Mutate Large, But Inherit Small! On the Analysis of Rescaled Mutations in $(\tilde{1}, \tilde{\lambda})$-ES with Noisy Fitness Data

Hans-Georg Beyer⋆

University of Dortmund, Department of Computer Science XI
D-44221 Dortmund, Germany

Abstract. The paper presents the asymptotical analysis of a technique for improving the convergence of evolution strategies (ES) on noisy fitness data. This technique that may be called "Mutate large, but inherit small", is discussed in light of the EPP (evolutionary progress principle). The derivation of the progress rate formula is sketched, its predictions are compared with experiments, and its limitations are shown. The dynamical behavior of the ES is investigated. It will be shown that standard self-adaptation has considerable problems to drive the ES in its optimum working regime. Remedies are provided to improve the self-adaptation.

1 Introduction

The evolution in noisy environments has some peculiarities over the non-perturbed fitness case. This is due to the deception introduced by the noisy fitness data, denoted by $\tilde{Q}$, which randomly mislead the selection process in the evolutionary algorithms. The deception can be so strong that even elitist selection schemes may exhibit divergence behavior. That is, the initial distance $R^{(0)}$ of the population to the optimum grows with increasing time g, $R^{(g)} > R^{(0)}$. This remarkable behavior has been first derived and experimentally verified in [1] for the $(\tilde{1} \overset{+}{,} \tilde{\lambda})$-ES[1] on the sphere model. In this paper we will resume the analysis of the $(\tilde{1}, \tilde{\lambda})$-ES. According to the work of Ostermeier and Rechenberg published in [6], there is a new technique for coping with noise deception, which is worth investigating here. The paper is organized as follows. The next section re-examines the $(\tilde{1}, \tilde{\lambda})$-ES progress rate theory in the light of the *evolutionary progress principle* (EPP) leading us to the idea of rescaled mutations in an natural way. Section 3 is devoted to the asymptotical analysis. Section 4 relates these results to the real time behavior of the $(\tilde{1}, \tilde{\lambda})$-ES including self-adaptation. We will see that *current σ-self-adaptation (σSA) methods do work for* $(\tilde{1}, \tilde{\lambda})$*-ES*, but that *they are not able to reach the domain of optimum performance* (i.e. the vicinity of the theoretically predicted maximal progress rate).

⋆ e-mail: beyer@ludo.cs.uni-dortmund.de

[1] The tilde indicates that the parental fitness data ($\tilde{1}$: there is only one parent) and the offspring values ($\tilde{\lambda}$: there are λ offspring) are disturbed by noise.

2 The $(\tilde{1}, \tilde{\lambda})$-ES, the EPP, and Rescaled Mutations

2.1 The $(\tilde{1}, \tilde{\lambda})$-ES Re-examined

Let us first reconsider the $(\tilde{1}, \tilde{\lambda})$-ES. It is assumed that the reader is familiar with the ES and the progress rate theory on the sphere model. Therefore, only a short summary is given here. The parent $\mathbf{y}_p$ at generation $g+1$ is the best offspring of generation g which is obtained by the mutation $\mathbf{z}_{\tilde{1};\tilde{\lambda}}$ that has produced the fittest individual, i.e. $\mathbf{y}_p^{(g+1)} := \mathbf{y}_p^{(g)} + \mathbf{z}_{\tilde{1};\tilde{\lambda}}$. The vectors $\mathbf{y}$ and $\mathbf{z}$ are from the N-dimensional real-valued space $\mathbb{R}^N$. The mutations are isotropic and Gaussian distributed with each component sampling from $\mathcal{N}(0, \sigma^2)$. As fitness model, the noisy sphere is considered, i.e., $\tilde{Q}(\mathbf{y}) = Q(R) + \varepsilon_Q$, with $R = \|\mathbf{y}\|$ the distance to the optimum and ε_Q a noise source which is assumed to be normally distributed $\varepsilon_Q \sim \mathcal{N}(0, \sigma_{\varepsilon_Q}^2)$, where σ_{ε_Q} can be a function of R. For this model, the progress rate φ, i.e., the expected distance change toward the optimum, can be calculated [1] (for the definition of φ, see Section 3). Using normalized quantities

$$\varphi^* := \varphi \frac{N}{R}, \quad \sigma^* := \sigma \frac{N}{R}, \quad \sigma_\varepsilon^* := \frac{\sigma_{\varepsilon_Q}(R)}{|Q'(R)|} \frac{N}{R} \quad \text{with} \quad Q'(R) := \frac{dQ}{dR} \tag{1}$$

the normalized progress rate reads

$$\varphi^*_{\tilde{1},\tilde{\lambda}}(\sigma^*, \sigma^*_\varepsilon) = c_{1,\lambda} \sigma^* \frac{1}{\sqrt{1 + (\sigma^*_\varepsilon / \sigma^*)^2}} - \frac{\sigma^{*2}}{2} + \mathcal{O}\left(\frac{1}{N}\right). \tag{2}$$

