big[Y_k\big] = 0$, the step-size σ is unbiased: for all $\alpha \neq 0$

$$\mathrm{E}\Big[\log\Big(\Big(\sigma^{(g+1)}\Big)^{\alpha}\Big)\Big|\sigma^{(g)}\Big] = \log\Big(\Big(\sigma^{(g)}\Big)^{\alpha}\Big) \quad , \tag{7}$$

but

$$\mathrm{E}\Big[\Big(\sigma^{(g+1)}\Big)^{\alpha}\Big|\sigma^{(g)}\Big] > \Big(\sigma^{(g)}\Big)^{\alpha} \quad . \tag{8}$$

Beyer and Deb (2001) oppose the evolution of $\log(\sigma)$ to the evolution of the *real* σ. Even though it is unclear why $\log(\sigma)$ can be considered as being *less real* than σ, the question remains whether to look at σ, σ^2, or $\log(\sigma)$ is more appropriate. For the case of the $(1,2)$-σ SA-ES, Proposition 1, complemented by the experimental results from Fig. 1 and Fig. 2, answers this question. On the one hand, $\log(\sigma)$ being constant reflects that the σ distribution remains identical before and after selection. On the other hand, $\log(\sigma)$ being constant is directly related to the collapse of fitness gain on f_{linear} and can predict the performance degradation of the $(1,2)$-σ SA-ES. In contrast, the exponential increase of σ and/or σ^2 turns out to be not sufficient to predict a satisfactory σ-dynamics on f_{linear}. Postulate 3 from Beyer and Deb (2001), that the expected population variance should increase exponentially, is not sufficient. The following postulate is more appropriate.

Postulate 1 *On f_{linear}, an evolutionary algorithm should increase the expected logarithm of the population variance linearly in time.*

What should be the conclusion from these observations? We would expect from any reasonable σ-control algorithm on a linear fitness function a linear increase of $\log(\sigma)$ over the time. The reason is threefold. First, in many cases f_{linear} is the best local approximation of the fitness function if σ is (too) small. Second, on f_{linear} a non-increasing $\log(\sigma)$ does not achieve a linear log-fitness gain. Third, σ-control *can* achieve the linear increase of $\log(\sigma)$ and of the log-fitness easily. (This was validated by the above results for $\lambda > 2$). The $(1,2)$-σ SA-ES fails to meet the expectation to increase $\log(\sigma)$.

The remainder of this paper is related to the question whether this failure carries over to other strategy variants. Is it a singular phenomenon of the $(1,2)$-σ SA-ES? Does it generalize to any $(\mu, 2\mu)$-σ SA-ES as stated in Ostermeier (1997, p. 11)? What if $\mu > \lambda/2$? If applied, does recombination play a role? These questions will be addressed in the following sections. It will be shown that, *depending on the applied mutation and recombination mechanism*, the phenomenon generalizes to the $(\mu, 2\mu)$-σ SA-ES. More formally, we will show that under certain assumptions the $(\mu, 2\mu)$-selection does not influence the expectation of certain transformations of σ. Slightly more restrictive assumptions are necessary to get conclusions for $\mu \neq \lambda/2$.

Prior to presenting the theoretical results, we need to define the Evolutionary Algorithm.

3 The Evolutionary Algorithm

We formulate a real coded evolutionary algorithm with (μ, λ)-truncation selection and with σ SA, i.e., mutative strategy parameter control of one global step-size σ. The new step-size $\mathfrak{S}_k$ and the new search point X_k are generated from the set of selected individuals, $(\boldsymbol{x}, \sigma)_{i:\lambda|i=1,\ldots,\mu}$, comprising object parameters $\boldsymbol{x}_{i:\lambda}$ and step-sizes $\sigma_{i:\lambda}$, where $i : \lambda$ denotes the index of the i-th best individual. For each descendant $k = 1, \ldots, \lambda$ we have independently

$$\mathfrak{S}_k^{\sigma_{i:\lambda|i=1,\ldots,\mu}} = \mathrm{Mut}_\sigma\left(\mathrm{Rec}_\sigma\left\{\sigma_{i:\lambda|i=1,\ldots,\mu}\right\}\right) \tag{9}$$

$$X_k^{(\boldsymbol{x},\sigma)_{i:\lambda|i=1,\ldots,\mu}} = \mathrm{Rec}_{\boldsymbol{x}}\left\{\boldsymbol{x}_{i:\lambda|i=1,\ldots,\mu}\right\} + \mathfrak{S}_k^{\sigma_{i:\lambda|i=1,\ldots,\mu}} \cdot Z_k \ , \tag{10}$$

where $\mathrm{Rec}_\sigma\{.\}$ and $\mathrm{Rec}_{\boldsymbol{x}}\{.\}$ denote recombination mechanisms, $\mathrm{Mut}_\sigma(.)$ denotes mutation of a step-size and $Z_k \in \mathbb{R}^n, k = 1, \ldots, \lambda$, are independent and identically distributed random vectors for the mutation of the object parameter vector. An example for $\mathrm{Rec}_{\boldsymbol{x}}\{.\}$ is *intermediate multi-recombination*, where $\mathrm{Rec}_{\boldsymbol{x}}\{\boldsymbol{x}_{i:\lambda|i=1,\ldots,\mu}\} = \frac{1}{\mu}\sum_{i=1}^{\mu} \boldsymbol{x}_{i:\lambda}$.

In the following $\mathfrak{S}_k^{\sigma_{i:\lambda|i=1,\ldots,\mu}} \in \mathbb{R}_+$ and $X_k^{(\boldsymbol{x},\sigma)_{i:\lambda|i=1,\ldots,\mu}} \in \mathbb{R}^n$ denote random variables whose distributions depend on the superscript values. The σ_k and $\boldsymbol{x}_k$ denote realized values of the previous generation. The number $i : \lambda$ denotes the index of the i-th best individual and the set $\{i : \lambda | i = 1, \ldots, \mu\}$ is the index set of the μ selected (best) individuals.

For the theory section of this paper two assumptions on the EA are introduced. Examples of mutation and recombination operators which satisfy these assumptions are formulated in the following.

The first assumption is concerned with the symmetry of the distribution of individuals.

Assumption 1 (Symmetry) *Recombination and mutation of object parameter vectors $\boldsymbol{x}$ in (10) yield a point-symmetrical distribution of descendants. That is, there exists a point $\boldsymbol{p} \in \mathbb{R}^n$, such that*

$$X_k^{(\boldsymbol{x},\sigma)_{i:\lambda|i=1,\ldots,\mu}} - \boldsymbol{p}$$

and

$$-\left(X_k^{(\boldsymbol{x},\sigma)_{i:\lambda|i=1,\ldots,\mu}} - \boldsymbol{p}\right)$$

are identically distributed, and also the distribution of $\mathfrak{S}_k^{\sigma_{i:\lambda|i=1,\ldots,\mu}} | X_k^{(\boldsymbol{x},\sigma)_{i:\lambda|i=1,\ldots,\mu}}$ is identical for the symmetry pair $X_k^{(\boldsymbol{x},\sigma)_{i:\lambda|i=1,\ldots,\mu}} = \boldsymbol{p} + \boldsymbol{a}$ and $X_k^{(\boldsymbol{x},\sigma)_{i:\lambda|i=1,\ldots,\mu}} = \boldsymbol{p} - \boldsymbol{a}$ for all $\boldsymbol{a} \in \mathbb{R}^n$.

Assumption 1 leads to a point-symmetrical distribution of the complete population before selection.

The following proposition gives a sufficient condition to satisfy the symmetry Assumption 1.

