Reducing random fluctuations in mutative self-adaptation

Thomas Philip Runarsson

Science Institute, University of Iceland
tpr@hi.is

Abstract. A simple method of reducing random fluctuations experienced in step-size control under mutative self-adaptation is discussed. The approach taken does not require collective learning from the population, i.e. no recombination. It also does not require knowledge about the instantiations of the actual random mutation performed on the object variables. The method presented may be interpreted as an *exponential recency-weighted average* of trial strategy parameters sampled by a lineage.

1 Introduction

The paper discusses a method of reducing mean step-size (σ) fluctuations under mutative self-adaptation. Any proposed method capable of reducing these fluctuations will contribute to improved performance [3, p. 325]. One means of achieving this goal is by using collective learning based on the current population, that is by using recombination in the strategy parameter space. The method presented relates to this technique, but by using an *exponential recency-weighted average* of trial strategy parameters sampled via the lineage, *not* by the population. Alternatively, a more sophisticated *derandomized approach to self-adaptation* [4] could be employed. The method presented here is also related to this technique, but *does not* require knowledge about the realized steps (z),

$$\begin{aligned} z_\iota^{(g+1)} &= \sigma_\iota^{(g+1)} N(0,1) \\ x_\iota^{(g+1)} &= x_{r;\lambda}^{(g)} + z_\iota^{(g+1)} \end{aligned} \tag{1}$$

for the given step-size variation $\sigma_\iota^{(g+1)}$. Here x denotes the object variable and index $r;\lambda$ represents the individual assigned the rank r (the rth best out of λ), at any given discrete generation (g).

Reductions in step-size fluctuations are especially important in the case where individual step sizes are used, i.e. each object variable has its own corresponding mean step-size. For simple evolution strategies with small populations the self-adaptation, of individual step sizes, will often fail [5]. Another important domain of application are noisy environments. For example, the fitness function may be noisy [2], or simply the ranking itself may be noisy [7].

The work presented is motivated by the need for a simple and efficient parallel implementation of evolutionary algorithms using mutative self-adaptation. What is implied by a more efficient implementation is that the communication between individuals in a population should be minimal. The paper is organized as follows. In section 2 a short overview will be given on mutative step-size self-adaptation. This is followed by a section describing the proposed technique for reducing random fluctuations in mutative self-adaptation. An experimental study is then presented in section 4, in order to evaluate the technique empirically. Finally, a discussion and some conclusions are given in section 5.

2 Mutative step-size self-adaptation

In mutative step-size self-adaptation the mutation strength is randomly changed. It is only dependent on the parent's mutation strength, the parent step-size multiplied by a random number. This random number is commonly log-normally distributed. Other distributions are equally plausible [2,4] and the techniques described in the following section will still be applicable.

The *isotropic* mutative self-adaptation for a (μ, λ) evolution strategy (σES), using the log-normal update rule, is as follows [8],

$$\begin{aligned} \eta_\iota &= \sigma_{r;\lambda}^{(g)} \exp(N(0, \tau_o^2)), \\ \boldsymbol{x}_\iota^{(g+1)} &= \boldsymbol{x}_{r;\lambda}^{(g)} + \boldsymbol{N}(0, \eta_\iota^2), \quad \iota = 1, \ldots, \lambda \end{aligned} \tag{2}$$

where the parent is randomly sampled, $r \in [1, \mu]$, anew for each ι, $\tau_o \simeq c_{(\mu,\lambda)}/\sqrt{n}$ and the step size is updated in the following discrete generation by setting $\sigma_{r;\lambda}^{(g+1)} = \eta_{r;\lambda}$. Similarly, the *non-isotropic* mutative self-adaptation rule is [8],

$$\begin{aligned} \eta_{\iota,j} &= \sigma_{(r;\lambda),j}^{(g)} \exp\left(N(0, \tau'^2) + N_j(0, \tau^2)\right), \\ x_{\iota,j}^{(g+1)} &= x_{(r;\lambda),j}^{(g)} + N_j(0, \eta_{\iota,j}^2), \ \iota = 1, \ldots, \lambda, j = 1, \ldots, n \end{aligned} \tag{3}$$

where $\tau' = \varphi/\sqrt{2n}$ and $\tau = \varphi/\sqrt{2\sqrt{n}}$. The step-size is updated as before by setting $\boldsymbol{\sigma}_{r;\lambda}^{(g+1)} = \boldsymbol{\eta}_{r;\lambda}$.