Here, $c_{1,\lambda}$ is the progress coefficient being the expectation of the λth order statistics of the standard normal variate (see [1], p. 171).[2]

2.2 The EPP and Convergence Improvement Techniques

It is interesting to discuss Eq. (2) from the viewpoint of the evolutionary progress principle (EPP), proposed in [4]. This principle states that the evolutionary progress toward the optimum is the result of two opposite forces:

EVOLUTIONARY PROGRESS	=	PROGRESS GAIN	−	PROGRESS LOSS

[2] A simple, but good approximation for $c_{1,\lambda}$ has been proposed by Rechenberg [5] $c_{1,\lambda} \approx \Phi^{-1}(\sqrt[\lambda]{1/2})$, with Φ^{-1} the inverse to the cdf (i.e., the quantile function) of the standard normal variate. The relative error is less than 3.5%.

As to the asymptotical ($N \to \infty$) formula (2), one can see that the *loss part* is a quadratic function of σ^*, whereas the *gain part* is given by $c_{1,\lambda}\sigma^*/\sqrt{1+(\sigma_\varepsilon^*/\sigma^*)^2}$. An important observation is that selection influences the gain part, but not the loss part. Because $\sigma_\varepsilon^* = 0$ yields as gain $c_{1,\lambda}\sigma^*$, one can further conclude that the effectiveness of selection is degraded by increasing the noise level σ_ε^*. If a fixed normalized mutation strength ($\sigma^* = \text{const.}$) is assumed then there are two possibilities to improve the convergence rate of the standard $(\tilde{1}, \tilde{\lambda})$-ES [1]:

a) Increase "selection pressure" by increasing λ, because $c_{1,\lambda} \sim \sqrt{2\ln(\lambda)}$.
b) Improve the effectiveness of the selection by reduction of σ_ε^* using m-times repeated fitness measurements with averaging. This produces a $\sigma_\varepsilon^{*\prime} = \sigma_\varepsilon^*/\sqrt{m}$.

It is to notice that both methods are not for free because they both afford additional fitness calculations and in case a) mutation and selection costs.

The aforementioned techniques work through the increase of the gain part; however, there are also methods for reducing the loss part of the EPP. Actually, recombination is such a technique; in case of intermediate (μ/μ) multi-parent recombination, the loss part reduces up to a factor of μ, however, recombination will not be considered here. As another approach, one could perhaps take advantage of the nonlinearity of the loss part in Eq. (2). Since the gain part is for small σ_ε^* ($\sigma_\varepsilon^*/\sigma^* \ll 1$) almost linear in σ^*, a reduction of σ^* by a factor of $1/\kappa$ reduces much more the loss part than the gain part, thus, providing a positive net progress. This is exactly the effect of a modified ES proposed by Ostermeier and Rechenberg [6] that follows the slogan: *"mutate large, but inherit small"*. That is, one produces λ offspring in the usual way by $\mathbf{z}_l$ mutations applied to the parent $\mathbf{y}_p^{(g)}$. The mutation $\mathbf{z}_{\tilde{1};\tilde{\lambda}}$ that generates the best offspring; however, is *rescaled* by a factor $1/\kappa$. This rescaled mutation is then added to the parent of generation g to produce the new $\mathbf{y}_p^{(g+1)}$. The algorithm reads

$$\mathbf{y}_p^{(g+1)} := \mathbf{y}_p^{(g)} + \frac{1}{\kappa}\mathbf{z}_{\tilde{1};\tilde{\lambda}}, \qquad \kappa \geq 1. \tag{3}$$

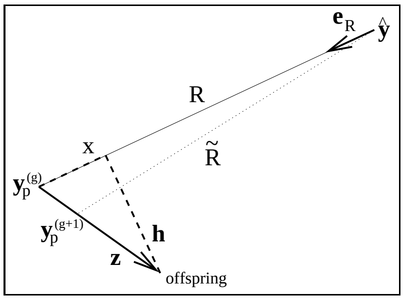
In the following section, it will be shown that the algorithm (3) indeed reduces the loss part of the progress rate formula.