Proposition 2 *The symmetry Assumption 1 is satisfied by (9) and (10) if $\mathrm{Rec}_{\boldsymbol{x}}\{\boldsymbol{x}_{i:\lambda|i=1,\ldots,\mu}\} = \boldsymbol{p}$ for all offspring $k = 1, \ldots, \lambda$ of one generation step, and Z_k is point-symmetrical w.r.t. $\mathbf{0}$, and Z_k is independent of $\mathfrak{S}_k^{\sigma_{i:\lambda|i=1,\ldots,\mu}}$. In particular an evolution strategy with intermediate multi-recombination, $(\mu/\mu_{\mathrm{I}}, \lambda)$-ES, satisfies the symmetry assumption.*

Without intermediate multi-recombination of the object parameters, symmetry cannot be taken for granted, because the selected population $\boldsymbol{x}_{1:\lambda}, \ldots, \boldsymbol{x}_{\mu:\lambda}$ is non-symmetrical in general.

The second assumption refers to the step-size variation. It concurrently defines a) what we refer to as *unbiased* step-size and b) an appropriate method to measure an *expected* step-size of the EA.

Assumption 2 (σ-Stationarity) *There exists a (strictly) monotonically increasing function* $h : \mathbb{R} \to \mathbb{R}$,[7] *for which recombination and mutation leave* $h(\mathfrak{S}_k^{\sigma_{i:\lambda|i=1,\ldots,\mu}})$ *unbiased, that is* $\mathrm{E}\big[h(\mathfrak{S}_k^{\sigma_{i:\lambda|i=1,\ldots,\mu}})\big] = \frac{1}{\mu}\sum_{i=1}^{\mu} h(\sigma_{i:\lambda})$*, for all* $k = 1, \ldots, \lambda$.

Lemma 1 *Let* $h : x \mapsto \log(x)$ *and let the mutation of* σ *be a multiplication with a factor* > 0*, that is, w.l.o.g.* $\mathrm{Mut}_\sigma(\sigma) = \sigma \cdot \exp(Y)$*. Then the biases of mutation and recombination add, that is*

$$\mathrm{E}\big[\log(\mathfrak{S}_k^{\sigma_{i:\lambda|i=1,\ldots,\mu}})\big] = \mathrm{E}\big[Y\big] + \mathrm{E}\big[\log\big(\mathrm{Rec}_\sigma\{\sigma_{i:\lambda|i=1,\ldots,\mu}\}\big)\big] \quad . \tag{11}$$

Proof

$$\begin{aligned}
\mathrm{E}\big[\log(\mathfrak{S}_k^{\sigma_{i:\lambda|i=1,\ldots,\mu}})\big] &= \mathrm{E}\big[\log\big(\mathrm{Mut}_\sigma\big(\mathrm{Rec}_\sigma\{\sigma_{i:\lambda|i=1,\ldots,\mu}\}\big)\big)\big] \\
&= \mathrm{E}\big[\log\big(\exp(Y)\cdot\mathrm{Rec}_\sigma\{\sigma_{i:\lambda|i=1,\ldots,\mu}\}\big)\big] \\
&= \mathrm{E}\big[\log(\exp(Y)) + \log\big(\mathrm{Rec}_\sigma\{\sigma_{i:\lambda|i=1,\ldots,\mu}\}\big)\big] \\
&= \mathrm{E}[Y] + \mathrm{E}\big[\log\big(\mathrm{Rec}_\sigma\{\sigma_{i:\lambda|i=1,\ldots,\mu}\}\big)\big] \quad .
\end{aligned}$$

□

Two examples for the operators $\mathrm{Rec}_\sigma\{.\}$ and $\mathrm{Mut}_\sigma(.)$ are given which satisfy the σ-stationarity Assumption 2 for a certain function h.

Proposition 3 *The σ-stationarity Assumption 2 is satisfied by* $h : x \mapsto \log(x)$*, if* $\mathrm{Mut}_\sigma(\sigma) = \sigma\cdot\exp(Y_k)$*, where* Y_k *is a random number with zero mean, e.g.* Y_k *is* $(0,1)$*-normally distributed, and*

1. *no recombination on the step-size is applied, that is* $\mathrm{Rec}_\sigma\{\sigma_{i:\lambda|i=1,\ldots,\mu}\} := \sigma_j$*, where* j *is uniformly distributed in* $\{i : \lambda | i = 1, \ldots, \mu\}$*, or*

2. *"geometric" recombination on the step-sizes is applied, that is* $\mathrm{Rec}_\sigma\{\sigma_{i:\lambda|i=1,\ldots,\mu}\} = \sqrt[\mu]{\prod_{i=1}^{\mu}\sigma_{i:\lambda}}$ *in the case of* μ *recombinants.*

Proof With (11) from Lemma 1 and $\mathrm{E}[Y_k] = 0$ we derive $\mathrm{E}\big[h(\mathfrak{S}_k^{\sigma_{i:\lambda|i=1,\ldots,\mu}})\big] = 0 + \mathrm{E}\big[\log\big(\mathrm{Rec}_\sigma\{\sigma_{i:\lambda|i=1,\ldots,\mu}\}\big)\big]$. From the definition of $\mathrm{Rec}_\sigma\{\sigma_{i:\lambda|i=1,\ldots,\mu}\}$ in both cases we derive $\mathrm{E}\big[\log\big(\mathrm{Rec}_\sigma\{\sigma_{i:\lambda|i=1,\ldots,\mu}\}\big)\big] = \frac{1}{\mu}\sum_{i=1}^{\mu}\log\sigma_{i:\lambda} = \frac{1}{\mu}\sum_{i=1}^{\mu}h(\sigma_{i:\lambda})$. □

Proposition 4 *The σ-stationarity Assumption 2 is satisfied by the identity function* $h : x \mapsto x$*, if* $\mathrm{Mut}_\sigma(\sigma) := \sigma + Y_k$*, where* Y_k *is a random number with zero mean,*[8] *and*

1. *no recombination on the step-size is applied, that is* $\mathrm{Rec}_\sigma\{\sigma_{i:\lambda|i=1,\ldots,\mu}\} := \sigma_j$*, where* j *is uniformly distributed in* $\{i : \lambda | i = 1, \ldots, \mu\}$*, or*

2. *"arithmetic" recombination on the step-sizes is applied, that is* $\mathrm{Rec}_\sigma\{\sigma_{i:\lambda|i=1,\ldots,\mu}\} := \frac{1}{\mu}\sum_{i=1}^{\mu}\sigma_{i:\lambda}$.

[7] For $h \equiv \mathrm{const}$ the theoretical results can be formally applied but become meaningless.

[8] This model allows, and demands, negative values for σ. With an object parameter mutation symmetrical around zero this is algorithmically insignificant. Consequently $\mathrm{E}[\sigma]$ does not equal $\mathrm{E}[|\sigma|]$ in general, that is $\mathrm{E}[\sigma]$ may not reflect the step length anymore.

Proof

$$\begin{aligned} \mathrm{E}\big[h(\mathfrak{S}_k^{\sigma_{i:\lambda|i=1,\ldots,\mu}})\big] &= \mathrm{E}\big[\mathfrak{S}_k^{\sigma_{i:\lambda|i=1,\ldots,\mu}}\big] \\ &= \mathrm{E}\big[\mathrm{Mut}_\sigma\left(\mathrm{Rec}_\sigma\{\sigma_{i:\lambda|i=1,\ldots,\mu}\}\right)\big] \\ &= \mathrm{E}\big[Y_k + \mathrm{Rec}_\sigma\{\sigma_{i:\lambda|i=1,\ldots,\mu}\}\big] \\ &= 0 + \mathrm{E}\big[\mathrm{Rec}_\sigma\{\sigma_{i:\lambda|i=1,\ldots,\mu}\}\big] \quad , \end{aligned}$$

and from the definition of $\mathrm{Rec}_\sigma\{\sigma_{i:\lambda|i=1,\ldots,\mu}\}$ in both cases we have $\mathrm{E}\big[\mathrm{Rec}_\sigma\{\sigma_{i:\lambda|i=1,\ldots,\mu}\}\big] = \frac{1}{\mu}\sum_{i=1}^{\mu}\sigma_{i:\lambda}$. □

In contrast to geometric recombination, arithmetic recombination of step-sizes introduces a bias on σ, given $h = \log$.