The primary aim of the step-size control is to tune the search distribution so that maximal progress in maintained. For this some basic conditions for achieving optimal progress must be satisfied. The first lesson in self-adaptation is taken from the 1/5-*success rule* [6, p. 367]. The rule's derivation is based on the probability w_e that the offspring is better than the parent. This probability is calculated for the case where the optimal standard deviation is used $\hat{w}_e$, from which it is then determined that the number of trials must be greater than or equal to $1/\hat{w}_e$ if the parent using the optimal step-size is to be successful. Founded on the sphere and corridor models, this is the origin of the 1/5 value.

In a mutative step-size control, such as the one given by (2), there is no single *optimal* standard deviation being tested, but rather a series of trial step

sizes η_ι, $\iota = 1, \ldots, \lambda/\mu$ centered[1] around the parent step size $\sigma_{r;\lambda}$. Consequently, the number of trials may need to be greater than that specified by the 1/5-success rule. If enough trial steps for success are generated near the optimal standard deviation then this trial step-size will be inherited via the corresponding offspring. This offspring will necessarily also be the most likely to achieve the greatest progress and hence be the fittest. The fluctuations on $\sigma_{r;\lambda}$ (the trial standard deviations η_ι) and consequently also on the optimal mutation strength, will degrade the performance of the ES. The theoretical maximal progress rate is impossible to obtain. Any reduction of this fluctuation will therefore improve performance [3, p. 315]. If random fluctuations are not reduced, then a larger number of trials must be used (the number of offspring generated per parent) in order to guarantee successful mutative self-adaptation[2]. This may especially be the case for when the number of free strategy parameters increases, as in the non-isotropic case.

3 Reducing random fluctuations

Random fluctuations are reduced for the strategy-parameters, generated by the mutative rules (2) or (3), by letting the following weighted average be inherited to the next generation,

$$\sigma_{\iota,j}^{(g+1)} = \sigma_{(r;\lambda),j}^{(g)} + \chi\big(\eta_{\iota,j}^{(g+1)} - \sigma_{(r;\lambda),j}^{(g)}\big), \quad j = 1, \ldots, n_\sigma \tag{4}$$

where $n_\sigma = 1$ or $n_\sigma = n$ respectively. By considering the $(1, \lambda)$ strategy, one notices that this is an exponential recency-weighted average for a given lineage,

$$\sigma_{(1;\lambda),j}^{(g+1)} = (1-\chi)^{g+1}\sigma_{(1;\lambda),j}^{(0)} + \sum_{i=1}^{g+1} \chi(1-\chi)^{g+1-i}\eta_{(1;\lambda),j}^{(i)} \quad j = 1, \ldots, n_\sigma \tag{5}$$

since $(1-\chi)^{g+1} + \sum_{i=1}^{g+1} \chi(1-\chi)^{g+1-i} = 1$, and when $(1-\chi)$ is less than 1, the weight decreases exponentially according to the exponent of $(1-\chi)$. When $\chi = 1$ the method is equivalent to the canonical approach presented in the previous section. As the average number of trials generated increases (λ/μ) the more likely it will become that the optimal step-size is generated and successful. In this case a value of χ closer to 1 is reasonable. However, if the generated step-size is only an approximation of the optimal one, then a value of χ around 0.2 is more appropriate.