3 Asymptotical Analysis

3.1 Derivation of the φ^* Formula

Due to space limitations only a sketch can be given that reflects the main ideas for the derivation of the asymptotically ($N \to \infty$) correct φ^* formula (see [3] for details). Let R be the distance of $\mathbf{y}_p^{(g)}$ to the optimum and $\tilde{R}$ that of $\mathbf{y}_p^{(g+1)}$. The progress rate φ is defined as the expectation $\varphi := \mathrm{E}\{R - \tilde{R}\}$. In order to calculate $\tilde{R}$ produced by $\mathbf{z}_{\tilde{1};\tilde{\lambda}}/\kappa$, according to (3), the mutations $\mathbf{z}_l$ are decomposed in such

Fig. 1. Decomposition of the $\mathbf{z}$ mutation in a part x in optimum direction and a perpendicular $\mathbf{h}$-vector, i.e., $\mathbf{z} = -x\mathbf{e}_R + \mathbf{h}$. R is the parental distance $\|\mathbf{y}_p^{(g)} - \hat{\mathbf{y}}\|$ to the optimum $\hat{\mathbf{y}}$. $\tilde{R}$ is the parental distance $\tilde{R} = \|\mathbf{y}_p^{(g+1)} - \hat{\mathbf{y}}\|$ for generation $g+1$ obtained *after* rescaling of that mutation which belongs to the best offspring. Due to the isotropy of the mutations, the x as well as the components of $\mathbf{h}$ are $\mathcal{N}(0, \sigma^2)$ distributed.



a way that one component, say x, points in optimum direction $-\mathbf{e}_R$ and the rest is carried by a vector $\mathbf{h}$. From Fig. 1, we read $\mathbf{y}_p^{(g)} - \hat{\mathbf{y}} = R\mathbf{e}_R$. Thus, we get for $\tilde{R} = \|\mathbf{y}_p^{(g+1)} - \hat{\mathbf{y}}\|$, using (3) and $\mathbf{z}_{\tilde{1};\tilde{\lambda}} = -x_{\tilde{1};\tilde{\lambda}}\mathbf{e}_R + \mathbf{h}_{\tilde{1};\tilde{\lambda}}$, the expression $\tilde{R} = \|(R - \frac{1}{\kappa}x_{\tilde{1};\tilde{\lambda}})\mathbf{e}_R + \frac{1}{\kappa}\mathbf{h}_{\tilde{1};\tilde{\lambda}}\|$. Therefore the progress rate is

$$\varphi_{\tilde{1},\tilde{\lambda}} = \mathrm{E}\left\{R - \sqrt{\left(R - x_{\tilde{1};\tilde{\lambda}}/\kappa\right)^2 + \mathbf{h}^2_{\tilde{1};\tilde{\lambda}}/\kappa^2}\right\}. \tag{4}$$

This expression can be further simplified if one takes the symmetry of the spherical model into account: The $\mathbf{h}$ parts of the mutations are selectively neutral as to their orientations in the $(N-1)$-dimensional subspace (isotropy). Therefore, the $\mathbf{h}^2$ of a single mutation is simply the sum over $N-1$ independent and squared $\mathcal{N}(0, \sigma^2)$ variates, i.e., $\mathbf{h}^2$ is χ^2 distributed with expectation $\mathrm{E}\{\mathbf{h}^2\} = (N-1)\sigma^2$ and standard deviation $\mathrm{D}\{\mathbf{h}^2\} = \sqrt{2(N-1)}\sigma^2$. Since $\mathrm{D}\{\mathbf{h}^2\}/\mathrm{E}\{\mathbf{h}^2\} \to 0$ for $N \to \infty$, the simplification $\mathbf{h}^2_{\tilde{1};\tilde{\lambda}} \sim \mathrm{E}\{\mathbf{h}^2\} = (N-1)\sigma^2 \sim N\sigma^2$ is justified. If this is inserted in (4), one obtains after a short calculation $\varphi_{\tilde{1},\tilde{\lambda}} \simeq R\,\mathrm{E}\left\{1 - \sqrt{1 - \frac{2x_{\tilde{1};\tilde{\lambda}}}{R\kappa} + \frac{x^2_{\tilde{1};\tilde{\lambda}}}{R^2\kappa^2} + \frac{N\sigma^2}{R^2\kappa^2}}\right\}$. Now, the square root is expanded into a Taylor series, i.e., $1 - \sqrt{1-2y} = y + \mathcal{O}(y^2)$, and the $(x_{\tilde{1};\tilde{\lambda}}/\kappa R)^2$ term is neglected. As result $\varphi_{\tilde{1},\tilde{\lambda}} \simeq R\,\mathrm{E}\left\{\frac{x_{\tilde{1};\tilde{\lambda}}}{R\kappa} - \frac{N\sigma^2}{R^2\,2\kappa^2}\right\}$ is obtained. Using the normalization from Eq. (1) we arrive at $\varphi^*_{\tilde{1},\tilde{\lambda}} \simeq \frac{N}{R}\frac{1}{\kappa}\mathrm{E}\{x_{\tilde{1};\tilde{\lambda}}\} - \frac{\sigma^{*2}}{2\kappa^2}$. As one can see, the rescaling technique (3) produces just the desired effect: The loss part is reduced by κ^2, whereas the gain part only by κ.