Proposition 5 *For $h : x \mapsto \log(x)$, the arithmetic recombination of σ, where $\mathrm{Rec}_\sigma\{\sigma_{i:\lambda|i=1,\ldots,\mu}\} := \frac{1}{\mu}\sum_{i=1}^{\mu}\sigma_{i:\lambda}$, does not satisfy Assumption 2 and biases σ towards increase.*

Proof Because $\mu \geq 2$ and σ obeys a density, we can assume that $\sigma_{i:\lambda} \neq \sigma_{j:\lambda}$ for some $i, j \leq \mu$. Therefore

$$\begin{aligned} \mathrm{E}\big[h\left(\mathrm{Rec}_\sigma\{\sigma_{i:\lambda|i=1,\ldots,\mu}\}\right)\big] &= \mathrm{E}\left[\log\left(\frac{1}{\mu}\sum_{i=1}^{\mu}\sigma_{i:\lambda}\right)\right] \\ &= \log\left(\frac{1}{\mu}\sum_{i=1}^{\mu}\sigma_{i:\lambda}\right) \\ &> \frac{1}{\mu}\sum_{i=1}^{\mu}\log\sigma_{i:\lambda} \\ &= \frac{1}{\mu}\sum_{i=1}^{\mu}h\left(\sigma_{i:\lambda}\right) \end{aligned}$$

□

This bias, introduced by arithmetic recombination of step-sizes, has important consequences on the strategy behavior, as we will see below.

All theoretical results in this paper are based on Assumption 1. Some results also require Assumption 2 (and hold for any function h satisfying Assumption 2). The experiments in Section 6 reveal the relevance of these assumptions in different algorithms.

4 Theoretical Results

From the symmetry Assumption 1 we derive the fundamental theorem that the step-size distributions of the i-th best and the i-th worst individual on f_{linear} are identical.

Theorem 1 *Given Assumption 1, the i-th best and the i-th worst individual have the same distribution of the step-size on f_{linear}, that is, for $i = 1, \ldots, \lambda$*

$$\mathfrak{S}_{i:\lambda}^{(\boldsymbol{x},\sigma)_{i:\lambda|i=1,\ldots,\mu}} \sim \mathfrak{S}_{\lambda-i+1:\lambda}^{(\boldsymbol{x},\sigma)_{i:\lambda|i=1,\ldots,\mu}} \quad . \tag{12}$$

Proof Because $\mathfrak{S}_k^{\sigma_{i:\lambda|i=1,\ldots,\mu}}|(X_k^{(\boldsymbol{x},\sigma)_{i:\lambda|i=1,\ldots,\mu}} = \boldsymbol{p} + \boldsymbol{a})$ is distributed like $\mathfrak{S}_k^{\sigma_{i:\lambda|i=1,\ldots,\mu}}|(X_k^{(\boldsymbol{x},\sigma)_{i:\lambda|i=1,\ldots,\mu}} = \boldsymbol{p} - \boldsymbol{a})$, selection on f_{linear} yields the same distribution for $\mathfrak{S}_{i:\lambda}^{(\boldsymbol{x},\sigma)_{i:\lambda|i=1,\ldots,\mu}}$ as selection on $-f_{\mathrm{linear}}$. Because the i-th best individual on $-f_{\mathrm{linear}}$ is identical to $\lambda - i + 1$-th best individual on f_{linear} the proof is complete. □

Corollary 1 *Given Assumption 1, on f_{linear} the correlation between fitness-value and step-size is zero.*

Proof Let denote $f_0(X_k) = f_{\text{linear}}(X_k^{(\boldsymbol{x},\sigma)_{i:\lambda|i=1,\ldots,\mu}}) - \mathrm{E}\big[f_{\text{linear}}(X_k^{(\boldsymbol{x},\sigma)_{i:\lambda|i=1,\ldots,\mu}})\big]$. Because $X_k^{(\boldsymbol{x},\sigma)_{i:\lambda|i=1,\ldots,\mu}}$ is symmetric, $f_0(X_k)$ is symmetric around zero and according to Assumption 1 the distribution of $\mathfrak{S}_k^{\sigma_{i:\lambda|i=1,\ldots,\mu}} \mid f_0(X_k) = a$ is identical to $\mathfrak{S}_k^{\sigma_{i:\lambda|i=1,\ldots,\mu}} \mid f_0(X_k) = -a$ for all $a \in \mathbb{R}$. Therefore $\mathrm{E}\big[f_0(X_k) \times \big(\mathfrak{S}_k^{\sigma_{i:\lambda|i=1,\ldots,\mu}} - \mathrm{E}\big[\mathfrak{S}_k^{\sigma_{i:\lambda|i=1,\ldots,\mu}}\big]\big)\big] = 0$, q.e.d. □

Theorem 1 and Corollary 1 indicate that on f_{linear} there is no clear cut relation between fitness (of the object parameters) and step-size quality. In the following, we will investigate some consequences of those results. For the remainder of this section, we use an EA which satisfies both assumptions given in the last section.

Consider two identical evolutionary algorithms with different (truncation-) selection schemes. One with (μ, λ)-selection, the other with (μ', λ)-selection. Assuming $\mu' = \lambda - \mu$ the algorithms are denoted by EA_μ and $\mathrm{EA}_{\lambda-\mu}$. We investigate now whether EA_μ and $\mathrm{EA}_{\lambda-\mu}$ increase $h(\sigma)$ on f_{linear}.

Let $\mathfrak{S}_{i:\lambda}^{\mathrm{EA}_\mu}$ and $\mathfrak{S}_{i:\lambda}^{\mathrm{EA}_{\lambda-\mu}}$ denote $\mathfrak{S}_{i:\lambda}^{(\boldsymbol{x},\sigma)_{i:\lambda|i=1,\ldots,\mu}}$ of EA_μ and $\mathrm{EA}_{\lambda-\mu}$, respectively, and let $\mathrm{E}_h[.]$ denote $\mathrm{E}[h(.)]$. We show that the set $\mathfrak{S}_{1\ldots\mu:\lambda}^{\mathrm{EA}_\mu}$ is "opposed" to the set $\mathfrak{S}_{1\ldots\lambda-\mu:\lambda}^{\mathrm{EA}_{\lambda-\mu}}$ on f_{linear}.

Theorem 2 *Assume identical baseline step-size, that is $\frac{1}{\mu}\sum_{i=1}^{\mu} h(\sigma_{i:\lambda})$ in EA_μ to be equal to $\frac{1}{\lambda-\mu}\sum_{i=1}^{\lambda-\mu} h(\sigma_{i:\lambda})$ in $EA_{\lambda-\mu}$. Assume the results of recombination, $\mathrm{Rec}_{\boldsymbol{x}}\{\boldsymbol{x}_{i:\lambda|i=1,\ldots,\mu}\}$ and $\mathrm{Rec}_\sigma\{\sigma_{i:\lambda|i=1,\ldots,\mu}\}$ identically distributed in EA_μ and $EA_{\lambda-\mu}$, respectively, apart from translation in search space. Then on f_{linear}, EA_μ enlarges the expected step-size if and only if $EA_{\lambda-\mu}$ reduces it, and vice versa. That is*

$$\frac{1}{\mu}\sum_{i=1}^{\mu} \mathrm{E}_h\left[\mathfrak{S}_{i:\lambda}^{\mathrm{EA}_\mu}\right] \gtreqless \frac{1}{\mu}\sum_{i=1}^{\mu} h\left(\sigma_{i:\lambda}\right) \gtreqless \frac{1}{\lambda-\mu}\sum_{i=1}^{\lambda-\mu} \mathrm{E}_h\left[\mathfrak{S}_{i:\lambda}^{\mathrm{EA}_{\lambda-\mu}}\right] \tag{13}$$

where $\gtreqless$ means $>$, or $=$, or $<$ holds equally at each occurrence.