This averaging has the effect of reducing the step-size variations passed on between generations although the variation within a generation remains the same. If one would like to retain the same variation between generations a larger learning rate must be used. That is $E\big[\exp\big(N(0,\tau^2)\big)\big]$ should be the same as

[1] The expected median is $\sigma_{r;\lambda}$.

[2] Some algorithms force a lower bound on the step-size to avoid search stagnation, at this point the mutative self-adaptation is ineffective.

$E\big[1+\chi(\exp\big(N(0,\bar{\tau}^2)\big)-1)\big]$. For the isotropic mutation the corrected learning rate would then be,

$$\bar{\tau}_o^2 = 2\ln\left(\frac{1}{\chi}\Big(\exp\left(\frac{\tau_o^2}{2}\right)-(1-\chi)\Big)\right) \tag{6}$$

and so if $\chi = 1$ then $\bar{\tau}_o = \tau_o$. Similarly, for the non-isotropic mutation,

$$\bar{\varphi}^2 = \frac{2}{\upsilon}\ln\left(\frac{1}{\chi}\Big(\exp\left(\frac{\varphi^2\upsilon}{2}\right)-(1-\chi)\Big)\right) \tag{7}$$

where $\upsilon = \frac{1}{2n} + \frac{1}{2\sqrt{n}}$, and for $\chi = 1$ then $\varphi = \bar{\varphi}$.

The new update rules are equivalent to that of (2) and (3), with the corrected learning rates (6) and (7), and the inclusion of (4). For example, the new isotropic mutative self-adaptive rules becomes,

$$\begin{aligned}
\eta_\iota &= \sigma_{r;\lambda}^{(g)}\exp\big(N(0,\bar{\tau}_o^2)\big)\\
x_{\iota,j}^{(g+1)} &= x_{(r;\lambda),j}^{(g)} + N_j(0,\eta_\iota^2), \quad j=1,\ldots,n\\
\sigma_\iota^{(g+1)} &= \sigma_{r;\lambda}^{(g)} + \chi\Big(\eta_\iota - \sigma_{r;\lambda}^{(g)}\Big), \quad \iota=1,\ldots,\lambda.
\end{aligned} \tag{8}$$

where $0 < \chi \leq 1$ and $r \in [1,\mu]$.

At this point it may appear more intuitive, given the multiplicative nature of the mutative self-adaptation, to use geometric averaging. In this case the geometrical equivalent of (4) is,

$$\sigma_{\iota,j}^{(g+1)} = \Big(\sigma_{(r;\lambda),j}^{(g)}\Big)^{(1-\chi)}\Big(\eta_{\iota,j}^{(g+1)}\Big)^{\chi}, \quad j=1,\ldots,n_\sigma \tag{9}$$

and the learning rates may be simply corrected by setting $\bar{\tau} = \tau/\chi$ or $\bar{\varphi} = \varphi/\chi$. This is again a weighted average over the entire lineage. The same technique is used by *rescaled mutations* [6, p. 197], but with some subtle differences. The aim here is to reduce random fluctuations by averaging. The intention of rescaled mutation is, however, to produce larger σ changes which should result in a more reliable detection of the direction of change [2] for noisy fitness functions. The rescaled mutations use a variable $\kappa_\sigma \equiv 1/\chi$ and scale up the trial mutation strength by a factor $k \approx \sqrt{n}$ during mutation. This will not be done here.

In a noisy environment it may be necessary to reduce random fluctuations for the estimated object variables. This can be done in an equivalent manner to that of the strategy parameters in (4). The inherited object variables are then,

$$x_{\iota,j} = x_{(r;\lambda),j}^{(g)} + N_j(0,\eta_\iota^2) \tag{10}$$

$$x_{\iota,j}^{(g+1)} = x_{(r;\lambda),j}^{(g)} + \chi\Big(x_{\iota,j} - x_{(r;\lambda),j}^{(g)}\Big) \tag{11}$$

where *the fitness is computed based on object vector* $\boldsymbol{x}_\iota$ in (10). Alternatively, one may view this as some form of *genetic repair* [3] based on the lineage, *not* on the population.