Due to space limitations, the calculation of $\mathrm{E}\{x_{\tilde{1};\tilde{\lambda}}\}$ must be omitted here. One finds $\mathrm{E}\{x_{\tilde{1};\tilde{\lambda}}\} = c_{1,\lambda}\sigma/\sqrt{1 + (\sigma^*_\varepsilon/\sigma^*)^2}$ which is similar to the linear part of Eq. (2). Thus, one finally obtains the asymptotical φ^* formula

$$\boxed{\varphi^*_{\tilde{1},\tilde{\lambda}}(\sigma^*, \sigma^*_\varepsilon) = \frac{1}{\kappa}\left[c_{1,\lambda}\sigma^*\,\frac{1}{\sqrt{1 + (\sigma^*_\varepsilon/\sigma^*)^2}} \;-\; \frac{\sigma^{*2}}{2\kappa}\right].} \tag{5}$$

3.2 Comparison with Experiments and Discussion

The progress rate formula (5) has been extensively tested. As an example the $(\tilde{1}, \tilde{5})$-ES will be simulated using 1-generation experiments. That is, for a fixed mutation strength σ the parent is randomly placed at distance R to the optimum. Then, algorithm (3) is performed for *one* generation and the parental distance change is calculated. This 1-generation experiment is repeated T times and the average distance change serves as an estimate for $\varphi(\sigma)$. Figures 2 and 3 show simulation results for a noise level $\sigma_\varepsilon^* = 3.5$ and different values of N and κ. The

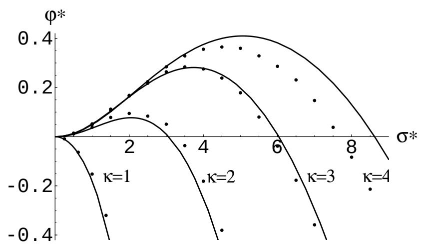


Fig. 2. On the approximation quality of the φ^* formula (5). The dots are obtained from the simulations of a $(\tilde{1}, \tilde{5})$-ES. Each dot is the average over $T = 100,000$ 1-generation experiments. The case $N = 30$ (parameter space dimensions) is displayed. For the cases $N = 100$ and $N = 500$, see Fig. 3.

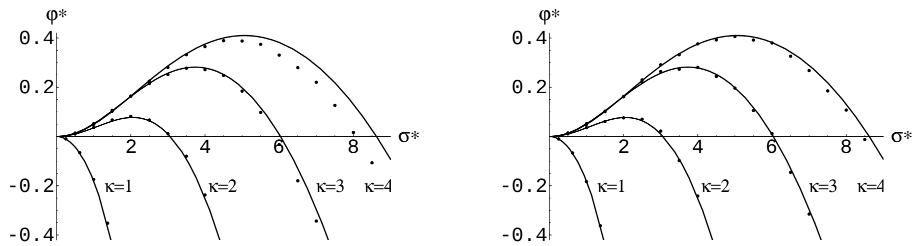


Fig. 3. Left picture: $N = 100$, $T = 100,000$. Right picture: $N = 500$, $T = 50,000$.

progress coefficient $c_{1,5} \approx 1.163$ and as fitness function $Q(\mathbf{y}) = \|\mathbf{y}\|^2$ has been chosen. As one can see, the approximation quality increases with larger N and decreases with increasing σ^*. Since there are considerable deviations for larger κ and σ^* there is still a demand for an N-dependent φ^* formula.