Proof With Assumption 2 we have for each $k = 1, \ldots, \lambda$

$$\frac{1}{\mu}\sum_{i=1}^{\mu} h\left(\sigma_{i:\lambda}\right) = \mathrm{E}_h\left[\mathfrak{S}_k^{\mathrm{EA}_\mu}\right] = \frac{1}{\lambda}\sum_{i=1}^{\lambda} \mathrm{E}_h\left[\mathfrak{S}_i^{\mathrm{EA}_\mu}\right] = \frac{1}{\lambda}\sum_{i=1}^{\lambda} \mathrm{E}_h\left[\mathfrak{S}_{i:\lambda}^{\mathrm{EA}_\mu}\right] \quad .$$

The last equation results from rearranging the sum after exchanging sum and expectation. Splitting the last sum yields

$$\begin{aligned} \frac{1}{\mu}\sum_{i=1}^{\mu} \mathrm{E}_h\left[\mathfrak{S}_{i:\lambda}^{\mathrm{EA}_\mu}\right] &\gtreqless \frac{1}{\mu}\sum_{i=1}^{\mu} h\left(\sigma_{i:\lambda}\right) \\ \iff \quad \frac{1}{\mu}\sum_{i=1}^{\mu} h\left(\sigma_{i:\lambda}\right) &\gtreqless \frac{1}{\lambda-\mu}\sum_{i=\mu+1}^{\lambda} \mathrm{E}_h\left[\mathfrak{S}_{i:\lambda}^{\mathrm{EA}_\mu}\right] \quad . \end{aligned} \tag{14}$$

Because the result of recombination is identical for EA_μ and $\mathrm{EA}_{\lambda-\mu}$, descendants are identically distributed in both strategies (beside their selection scheme both algorithms are identical). That is, $\mathfrak{S}_{i:\lambda}^{\mathrm{EA}_\mu}$ has the same distribution as $\mathfrak{S}_{i:\lambda}^{\mathrm{EA}_{\lambda-\mu}}$ and with Theorem 1 we can replace i by $\lambda - i + 1$. Therefore the RHS of (14) equals to $\frac{1}{\lambda-\mu}\sum_{i=\mu+1}^{\lambda} \mathrm{E}_h\left[\mathfrak{S}_{\lambda-i+1:\lambda}^{\mathrm{EA}_{\lambda-\mu}}\right] = \frac{1}{\lambda-\mu}\sum_{i=1}^{\lambda-\mu} \mathrm{E}_h\left[\mathfrak{S}_{i:\lambda}^{\mathrm{EA}_{\lambda-\mu}}\right]$, q.e.d. □

Theorem 2 shows that, under slightly tighter assumptions, only one of the two algorithms EA_μ and $\mathrm{EA}_{\lambda-\mu}$ increase the step-size on f_{linear}. That is, only one of these algorithms work properly on f_{linear}. Sufficient conditions for the assumptions in Theorem 2 can be given.

Proposition 6 *The additional assumption in Theorem 2 on the recombination result* $\mathrm{Rec}_{\boldsymbol{x}}\{\boldsymbol{x}_{i:\lambda|i=1,\ldots,\mu}\}$ *is satisfied by Proposition 2. The additional assumption in Theorem 2 on the recombination result of* $\mathrm{Rec}_\sigma\{\sigma_{i:\lambda|i=1,\ldots,\mu}\}$ *is satisfied by the multi-recombination from Proposition 3, point 2, and from Proposition 4, point 2.*

Finally, we derive from Theorem 2 the answer to the original question, whether our observations from the $(1,2)$-σ SA can be generalized to any μ, where $\lambda = 2\mu$.

Theorem 3 *Given Assumption 1 and 2, the expected step-size of a $(\mu, 2\mu)$-EA on f_{linear} is constant. That is, for $\lambda = 2\mu$*

$$\frac{1}{\mu}\sum_{i=1}^{\mu} \mathrm{E}_h\left[\mathfrak{S}_{i:\lambda}^{(\boldsymbol{x},\sigma)_{i:\lambda|i=1,\ldots,\mu}}\right] = \frac{1}{\mu}\sum_{i=1}^{\mu} h\left(\sigma_{i:\lambda}\right) \quad . \tag{15}$$

Proof Because $\lambda = 2\mu$, EA_μ and $\mathrm{EA}_{\lambda-\mu}$ in Theorem 2 denote the same strategy. Therefore, the assumption of identical recombination result in Theorem 2 is satisfied. We have $\frac{1}{\mu}\sum_{i=1}^{\mu} \mathrm{E}_h\left[\mathfrak{S}_{i:\lambda}^{\mathrm{EA}_\mu}\right] = \frac{1}{\lambda-\mu}\sum_{i=1}^{\lambda-\mu} \mathrm{E}_h\left[\mathfrak{S}_{i:\lambda}^{\mathrm{EA}_{\lambda-\mu}}\right]$, and with (13) follows (15). □

Theorem 3 generalizes (7) from the $(1,2)$-σ SA-ES to any $(\mu, 2\mu)$-EA that satisfies both assumptions from Section 3. Furthermore, Theorem 3 gives even stronger evidence that Theorem 2 is of relevance for any (μ,λ)-EA from Section 3. It seems unlikely that the step-size remains constant for $\mu = \lambda/2$, as shown in Theorem 3, while it increases for $\mu < \lambda/2$ *and* for $\mu > \lambda/2$. This will be confirmed in the following, where a few EA variants are formulated and the predictions of the theoretical results are compared with experimental results.

5 Algorithms Used for Experiments

For our experiments we used the following $(\mu/\rho_{\mathrm{I}},\lambda)$-evolution strategy, where we use a different, more common notation from now on. The transition from generation g to $g+1$ for step-sizes $\sigma \in \mathbb{R}^+$ and object parameter vectors $\boldsymbol{x} \in \mathbb{R}^n$ read for each descendant $k = 1, \ldots, \lambda$

$$\sigma_k^{(g+1)} = \langle\sigma\rangle_k^{(g)} \exp\left(\mathcal{N}_k\left(0,\tau^2\right)\right) \tag{16}$$

$$\boldsymbol{x}_k^{(g+1)} = \langle\boldsymbol{x}\rangle_k^{(g)} + \sigma_k^{(g+1)}\mathcal{N}_k(\mathbf{0},\mathbf{I}) \quad , \tag{17}$$

where $\mathcal{N}_k\left(0,\tau^2\right)$ denotes a normally distributed random number with zero mean and standard deviation $\tau = 1/\sqrt{2n}$. $\mathcal{N}_k(\mathbf{0},\mathbf{I})$ denotes a $(\mathbf{0},\mathbf{I})$-normally distributed random vector, where $\mathbf{I}$ denotes the identity matrix.

By defining the recombination operators, i.e., $\langle\sigma\rangle_k^{(g)}$ and $\langle\boldsymbol{x}\rangle_k^{(g)}$ in different ways the following algorithm variants are specified ($i : \lambda$ denotes the index of the i-th best out of λ individuals):

$\mathrm{ES}_{\mathrm{w/o}}$: ES without recombination ($\rho = 1$). We have that

$$\langle\sigma\rangle_k^{(g)} = \sigma_i^{(g)} \tag{18}$$

$$\langle\boldsymbol{x}\rangle_k^{(g)} = \boldsymbol{x}_i^{(g)}, \tag{19}$$

where i is uniformly distributed in $\{j : \lambda | j = 1, \ldots, \mu\}$, drawn for each $k = 1, ..., \lambda$ independently. This scheme satisfies the σ-stationarity Assumption 2 for $h \equiv \log$ (see Proposition 3). We suspect that the symmetry Assumption 1 is usually not satisfied for two reasons. It seems unlikely for most objective functions that the μ selected individuals of a population are point-symmetrically distributed. Additionally the linkage between object parameter $\boldsymbol{x}$ and σ should compromise the symmetry assumption w.r.t. the distribution of σ.

ES$_{\text{unlink}}$: ES without recombination of $\boldsymbol{x}$ or σ ($\rho = 1$), while $\boldsymbol{x}$ and σ are chosen independently. We have that

$$\begin{aligned} \langle\sigma\rangle_k^{(g)} &= \sigma_i^{(g)} & (20) \\ \langle\boldsymbol{x}\rangle_k^{(g)} &= \boldsymbol{x}_j^{(g)}, & (21) \end{aligned}$$

where i and j are chosen independently uniformly drawn in $\{i : \lambda | i = 1, \ldots, \mu\}$. This mechanism unlinks σ from $\boldsymbol{x}$ and can also be regarded as recombination between $\boldsymbol{x}$ and σ blocks.