4 Experimental Studies

In this section the behavior of the proposed approach will be examined for the sphere model,

$$\min \quad f(\boldsymbol{x}) = \sum_{k=1}^{n} x_k^2 \tag{12}$$

where $n = 30$, and the noisy sphere model [2],

$$\max \quad g(\boldsymbol{x}) = 1 - \sum_{k=1}^{n} x_k^2 + N(0, \sigma_\epsilon^2) \tag{13}$$

where $n = 10$. Although these models are simple, they do serve the present purpose of verifying the technique.

4.1 Noiseless sphere model

Isotropic mutation The first experiment aims to examine whether the canonical $(1, \lambda)\sigma$ES using the isotropic mutative self-adaptive rule, (2) or (8) for $\chi = 1$, can be improved by setting χ to a smaller value, say $\chi = 0.2$. The runs using

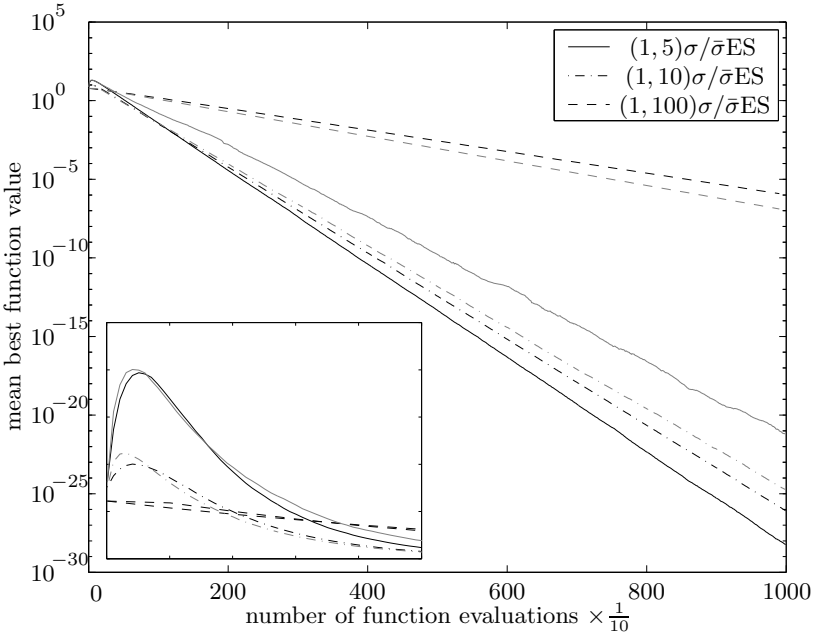

Fig. 1. Noiseless sphere model experiment using the isotropic mutation. The dark lines correspond to the new scheme $\bar{\sigma}$ES ($\chi = 0.2$) and the gray lines to the canonical σES. The sub-figure shows the first few generation in linear scale.

the averaging of σ according to (4) are denoted by $\bar{\sigma}$. The initial step-size $\sigma^{(0)}$ is $2/\sqrt{12}$ and the object variables are randomly initialized in $[-1, 1]$.

A total of 1000 independent runs are performed and the average best function value versus number of function evaluations plotted in figure 1, for $\lambda = 5, 10,$ and 100. For $\chi = 0.2$ there is a performance enhancement for the $(1, 5)\bar{\sigma}$ES and $(1, 10)\bar{\sigma}$ES strategies. However, as the number of trials is increased to 100, the canonical approach $(1, 100)\sigma$ES (or $(1, 100)\bar{\sigma}$ES with $\chi = 1$) is better. This is as predicted, the greater the number of trials performed the likelier it becomes that the trial mutation created is close the optimal. As a consequence more weight should be put on this estimate, that is χ should be closer to 1.