Let us discuss Eq. (5) in detail. Again for the $(\tilde{1}, \tilde{5})$-ES, φ^* is plotted by contour lines in Fig. 4. It seems that φ^* increases monotonously when κ and σ^* are appropriately increased. This can be shown analytically by maximizing (5) with respect to κ and fixed σ^*. One finds

$$\kappa_{\text{opt}} = \frac{1}{c_{1,\lambda}} \sqrt{\sigma^{*2} + \sigma_\varepsilon^{*2}} \qquad \Longrightarrow \qquad \hat{\varphi}^*_{\tilde{1},\tilde{\lambda}} = \frac{1}{1 + (\sigma_\varepsilon^*/\sigma^*)^2} \, \frac{c_{1,\lambda}^2}{2}. \tag{6}$$

The limit case $\kappa_{\text{opt}} \to \infty$ yields $\hat{\varphi}^* \to c_{1,\lambda}^2/2$. That is, for the asymptotical limit, the progress rate of the noisy ES approaches the performance of the ES without

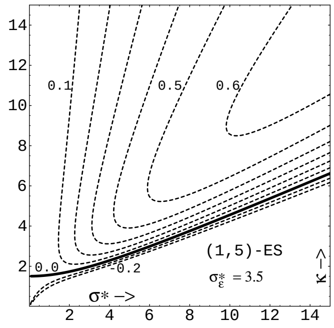


Fig. 4. Contour plot of $\varphi^*_{\tilde{1},\tilde{5}}(\sigma^*, \kappa)$. The solid curve $\varphi^* = 0.0$ separates the domain of convergence from the domain of divergence.

noise! However, it is really important to keep in mind that this result holds for $N \to \infty$. The real-world case $N < \infty$ gives smaller maximal progress and κ is to be chosen appropriately.

As for each ES, the optimum mutation strength $\hat{\sigma}^*$ that maximizes φ^* is of interest. Unlike the determination of κ_{opt} by Eq. (6), maximizing φ^* with respect to σ^* requires a third order equation $y^3 - \kappa c_{1,\lambda} y^2 - \kappa c_{1,\lambda} \sigma_\varepsilon^{*2} = 0$ to be solved for y. The optimal $\sigma^* = \hat{\sigma}^*$ is then obtained by $\hat{\sigma}^* = \sqrt{y^2 - \sigma_\varepsilon^{*2}}$. Since the only real-valued solution is rather lengthy, an approximation is given instead

$$\hat{\sigma}^* \approx \kappa c_{1,\lambda} \sqrt{(1 + 1/(1 + (\kappa c_{1,\lambda}/\sigma_\varepsilon^*)^2))^2 - (\sigma_\varepsilon^*/\kappa c_{1,\lambda})^2}. \qquad (7)$$

The derivation of (7) using asymptotic iteration techniques must be omitted here (space restrictions). If $\sigma_\varepsilon^*/\kappa c_{1,\lambda} \ll 1$ we get again $\hat{\sigma}^* \approx \kappa c_{1,\lambda}$.

4 Dynamical Aspects

4.1 The Thales Criterion of Evolution

As we will see later on, one cannot always expect the ES working in its optimal performance interval $\sigma^* \approx \hat{\sigma}^*$. Therefore, an *evolution criterion* can be helpful to estimate the convergence/divergence of the ES. Such a criterion is easily obtained from the condition $\forall g,\ g > g_0 : R^{(g+1)} < R^{(g)}$. Since $R^{(g+1)} = \mathrm{E}\{\tilde{R}\}$, we have $R^{(g+1)} < R^{(g)} \Leftrightarrow \mathrm{E}\{\tilde{R}\} < R^{(g)} \Leftrightarrow \mathrm{E}\{R - \tilde{R}\} > 0 \Leftrightarrow \varphi^* > 0$. Thus, by means of Eq. (5) one obtains the

$$\boxed{\textsc{Evolution Criterion:} \qquad 4\kappa^2 c_{1,\lambda}^2 > \sigma^{*2} + \sigma_\varepsilon^{*2}.} \qquad (8)$$

If the relation sign in (8) is substituted by an equal sign, a Thales circle with radius $\kappa c_{1,\lambda}$ is prescribed. This allows for a graphical interpretation of the $(\tilde{1}, \tilde{\lambda})$ convergence properties, given in Fig. 5. As one can see, the normalized σ_ε must *necessarily fulfill the condition* $\sigma_\varepsilon^* < 2\kappa c_{1,\lambda}$ in order to have convergence. If