This algorithm satisfies the σ-stationarity Assumption 2 for $h \equiv \log$ (see Proposition 3). Again the symmetry Assumption 1 might be violated due to an asymmetrical distribution of the population.

ES$_{\text{IA}}$: ES with intermediate μ/μ-recombination ($\rho = \mu$) of the object parameters and intermediate *arithmetic* μ/μ-recombination of the step-sizes. We have that

$$\begin{aligned} \langle\sigma\rangle_k^{(g)} &= \frac{1}{\mu}\sum_{i=1}^{\mu}\sigma_{i:\lambda}^{(g)} & (22) \\ \langle\boldsymbol{x}\rangle_k^{(g)} &= \frac{1}{\mu}\sum_{i=1}^{\mu}\boldsymbol{x}_{i:\lambda}^{(g)}\ . & (23) \end{aligned}$$

This algorithm satisfies the symmetry Assumption 1 (see Proposition 2). The σ-stationarity Assumption 2 is satisfied (only) for $\mu = 1$ and $h \equiv \log$.[9] For $\mu > 1$ and $h \equiv \log$, the σ-stationarity is not preserved; $\mathrm{E}\big[\log(\sigma)\big]$ increases under the recombination operator (Proposition 5).

ES$_{\text{IG}}$: ES with intermediate μ/μ-recombination ($\rho = \mu$) of object parameters and intermediate *geometric* μ/μ-recombination of the step-sizes. We have that

$$\begin{aligned} \langle\sigma\rangle_k^{(g)} &= \exp\left(\frac{1}{\mu}\sum_{i=1}^{\mu}\log\sigma_{i:\lambda}^{(g)}\right) = \left(\prod_{i=1}^{\mu}\sigma_{i:\lambda}^{(g)}\right)^{\frac{1}{\mu}} & (24) \\ \langle\boldsymbol{x}\rangle_k^{(g)} &= \frac{1}{\mu}\sum_{i=1}^{\mu}\boldsymbol{x}_{i:\lambda}^{(g)}\ . & (25) \end{aligned}$$

According to Propositions 2 and 3 this algorithm satisfies both assumptions from Section 3 for $h = \log$.

[9] This result explains, why for the ES$_{\text{IA}}$ in a noisy environment σ seems to converge to zero only for $\mu = 1$ (Beyer and Sendhoff, 2005).

ES$_{\mathrm{I\,w/o}}$: ES with intermediate μ/μ-recombination ($\rho = \mu$) of the object parameters as in ES$_{\mathrm{IA}}$ and ES$_{\mathrm{IG}}$ (Equation (25)) and without recombination of the step-sizes as in ES$_{\mathrm{w/o}}$ and ES$_{\mathrm{unlink}}$(Equation (20)). According to Propositions 2 and 3 this algorithm satisfies both assumptions from Section 3 like ES$_{\mathrm{IG}}$.

For ES$_{\mathrm{IG}}$ also the additional assumption for Theorem 2 of identical population distribution of EA$_{\mu}$ and EA$_{\lambda-\mu}$ is satisfied.

The described mutation and recombination mechanisms for the object parameter vector $\boldsymbol{x}$ are frequently used in ESs. While the arithmetic mutation operator for step-size σ in (16) is commonly used, the geometric recombination operator in (24) is fairly uncommon. Nevertheless, from a fundamental viewpoint, the geometric mean seems to be the more sensible operator for the following reason: Let, for example, step-sizes $\sigma_1 = 10^{-2}$ and $\sigma_2 = 10^{-4}$ been "tested" already. If one would like to choose a third step-size for testing "in the middle" of σ_1 and σ_2 this would be, naturally, $\sigma_3 = 10^{-3} = \sqrt[2]{\sigma_1\sigma_2}$. Using the arithmetic mean would result in $\sigma_3 = 5.05 \cdot 10^{-3}$—a fairly strange choice: a factor of 2 less than σ_1, but a factor of 50 greater than σ_2. Also the commonly used *mutation* operator reflects the feasibility of the "geometrical approach" to operate with step-sizes. Here, after mutation of σ the numbers $\sigma \cdot \alpha$ and σ/α appear with the same probability for all $\alpha > 0$. The original σ is the geometric mean of the mutated ones. Nevertheless, we will see in the next section that the arithmetic operator can be more useful from a practical point of view.

6 Experimental Results

We perform computer simulations for the following two purposes. First, we want to confirm that the predictions made by the theoretical results can actually be observed in the simulation. This can be done by simulating ES$_{\mathrm{IG}}$ and ES$_{\mathrm{I\,w/o}}$ that satisfy the assumptions made for the theoretical results. These simulations must be in accordance with Theorem 2 and 3. Second, we want to reveal the importance of the assumptions made in Section 3. This is done by simulating strategies that violate one or the other assumption.

We performed simulations with ES$_{\mathrm{IG}}$ and ES$_{\mathrm{I\,w/o}}$ on f_{linear} for various μ and λ. The results are in accordance with the theoretical results: $\sum_{i=1}^{\mu} \mathrm{E}\big[\log(\sigma_{i:\lambda}^{(g+1)})\big]$ equals $\sum_{i=1}^{\mu} \log \sigma_{i:\lambda}^{(g)}$ for $\mu = \lambda/2$, while $\log(\sigma)$ increases for $\mu < \lambda/2$ and $\log(\sigma)$ decreases for $\mu > \lambda/2$. In Figs. 3 (a) and 4, data are given for ES$_{\mathrm{IG}}$ and $\lambda = 20$ and $n = 2$. Expectation values were estimated from data of about 50000 generations. While the results are actually independent of the problem dimension n, they depend on the strategy parameter $\tau(n)$ which simply rescales the y-axis.

In Fig. 3 (a) the expected change of $\lg(\sigma)$ of the i-th best individual in the $(\mu/\mu_{\mathrm{IG}}, 20)$-ES relates to the lowest graph (○). The individual σ-change is independent of the chosen μ. Theorem 3 states that the sum over $i = 1 \ldots 10$ is zero. In the data shown the value is $-3.56 \cdot 10^{-4}$. The symmetry of the graph shows that the i-th best and i-th worst individual have the same expected $\lg(\sigma)$. Note that already the 5-th best out of 20 individuals has a decreased expected $\lg(\sigma)$ value, compared to the mean parental value (zero at the y-axis), and therefore carries a worse strategy parameter than the parental population.

In Fig. 4 the expected $\lg(\sigma)$ change of ES$_{\mathrm{IG}}$ (○) between two generations is shown for $\mu = 1; 5; 10; 15; 19$. The points correspond to the mean value of the first μ points in Fig. 3. For $\mu < 10$ the $\lg(\sigma)$ change for the $(\mu/\mu_{\mathrm{IG}}, 20)$-ES (○) is positive. According to Theorem 2, the expected change is negative for $11 \leq \mu \leq 19$.