Notice that the performance of the $(1, 10)$ is better than that of the $(1, 5)$ strategy[3] in the canonical case, but this is not the case when the random fluctuations have been reduced by setting $\chi = 0.2$.

The same experiment is again repeated and this time the arithmetical style averaging of (4) is compared with the geometrical style averaging of (9). This version is denoted by $(1, \lambda)\hat{\sigma}$ES. In this experiment an attempt is made to optimize χ. It is found that in the case of $\bar{\sigma}$ES the performance is best and

[3] Theoretically, for the sphere model, optimal progress is obtained when $\lambda \approx 5$ and the mutation strength is optimal.

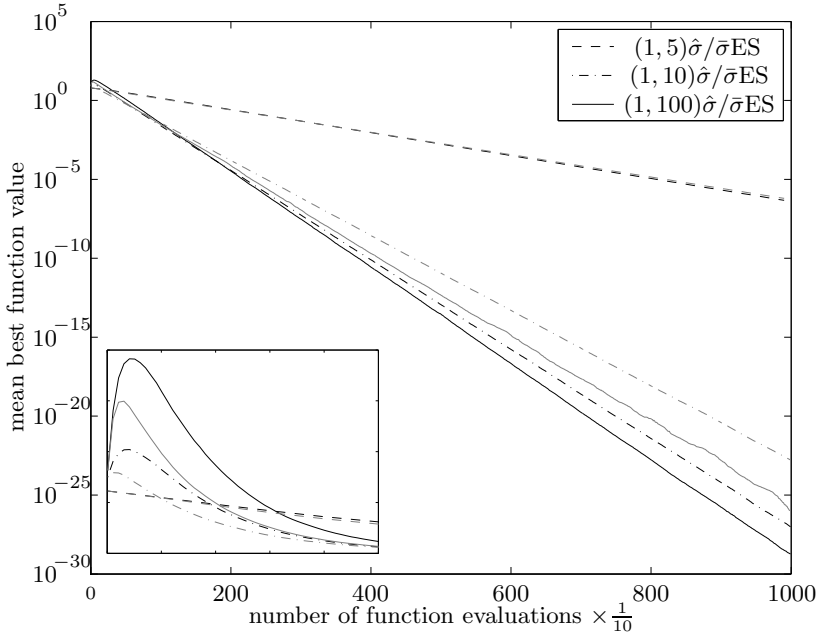

Fig. 2. Noiseless sphere model experiment using the isotropic mutation. The dark lines correspond to the weighted average $\bar{\sigma}$ES ($\chi = 0.3$) and the gray lines to the geometrical average $\hat{\sigma}$ES ($\chi = 0.5$). The sub-figure shows the first few generation in linear scale.

similar in the range $0.2 < \chi < 0.4$. For the $\hat{\sigma}$ES the best performance is in the range $0.4 < \chi < 0.6$. The $\bar{\sigma}$ES ($\chi = 0.3$) and $\hat{\sigma}$ES ($\chi = 0.5$) are therefore compared. The average of 1000 independent runs are plotted in figure 2, where one can see that, for the noiseless sphere model, the $\bar{\sigma}$ES is slightly better.

Non-isotropic mutation In the second experiment the first experiment is repeated using the non-isotropic mutation with $\varphi = 1$. Again 1000 independent runs are performed and the result plotted in figure 3. This time the mutative self-adaptation fails for the canonical $(1, 100)\sigma$ES – the search stagnates. In order to avoid this problem two new strategies are introduced, both using recombination in the strategy parameter space. The first one uses intermediate recombination,

$$\eta_{\iota,j} = \Big(\frac{1}{\mu}\sum_{k=1}^{\mu}\sigma^{(g)}_{(k;\lambda),j}\Big)\exp\big(N(0,\tau'^2) + N_j(0,\tau^2)\big) \tag{14}$$

and is denoted by $(\mu/\mu_{(i)}, \lambda)$. The non-isotropic mutation is included on the right-hand-side of (14). The second is an arithmetical recombination, including the mutation,

$$\eta_{\iota,j} = \tfrac{1}{2}\big(\sigma^{(g)}_{(r;\lambda),j} + \sigma^{(g)}_{(k;\lambda),j}\big)\exp\big(N(0,\tau'^2) + N_j(0,\tau^2)\big) \tag{15}$$