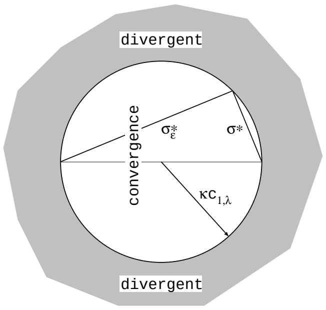


Fig. 5. The Thales criterion of convergence. All σ^*-, σ_ε^*-length combinations lying within the circle of radius $\kappa c_{1,\lambda}$ are convergent states, i.e., the ES reduces on average the distance to the optimum. All states outside are divergent.

this necessary condition is fulfilled, convergence depends on the actual mutation strength σ. That is, one needs a technique that tunes σ in such a way that σ^* and σ_ε^* are in the Thales circle. σ-self-adaptation (σSA) can be used for this purpose.

4.2 Experiments with σ-Self-Adaptation (σSA)

The self-adaptation theory is still in the beginning. Up to now, only the $(1, \lambda)$-ES without noise and $\kappa = 1$ has been analyzed [2]. It is assumed that the reader is familiar with the σSA-algorithm and the notations used in [2]. The mutation of the σ values is performed by multiplying the parental σ with a random number ξ. The ξ variates which have been investigated are the log-normal mutation ξ_{LN} with learning parameter τ [7] and the *generalized two-point mutation* ξ_γ [2] with τ (defined below) and probability parameter γ

$$\xi_{\mathrm{LN}} := \mathrm{e}^{\tau \mathcal{N}(0,1)} \qquad \text{and} \qquad \xi_\gamma := \begin{cases} 1+\tau, & \text{if} \quad \mathtt{u}(0,1] > \gamma \\ 1/(1+\tau), & \text{if} \quad \mathtt{u}(0,1] \leq \gamma \end{cases}. \tag{9}$$

$\mathcal{N}(0,1)$ is a standard normal random number and $\mathtt{u}(0,1]$ is a sample from the uniform distribution $0 < \mathtt{u} \leq 1$. As learning parameter, $\tau = c_{1,\lambda}/\sqrt{N}$ has been chosen and $\gamma = 1/2$ (symmetrical two-point mutations).

All experiments minimize a fitness function $Q(\mathbf{y}) = \|\mathbf{y}\|^2$ using a $(\tilde{1}, \tilde{5})$-ES in an $N = 100$-dimensional space, $\sigma_\varepsilon^* = 3.5$, with the initial conditions $\mathbf{y}^{(0)} = (1.5, \ldots, 1.5)$ and $\sigma^{(0)} = 1$. As rescaling factor, $\kappa = 3$ has been chosen (if not explicitly stated). All figures show the evolution of the residual distance to the optimum and the effect of the σSA by displaying the *normalized* mutation strength σ^*. The left picture of Fig. 6 shows the standard σSA with log-normal mutation of σ and *no* rescaling, i.e., $\kappa = 1$. Since $\sigma_\varepsilon^* = 3.5 > 2c_{1,5} \approx 2.3$, Eq. (8) is not fulfilled and the ES *must* diverge, however, *the σSA does work:* It reduces σ in such a way that $\sigma^* \approx 0$. The right picture shows the ES with rescaling $\kappa = 3$. Unlike the left picture *no* σSA was used, instead an "evolution daemon" tunes the σ by measuring the actual distance $R^{(g)}$ such that $\sigma^{(g)} = \hat{\sigma}^* R^{(g)}/N$ (in real applications this is, of course, excluded). As one can see, the use of the

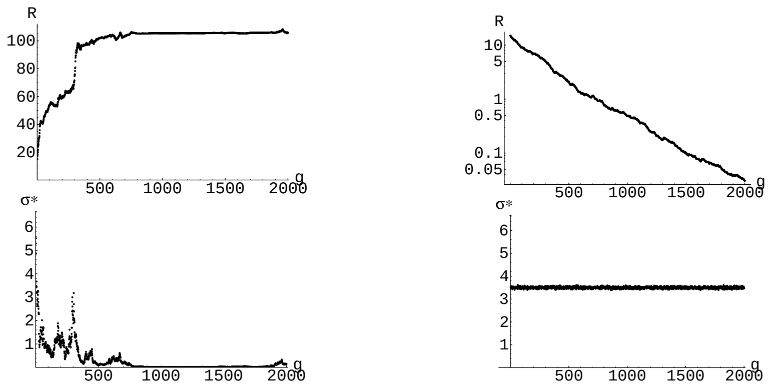


Fig. 6. Left picture: performance of the $(\tilde{1}, \tilde{5})$-σSA-ES with $\kappa = 1$ applying ξ_{LN}. Right picture: performance of the $(\tilde{1}, \tilde{5})$-ES with $\kappa = 3$, *but* σ is dynamically tuned such that $\sigma^* \approx \hat{\sigma}^* \approx \kappa c_{1,5} \approx 3.49$. $\varphi^* \approx 0.28$, measured over $g = 1 \ldots 2000$.