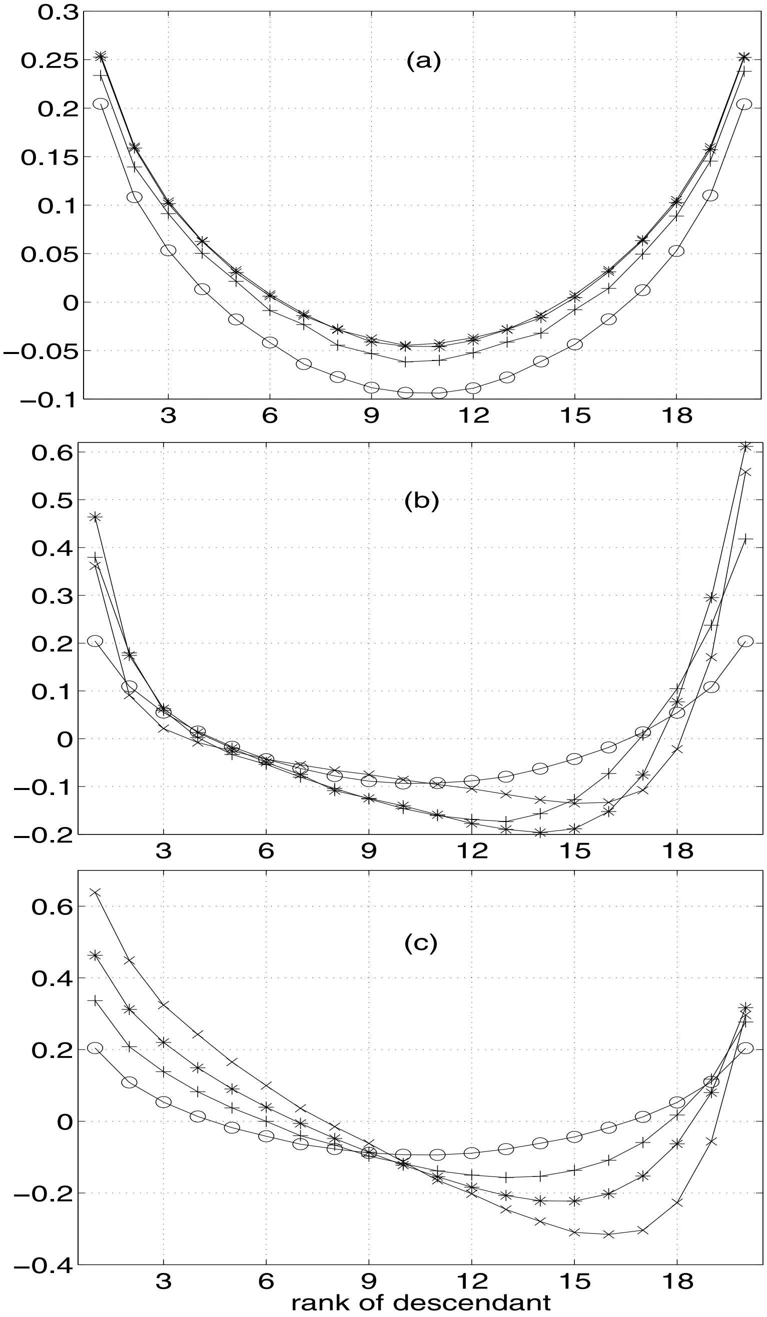

Figure 3: Expected logarithmic step-size change of single individuals, i.e., $\mathrm{mean}_{1e3\leq g<5e4}(\lg(\sigma_{i:20}^{(g+1)})-\mathrm{mean}_{j\in\{i:\lambda|i=1,\ldots,\mu\}}(\lg(\sigma_j^{(g)})))$ on f_{linear} versus fitness rank of descendants from left (best) to right (worst). (a): $(\mu/\mu,20)$-ES$_{\mathrm{IA}}$, (b): $(\mu/1,20)$-ES$_{\mathrm{unlink}}$, (c): $(\mu/1,20)$-ES$_{\mathrm{w/o}}$. Results given for $\mu=1;5;10;15$ ($\circ;+;*;\times$), where problem dimension $n=2$.

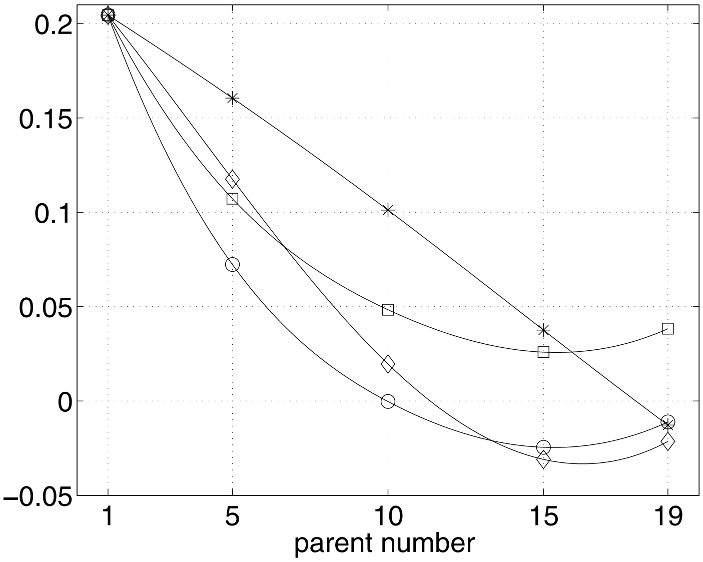

Figure 4: Expected logarithmic step-size change within one generation step, i.e., $\mathrm{mean}_{1e3\leq g<5e4}(\mathrm{mean}_{j\in\{i:\lambda|i=1,\ldots,\mu\}}(\lg(\sigma_j^{(g+1)}))-\mathrm{mean}_{j\in\{i:\lambda|i=1,\ldots,\mu\}}(\lg(\sigma_j^{(g)})))$ on f_{linear} versus parent number μ for $\mu=1;5;10;15;19$. ○: $(\mu/\mu,20)$-ES$_{\mathrm{IG}}$, □: $(\mu/\mu,20)$-ES$_{\mathrm{IA}}$, ◇: $(\mu/1,20)$-ES$_{\mathrm{unlink}}$, $*$: $(\mu/1,20)$-ES$_{\mathrm{w/o}}$. Problem dimension $n=2$. Values ≤ 0 reveal a failure of the strategy.

The situation changes significantly for ES$_{\mathrm{IA}}$. The graphs can be seen in Fig. 3 (a) for $\mu=1;5;10;15$ and $\lambda=20$. For $\mu=1$ the graph must be obviously identical with that of ES$_{\mathrm{IG}}$. With increasing μ the graphs move upward. For $\mu=10$ still four of the new parents have a smaller $\lg(\sigma)$ value than the former population. Nevertheless, the sum over $i=1\ldots\mu$ is greater than zero in every single graph. These mean values can be seen in Fig. 4 (□) for $\mu=1;5;10;15;19$ respectively. They correspond to σ-changing factors between 1.60 and 1.06 per generation. Therefore step-sizes increase in any of these selection schemes on f_{linear}. Due to the violation of the σ-stationarity Assumption 2 neither Theorem 2 nor Theorem 3 holds.

Results for ES$_{\mathrm{unlink}}$, where $\boldsymbol{x}$ and σ are chosen independently for each offspring, are shown in Fig. 3(b), again for $\mu=1;5;10;15$. Graphs of the expected $\lg(\sigma)$ are clearly asymmetrical for $\mu>1$ and have their minimum at offspring number $13;14;15$ for $\mu=5;10;15$. In all cases the 5-th best individual already has a decreased $\lg(\sigma)$. According to the σ-stationarity Assumption 2, the sum over all points of any single graph deviates only stochastically from zero. That is, under random selection or in a flat fitness landscape the expected $\lg(\sigma)$ change is zero for any μ. Due to the violation of the symmetry Assumption 1, Theorem 3 does not hold. As can be seen in Figure 4 (◇) the $(10,20)$-scheme increases $\lg(\sigma)$. While the $\lg(\sigma)$ change is 0.002 for the $(11,20)$-scheme, it is smaller than zero for $12\leq\mu\leq 19$.

Results for ES$_{\mathrm{w/o}}$ without recombination are shown in Fig. 3(c). The graph for $\mu=1$ is identical with that one in (a) and (b). For $\mu>1$ graphs become even more asymmetrical than for ES$_{\mathrm{unlink}}$. Again, according to the σ-stationarity Assumption 2, the sum over all points of any single graph deviates only stochastically from zero, and under random selection or in a flat fitness landscape the expected $\lg(\sigma)$ change is zero.