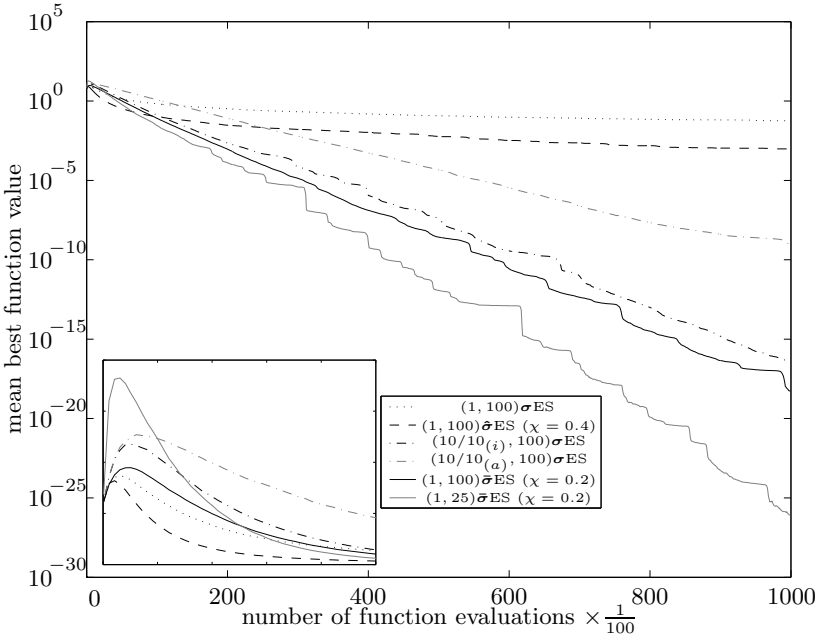

Fig. 3. Noiseless sphere model experiment using the non-isotropic mutation. The subfigure shows the first few generation in linear scale. The legend lists the different strategies in order from worst to best.

denoted $(\mu/\mu_{(a)}, \lambda)$, where $k \in [1, \mu]$ is sampled randomly and anew for each j. By using recombination on the strategy parameters, search stagnation is avoided. It is also noticed that arithmetical recombination performs better than intermediate recombination. However, in terms of the number of function evaluation needed, the newly proposed scheme still performs best. For the $(1, 100)$ strategy the search does not stagnate for the case when $\chi = 0.2$. In order to achieve faster progress the number of trial is reduced down to a $(1, 25)$ strategy, but going down to $(1, 10)$ will result in failure[4]. Clearly a greater number of trials will be needed for success as the number of free parameters increases. An attempt to use the geometrical averaging of (9) resulted in faster initial progress and then search stagnation. It seems that arithmetical averaging has a tendency towards larger mutation strengths than geometrical averaging.

The final experiment introduces recombination in the object parameter space. Two versions are tested. The first is again an intermediate recombination and mutation,

$$x_{\iota,j}^{(g+1)} = \frac{1}{\mu} \sum_{k=1}^{\mu} x_{(k;\lambda),j}^{(g)} + N_j(0, \eta_{\iota,j}^2) \tag{16}$$

[4] Unpredictable behavior is observed, where the step-size may keep growing in size.