right mutation strength guarantees nearly optimal performance (measured, as well as theoretically: $\varphi^* \approx \hat{\varphi}^* \approx 0.28$).

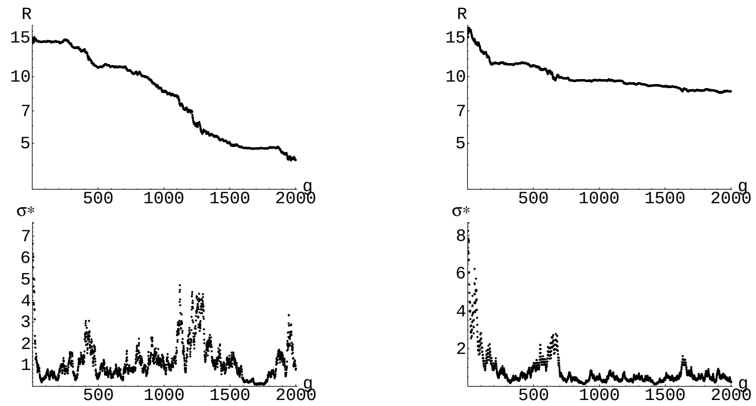


Fig. 7. General setup: $\tau = c_{1,5}/\sqrt{100} \approx 0.116$, $\kappa = 3$. Left picture: σSA with ξ_{LN}. $\varphi^* \approx 0.078$, average of $\sigma^* \approx 1.18$ (measured over $g = 501 \ldots 2000$). Right picture: σSA with $\xi_{\gamma=1/2}$. $\varphi^* \approx 0.014$, average of $\sigma^* \approx 0.60$ (measured over $g = 501 \ldots 2000$).

In Fig. 7 the performance of the σSA using log-normal ξ_{LN}, left picture, and symmetrical two-point ξ_γ (i.e. $\gamma = 1/2$), right picture, for $\kappa = 3$ is compared. By using rescaled mutations, the σSA is able to drive the ES into the convergence regime. However, the progress rates obtained are rather disappointing when compared with the theoretical optimum $\hat{\varphi}^* \approx 0.28$. It seems that the log-normal mutation rule is superior to the symmetrical two-point variant.

4.3 Thoughts About the Poor σSA Performance and Remedies

Let us look once more at Fig. 7. The average performance is well predicted by Eq. (5) if the actually measured averages of σ^* are used. Therefore, the poor performance of the ES is due to the too small steady state σ^* averages of 1.18 and 0.6, respectively, instead of the desirable $\sigma^* = \hat{\sigma}^* \approx 3.49$. Obviously, σSA is not able to drive the ES into the vicinity of the optimum σ^*. Since there is no theory for σSA in noisy environments, only a qualitative explanation for this failure can be given here. To understand the stationary σ^* value (to be more correct, one should speak of the steady state σ^* distribution) one has to recall that the actual parental σ^* is inherited from the offspring with the best $\tilde{Q}$ value obtained by $(1, \lambda)$-selection. However, the rescaling takes place *after* the selection. That is, there is no feedback from the fitness of the parent actually produced by the rescaling Eq. (3). Thus, the σSA learns the optimum σ^* *without* rescaling. Up to now there is no simple way – other than Meta-ES – to learn the optimal σ^* of the rescaling algorithm. However, two "remedies" aiming at an increase of the steady state σ^* shall be mentioned and discussed in short.