Table 1: μ values of the $(\mu, 20)$-σ SA-ES on f_{linear} classified by $\log \sigma$-changes

	$\text{ES}_{\text{w/o}}$	$\text{ES}_{\text{unlink}}$	ES_{IA}	ES_{IG}	$\text{ES}_{\text{I w/o}}$
$\log \sigma$ increases	$1-17$	$1-11$	$1-20$	$1-9$	$1-9$
$\log \sigma$ decreases	$18, 19$	$12-19$	$\emptyset$	$11-19$	$11-19$
$\log \sigma$ is unbiased	20	20	$\emptyset$	$10, 20$	$10, 20$

For $\mu = 1$ and $\mu = 5$ *all* parents have a positive $\lg(\sigma)$ change. For $\mu = 10$ the best 6, for $\mu = 15$ the best 7 individuals show a positive $\lg(\sigma)$ change. Figure 4 shows $(*)$ that for $\mu < 18$ the overall population $\lg(\sigma)$ change is greater than zero. It is smaller than zero only for $\mu = 18$ ($\lg(\sigma) = -0.002$) and $\mu = 19$. The violation of the symmetry Assumption 1 is here even more significant. Neither Theorem 2 nor Theorem 3 holds.

Table 1 classifies the μ values by $\log \sigma$-changes for all $(\mu, 20)$-σ SA-ESs. Only the ES_{IA} increases μ for all $\mu < \lambda$ and for random selection ($\mu = 20$). The latter is not necessarily desirable, because it can cause the step-size to diverge under weak selection, e.g. in a noisy environment.

How these results relate to single runs on f_{linear} can be observed in Fig. 5 for different $(12, 20)$-σ SA-ESs, where $n = 10$. While in ES_{IA} (a) and $\text{ES}_{\text{w/o}}$ (c) there is a linear increase of $\lg(\sigma)$ over the time, in ES_{IG} (b) and $\text{ES}_{\text{unlink}}$ (d) it linearly decreases. The increase of $\lg(\sigma)$ is considerably larger in $\text{ES}_{\text{w/o}}$ than in ES_{IA}. This difference becomes larger with increasing n. The remarkable difference between $\text{ES}_{\text{w/o}}$ (c) and $\text{ES}_{\text{unlink}}$ (d) is due to the existing/missing direct linkage between $\boldsymbol{x}$ and σ for the transition between parents and descendants. Note that in a flat or completely random fitness landscape one would observe an unbiased random walk of $\lg(\sigma)$ for all but ES_{IA}, where $\mu > 1$.

7 Summary and Conclusions

We investigated σ-self-adaptation (σ SA), i.e., the mutative control of one global step-size σ of an evolutionary algorithm on the linear fitness function f_{linear}. We find that $\log \sigma$ is the adequate measure to examine σ SA. Based on this observation we call an operator *unbiased*, if $\log \sigma$ is not changed in expectation. Naturally, random selection or selection in a flat fitness landscape is unbiased. The mutation operator that is most commonly applied to the step-size, i.e. $\sigma \mapsto \sigma \cdot \exp(Y)$, where $\mathrm{E}[Y] = 0$, is unbiased as well. The frequently applied arithmetic recombination of step-sizes is biased towards an increase of $\log \sigma$ (Proposition 5).

As an elementary demand on step-size control on f_{linear} we identify the (expected) increase of $\log \sigma$, linear in time (Postulate 1). The σ SA does not meet this demand in general. For example, in the $(1, 2)$-σ SA-EA, selection on f_{linear} leaves the σ-distribution unchanged. Therefore, even after a complete $(1, 2)$-σ SA-EA generation step, σ remains unbiased on f_{linear} and $\log \sigma$ does not increase.

More general, in a (μ, λ)-σ SA-EA on f_{linear}, the i-th best and the i-th worst individual of the population have the same σ-distribution, given that the distribution of new search points (generated by recombination and mutation) is point-symmetrical in the object parameter space. We can derive two consequences.

- On f_{linear}, the correlation between fitness and step-size is zero. A necessary prerequisite for self-adaptation is that better individuals (i.e., individuals with a higher fitness of their object parameters) also inherit better *strategy parameters*. Because the correlation between fitness and step-size is zero, the link between fitness and step-size quality is less evident than one would expect.

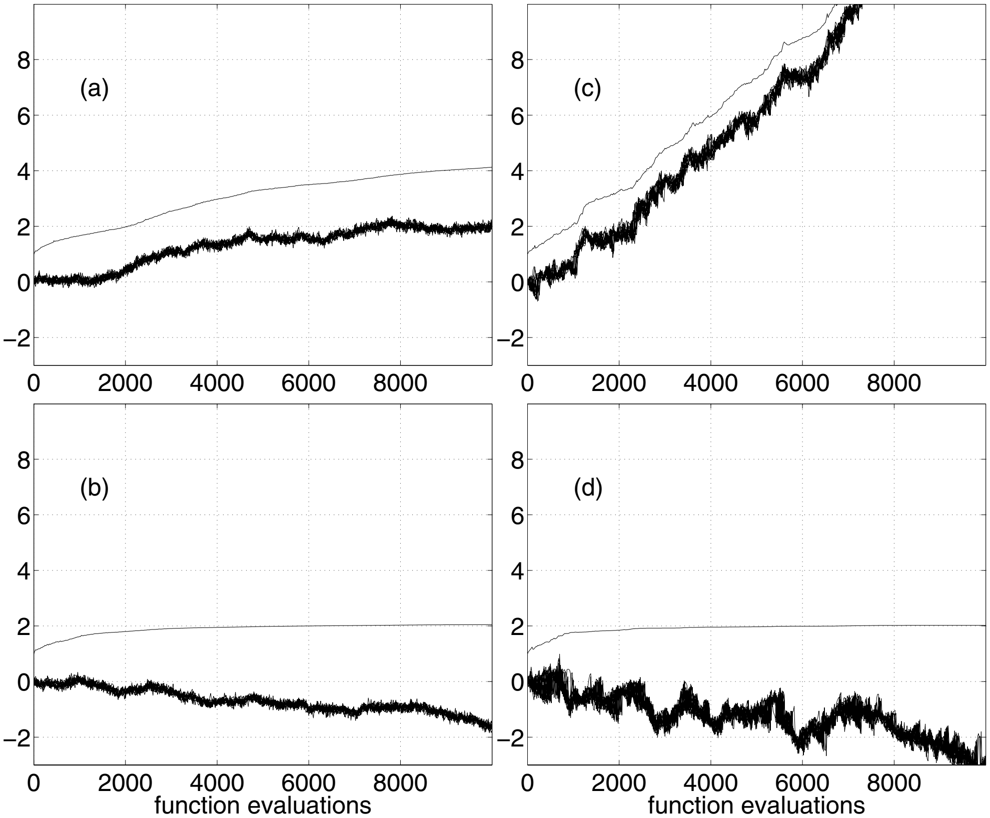

Figure 5: Best $\lg(f_{\text{linear}})$ (upper graph) and $\lg(\sigma)$ of all parents versus number of function evaluations. Shown is a single run of (a): $(12/12_{\text{IA}}, 20)$-σ SA-ES, (b): $(12/12_{\text{IG}}, 20)$-σ SA-ES, (c): $(12/1, 20)$-σ SA-ES$_{\text{w/o}}$, (d): $(12/1, 20)$-σ SA-ES$_{\text{unlink}}$. Problem dimension $n = 10$. Only in (a) and (c) an increase of $\lg(\sigma)$ and a corresponding constant fitness gain can be observed.

- Provided that recombination and mutation leave the step-size unbiased (stationarity assumption), on f_{linear} the $(\mu/\mu_{\text{I}}, \lambda)$-$\sigma$ SA-EA increases the expected step-size for $\mu < \lambda/2$, decreases the expected step-size for $\mu > \lambda/2$, and leaves the step-size unbiased for $\mu = \lambda/2$.

Given a symmetrical distribution of the new population before selection and an unbiased generation of the step-size, the (μ, λ)-σ SA-EA fails to increase σ on f_{linear} for $\mu \geq \lambda/2$. In practice, this failure on f_{linear} does not play an important role. Usually applied strategy variants do not fail, because μ is chosen $< \lambda/2$, or the recombination of σ is biased, or a recombination scheme is chosen that does not lead to a symmetrical population distribution. Nevertheless, the analysis helps to *understand why* σ SA usually works adequately on f_{linear}. The results also indicate that a low selection pressure does *not necessarily* lead to a failure of σ SA (see Table 1).