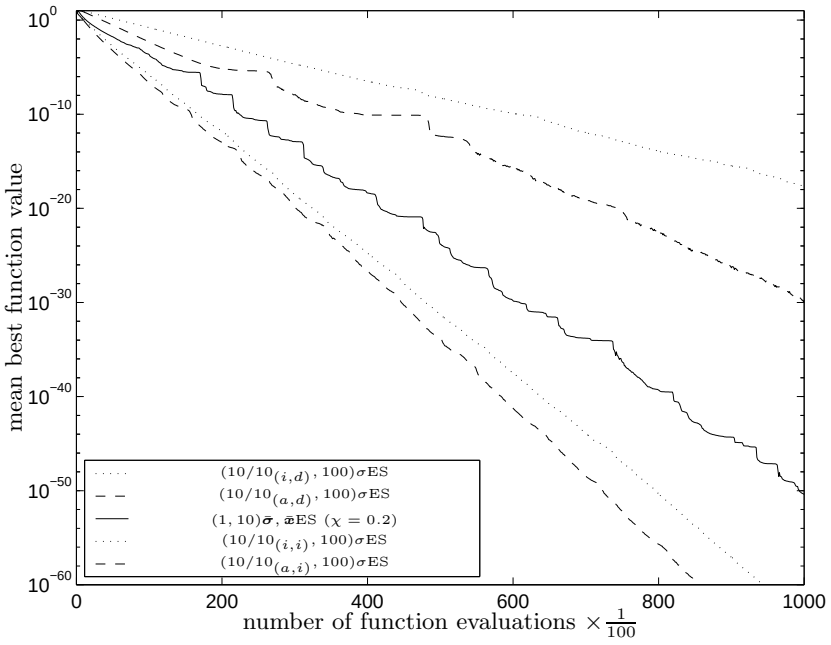

Fig. 4. Noiseless sphere model using non-isotropic mutation and recombination in both strategy and object parameter space. The versions listed in the legend are ordered based on performance, from worst to best.

denoted here by $(\mu/\mu_{(,i)}, \lambda)$ and the second dominant recombination and mutation,

$$x_{\iota,j}^{(g+1)} = x_{(k;\lambda),j}^{(g)} + N_j(0, \eta_{\iota,j}^2) \tag{17}$$

where $k \in [1, \mu]$ is sampled randomly and anew for each j. This version is denoted by $(\mu/\mu_{(,d)}, \lambda)$. The recombinations in object parameter space are used in conjunction with recombination in strategy parameter space resulting in four different strategies: $(\mu/\mu_{(i,d)}, \lambda)$, $(\mu/\mu_{(a,d)}, \lambda)$, $(\mu/\mu_{(i,i)}, \lambda)$, and $(\mu/\mu_{(a,i)}, \lambda)$. In order to make a fair comparison, the new technique is allowed to use the weighted average for the object variables as defined by (11). It is interesting to observe that now it becomes possible to switch to a $(1, 10)$ scheme without any problems. Using geometrical averaging (9) will also result in search stagnation for this experiment and is therefore omitted. The results, averaged over 1000 independent runs are presented in figure 4. The best results are, nevertheless, achieved using intermediate recombination on the object variables.

4.2 Noisy sphere model

The final experiment is based on simulations performed in [2]. This involves the noisy sphere model (13), where the object parameters are initialized in $[-\frac{1}{\sqrt{10}}, \frac{1}{\sqrt{10}}]$, and an isotropic mutation with $\sigma^{(0)} = 1$ and $\tau_o = 0.7$ is used.