The first technique is after an idea of Rechenberg [6], however slightly modified, and uses a rescaling approach similar to Eq. (3). Since the mutation of σ is done by multiplication that maps to the addition when the logarithm is taken, $\ln \tilde{\sigma}_l := \ln \sigma_p^{(g)} + \ln \xi$, the slogan "mutate large, but inherit small" leads to $\ln \tilde{\sigma}_p^{(g+1)} := \ln \sigma_p^{(g)} + \frac{1}{\kappa_\sigma} \ln \xi_{\tilde{1};\tilde{\lambda}}$. Therefore, the new σ update rule reads

$$\tilde{\sigma}_p^{(g+1)} := \sigma_p^{(g)} \cdot \xi_{\tilde{1};\tilde{\lambda}}^{1/\kappa_\sigma}, \qquad \kappa_\sigma \geq 1. \tag{10}$$

The second technique, proposed in this paper, uses ξ_γ mutations with $\gamma < 1/2$. That is, the probability of a σ-increase is greater than that of a σ-decrease. Thus, the steady state σ is shifted toward larger mutation strengths. Both techniques are displayed in Fig. 8. The left picture is a sample generated by an ES using rule (10). The right picture applies non-symmetrical two-point mutations. Both techniques can produce a larger average σ^* than the standard σSA. Up until now there is no theory how to choose the new strategy parameters κ_σ and γ. Also the idea of combining both techniques seems promising; however, this "hybridism" is still under investigation.

5 Outlook

The theoretical investigation of the ES in noisy fitness landscapes as well as the σSA and the rescaling techniques are still *in statu nascendi*. As to the $(\tilde{1}, \tilde{\lambda})$-ES, there are two main directions which should be followed in future research. First, the investigation of the rescaling Eq. (3) in case of finite N. This will allow for the estimation of optimal κ not available from the asymptotic theory. Second, the analysis of the σSA with noisy fitness and generalized σ update rules (10)

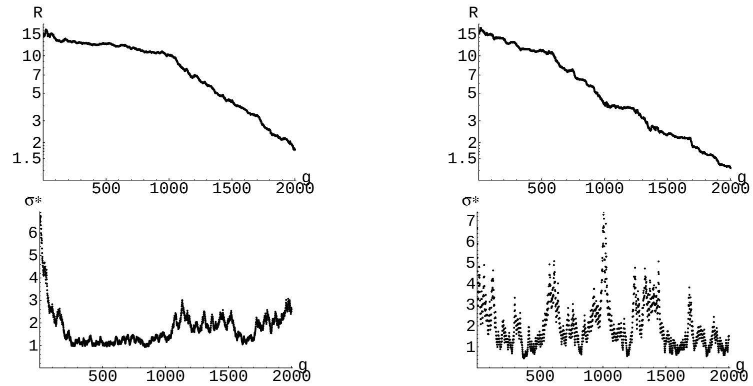


Fig. 8. General setup: $\tau \approx 0.116$, $\kappa = 3$. Left picture: σSA with ξ_γ, $\gamma = 1/2$, $\kappa_\sigma = 3$. $\varphi^* \approx 0.17$, average of $\sigma^* \approx 1.92$. Right picture: σSA with ξ_γ, $\gamma = 0.45$, $\kappa_\sigma = 1$. $\varphi^* \approx 0.15$, average of $\sigma^* \approx 2.01$. Interval of measurement: $g = 1000 \dots 2000$.

and/or non-symmetrical two-point rules. A progress in this direction would open the possibility to determine γ and κ_σ on a solid ground.

6 Acknowledgments

The author is grateful to A. Irfan Oyman and Ralf Salomon for helpful comments. The author is Heisenberg fellow of the DFG, grant Be 1578/4-1.

References

1. H.-G. Beyer. Toward a Theory of Evolution Strategies: Some Asymptotical Results from the $(1,^+ \lambda)$-Theory. *Evolutionary Computation*, 1(2):165–188, 1993.
2. H.-G. Beyer. Toward a Theory of Evolution Strategies: Self-Adaptation. *Evolutionary Computation*, 3(3):311–347, 1996.
3. H.-G. Beyer. *Zur Analyse der Evolutionsstrategien.* Habilitationsschrift, University of Dortmund, 1996.
4. H.-G. Beyer. An Alternative Explanation for the Manner in which Genetic Algorithms Operate. *BioSystems*, 41:1–15, 1997.
5. I. Rechenberg. Evolutionsstrategien. In B. Schneider and U. Ranft, editors, *Simulationsmethoden in der Medizin und Biologie*, pages 83–114. Springer-Verlag, Berlin, 1978.
6. I. Rechenberg. *Evolutionsstrategie '94.* Frommann–Holzboog Verlag, Stuttgart, 1994.
7. H.-P. Schwefel. *Evolution and Optimum Seeking.* Wiley, New York, NY, 1995.