The bias introduced by arithmetic recombination of step-sizes can explain previously reported "surprising" observations.

- Kursawe (1995) gives an example for a failure of *individual* σ SA, i.e., self-adaptation of n individual step-sizes, where $n = 100$. The strategy diverges in several cases, but only when arithmetic multi-recombination of step-size is applied (as in $\mathrm{ES}_{\mathrm{IA}}$, see Section 3). In the light of this paper Kursawe's observation can be explained as follows. The bias on the step-sizes introduced by arithmetic recombination, that increases $\log \sigma$ in expectation, must be compensated. Compensation can be accomplished in two ways.

- By selection. Selection can compensate the bias best, the smaller the number of self-adapted step-sizes. With individual step-sizes in large dimensions, the number of step-sizes becomes too large and σ SA can diverge even on the sphere model.
- By intermediate recombination of the object parameters. Intermediate recombination reduces the finally realized step *lengths* and therefore can compensate a too large step-size σ. The largest step-sizes are still discarded by selection, because they tend to produces much lower fitness values with a high probability.

Divergence is observed if neither selection sufficiently compensates the bias nor intermediate recombination is applied. Our interpretation can be supported by additional simulations, not shown here.

- Arnold and Beyer (2000) investigate the $(3/3_{\mathrm{I}}, 10)$-σ SA-$\mathrm{ES}_{\mathrm{IA}}$ on a noisy sphere model. They observe divergence of the step-size when the (proportional) noise level becomes large. Their observation can be explained as follows. If selection is highly disturbed by noise, i.e. selection becomes almost random, the biased recombination of step-sizes cannot be compensated and leads to divergence. The divergence disappears when the unbiased geometric recombination is used ($\mathrm{ES}_{\mathrm{IG}}$), but then the step-size becomes much too small.

- The dynamics of the $(\mu/\mu_{\mathrm{I}}, \lambda)$-$\sigma$ SA-$\mathrm{ES}_{\mathrm{IA}}$on noisy functions depends significantly on μ. Only for $\mu = 1$ step-sizes seem to converge to zero where the noise level is large (Beyer and Sendhoff, 2005). The dependency can be explained as follows. For $\mu = 1$ no recombination takes place and therefore σ is unbiased before selection. Because, for large noise levels, the selection is almost random, we expect an almost unbiased random walk of $\log \sigma$. Because, in the given noise model, too large step-sizes lead to a remarkable impairment of the fitness, the random walk is bounded from above and can appear as convergence to zero. For $\mu > 1$ in the $\mathrm{ES}_{\mathrm{IA}}$ the random walk becomes biased and σ will not become arbitrarily small.

Results from an analysis, as presented in this paper, can be most useful to derive design criteria and specifications for (new) search algorithms. Two such conclusions are drawn on the dynamics of the population variance.

- As a basic specification on f_{linear}, the population variance should increase linearly on the log scale (Postulate 1).[10] To our intuition, an increase by a factor of at least 1.1^2 after n function evaluations, or, for large populations, by a factor of at least 1.5^2 after one generation, seems adequate.

[10] In our case the population variance corresponds to the squared step-size σ^2. To perform on f_{linear}, an increase of the population variance in gradient direction is sufficient. To maintain diversity and exploration an increase in all directions seems preferable.

- Under random selection the population variance should be unbiased, because any bias entails the danger of divergence or premature convergence. In a noisy or flat environment, a bias towards increase can be useful and desirable, but it should be carefully quantified. We believe, the increase of the population variance under random selection should be small, *and in particular smaller than its increase on* f_{linear}.

Finally, to recapitulate, a practitioner will, in our opinion, most frequently face the following drawback of σ SA. For $\mu > 1$ the target step-size of σ SA is usually (far) smaller than the optimal step-size, because σ SA is based on selection of individuals, while optimality of step-size wisely refers to the advance of the whole population. Therefore, after the adaptation has taken place, the algorithm will usually operate with step lengths considerably smaller than optimal. For the same reason, too small step lengths must be expected from any mechanism that determines (explicitly or implicitly) the population variance *based on individual selection*.

Acknowledgments

The author is deeply grateful to Anne Auger for her many invaluable comments on the manuscript.

References

Arnold, D. and Beyer, H.-G. (2000). Efficiency and mutation strength adaptation of the $(\mu/\mu_I, \lambda)$-ES in a noisy environment. In Schoenauer, M., Deb, K., Rudolph, G., Yao, X., Lutton, E., Merelo, J., and Schwefel, H.-P., editors, *Parallel Problem Solving from Nature—PPSN VI, Proceedings*, pages 39–48, Paris. Springer, Berlin.

Bäck, T. and Schwefel, H.-P. (1993). An overview of evolutionary algorithms for parameter optimization. *Evolutionary Computation*, 1(1):1–23.

Beyer, H.-G. (1996). Toward a theory of evolution strategies: Self-adaptation. *Evolutionary Computation*, 3(3):311–347.

Beyer, H.-G. (2001). *The Theory of Evolution Strategies*. Natural Computing Series. Springer, Heidelberg.

Beyer, H.-G. and Deb, K. (2001). On self-adaptive features in real-parameter evolutionary algorithms. *IEEE Transactions on Evolutionary Computation*, 5(3):250–270.

Beyer, H.-G. and Sendhoff, B. (2005). Functions with noise-induced multi-modality: A test for evolutionary robust optimization—properties and performance analysis. *IEEE Transactions on Evolutionary Computation*. forthcoming.

Grünz, L. and Beyer, H.-G. (1999). Some Observations on the Interaction of Recombination and Self-Adaptation in Evolution Strategies. In Angeline, P., editor, *Proceedings of the CEC'99 Conference*, pages 639–645, Piscataway, NJ. IEEE.

Hansen, N. (1998). *Verallgemeinerte individuelle Schrittweitenregelung in der Evolutionsstrategie. Eine Untersuchung zur entstochastisierten, koordinatensystemunabhängigen Adaptation der Mutationsverteilung.* Mensch und Buch Verlag, Berlin. ISBN 3-933346-29-0.

Hansen, N., Gawelczyk, A., and Ostermeier, A. (1995). Sizing the population with respect to the local progress in (1,λ)-evolution strategies — a theoretical analysis. In *1995 IEEE International Conference on Evolutionary Computation Proceedings*, pages 80–85.

Hansen, N. and Ostermeier, A. (1996). Adapting arbitrary normal mutation distributions in evolution strategies: The covariance matrix adaptation. In *Proceedings of the 1996 IEEE International Conference on Evolutionary Computation*, pages 312–317.

Hansen, N. and Ostermeier, A. (2001). Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195.

Herdy, M. (1993). The number of offspring as strategy parameter in hierarchically organized evolution strategies. *SIGBIO Newsletter*, 13(2):2–7.

Kursawe, F. (1995). Towards self-adapting evolution strategies. In *1995 IEEE International Conference on Evolutionary Computation Proceedings*, pages 283–288. IEEE Press.

Ostermeier, A. (1997). *Schrittweitenadaptation in der Evolutionsstrategie mit einem entstochastisierten Ansatz*. PhD thesis, Technische Universität Berlin.

Ostermeier, A., Gawelczyk, A., and Hansen, N. (1994). A derandomized approach to self-adaptation of evolution strategies. *Evolutionary Computation*, 2(4):369–380.

Rechenberg, I. (1994). *Evolutionsstrategie '94*. frommann-holzboog, Stuttgart.

Schwefel, H.-P. (1995). *Evolution and Optimum Seeking*. Sixth-Generation Computer Technology Series. John Wiley & Sons Inc., New York.

Yao, X., Liu, Y., and Lin, G. (1999). Evolutionary programming made faster. *IEEE Transactions on Evolutionary Computation*, 3(2):82–102.