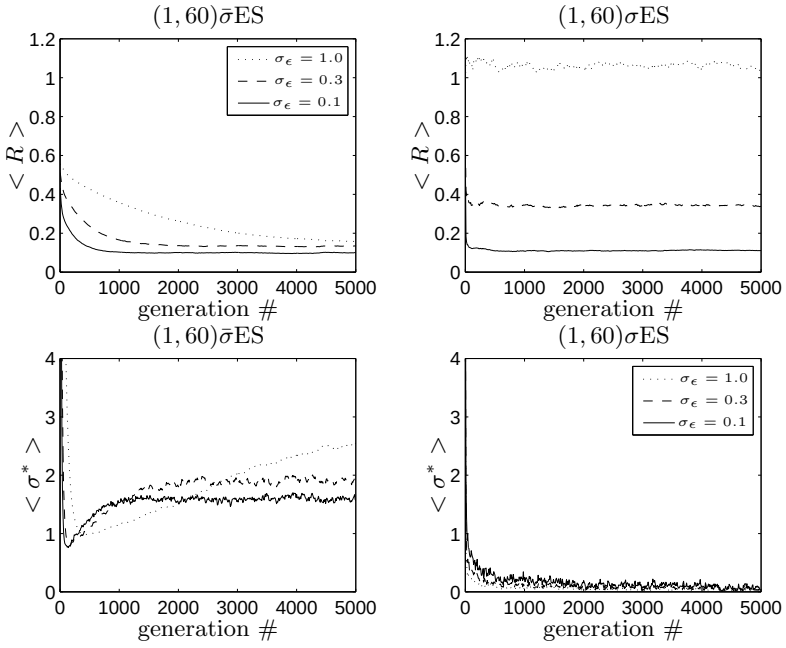

Fig. 5. Noisy sphere model experiments. The plots on the left are from the new technique $\bar{\sigma}$ES and on the right the canonical approach σES.

In this experiment the value of χ is varied as a function of the noise level σ_ϵ in (13). The noise levels σ_ϵ are 0.1, 0.3, and 1.0 and the corresponding values of χ chosen are 0.1, 0.06, and 0.02. For the three different noise levels 300 independent runs are performed using a $(1, 60)$ strategy. The new method uses the averaging (11) for the object variables.

Instead of plotting mean function values, the mean distance to the optimum, $R = \|x\|$, is depicted versus the generation number. Additionally, the mean normalized step-size, $\sigma^* = n\sigma/R$, is plotted against the generation number. The behavior of the new scheme is similar to that of the rescaled mutations technique in [2], as seen in figure 5. The canonical approach on the other hand has difficulties coping with the noisy function.

5 Discussion and conclusion

The update rule of (4) is not unlike the update rules used in reinforcement learning, where χ is a step-size parameter. When using a constant χ more weight is put on the strategy parameters used recently rather than long-past ones. This makes sense when tracking the nonstationary problem of self-adaptation. Furthermore, the greater the number of trials generated ($\approx \lambda/\mu$) the more reliable the selected strategy parameters become.

For noisy functions smaller χ values were used. At the later stage of search, say after generation g', the problem of self-adaptation becomes essentially stationary. In this case it may be reasonable to set $\chi = 1/(g - g')$, and so all estimates after g' are equally weighted. That is, (4) becomes an incremental implementation of a simple-average.

A technique for reducing of random fluctuations in mutative self-adaptation has been presented and some experimental results given. Self-adaptive rules using individual step sizes commonly fail. The proposed method may be a simple means of alleviating this problem.

References

1. H.-G. Beyer. Toward a theory of evolution strategies: self-adaptation. *Evolutionary Computation*, 3(3):311–347, 1996.
2. H.-G. Beyer. Evolutionary algorithms in noisy environments: theoretical issues and guidelines for practice. *Computer Methods in Applied Mechanics and Engineering*, 186(2–4):239–267, 2000.
3. H.-G. Beyer. *The Theory of Evolution Strategies*. Springer-Verlag, Berlin, 2001.
4. N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution stategies. *Evolutionary Computation*, 2(9):159–195, 2001.
5. A. Ostermeier, A. Gawelczyk, and N. Hansen. A derandomized approach to self-adaptation of evolution strategies. *Evolutionary Computation*, 2(4):369–380, 1995.
6. I. Rechenberg. *Evolutionstrategie '94*. Frommann-Holzboog, Stuttgart, 1994.
7. T. P. Runarsson and X. Yao. Stochastic ranking for constrained evolutionary optimization. *IEEE Transactions on Evolutionary Computation*, 4(3):284–294, September 2000.
8. H.-P. Schwefel. *Evolution and Optimum Seeking*. Wiley, New-York, 1995.