On multiplicative noise models for stochastic search

Mohamed Jebalia, Anne Auger

On Multiplicative Noise Models for Stochastic Search

Mohamed Jebalia[1] and Anne Auger[1,2]

[1] TAO Team, INRIA Saclay, Université Paris Sud, LRI, 91405 Orsay cedex, France
[2] Microsoft Research-INRIA Joint Centre 28, rue Jean Rostand, 91893 Orsay Cedex, France
mohamed.jebalia@lri.fr, anne.auger@inria.fr

Abstract. In this paper we investigate multiplicative noise models in the context of continuous optimization. We illustrate how some intrinsic properties of the noise model imply the failure of reasonable search algorithms for locating the optimum of the noiseless part of the objective function. Those findings are rigorously investigated on the $(1+1)$-ES for the minimization of the noisy sphere function. Assuming a lower bound on the support of the noise distribution, we prove that the $(1+1)$-ES diverges when the lower bound allows to sample negative fitness with positive probability and converges in the opposite case. We provide a discussion on the practical applications and non applications of those outcomes and explain the differences with previous results obtained in the limit of infinite search-space dimensionality.

1 Introduction

In many real-world optimization problems, objective functions are perturbed by noise. Evolutionary Algorithms (EAs) have been proposed as effective search methods in such contexts [5, 10]. A noisy optimization problem is a rather general optimization problem where for each point x of the search space, we can observe $f(x)$ perturbed by a random variable or in other words for a given x we can observe a distribution of possible objective values. The goal is in general to converge to the minimum of the averaged value of the observed random variable. One type of noise encountered in real-world problems is the so-called multiplicative noise where the noiseless objective function $f(x)$ is perturbed by the addition of a noise term proportional to f, ie. the noisy objective function $\mathcal{F}$ reads

$$\mathcal{F}(x) = f(x)(1+\mathcal{N}) \tag{1}$$

where $\mathcal{N}$ is the noise random variable, sampled independently at each new evaluation of a point. Such noise models are in particular used to benchmark robustness of EAs with respect to noise [12]. The focus here is continuous optimization (that will be minimization) where f maps a continuous search space, ie. a subset of $\mathbb{R}^d$, into $\mathbb{R}$. The EAs specifically designed for continuous optimization are usually referred as Evolution Strategies (ES), where a set of candidate solutions evolves by first applying Gaussian perturbations (mutations) to the current solutions then selection. ES in noisy environments have been studied by Arnold and Beyer [8, 3, 1]. Multiplicative noise has been investigated in the case of $\mathcal{N}$ being normally distributed with a standard deviation scaled by $1/d$ for a $(1+1)$-ES [4], (μ, λ)-ES [3, 7], $(\mu/\mu_{\mathrm{I}}, \lambda)$-ES [2] and f being the sphere function $f(x) = \|x\|^2$. Under the assumption that d goes to infinity, Arnold and Beyer

show, for $f(x) = \|x\|^2$, positive expected fitness gain for the elitist $(1+1)$-ES (if the fitness of the parent is not reevaluated in the selection step which is the case of our study). This implies a decrease of the expectation of the square distance to the optimum (here zero). However, convergence of the $(1+1)$-ES to the optimum of the noiseless part of the noisy objective function seems to be unlikely if the noise random variable takes values smaller than -1 as we illustrate now on a simple example. Assume indeed that $\mathcal{N}$ takes three distinct values (each with probability $1/3$) $+\gamma$, 0 and $-\gamma$ where γ satisfies $\gamma > 1$. For a given $x \in \mathbb{R}^d$, the objective function $\mathcal{F}(x)$ takes 3 different values (each with probability $1/3$) $(1+\gamma)\|x\|^2$, $\|x\|^2$, $(1-\gamma)\|x\|^2$. The last term is strictly negative for x non equal to zero. Therefore, if one negative objective function value is reached, the $(1+1)$-ES that can only accept solutions having a lower objective function value will never accept solutions closer to the optimum since they have higher objective function values[1]. On the contrary the $(1+1)$-ES will diverge log-linearly [2], i.e. the logarithm of the distance to the optimum will increase linearly.

Starting from this observation, we investigate how the properties of the support of the noise distribution relate to convergence or divergence of stochastic search algorithms and can make the convergence to the optimum of the noiseless part of the objective function hopeless for reasonable search algorithms. Compared to previous approaches, we do not make use of asymptotic assumptions, trying to capture effects that were not observed before [4]. In Section 2, we detail the noise model considered and show experimentally on a $(1+1)$-ES that divergence and convergence is determined by the probability to sample noise values smaller than -1. In Section 3, we provide some simple proofs of convergence and divergence for the $(1+1)$-ES. In Section 4 we discuss the results and explain where the difference with the results in [4] stems from.

2 Motivations

Elementary remarks on the noise model We investigate multiplicative noise models as defined in Eq. 1 where $\mathcal{N}$ is a random variable with finite mean and $f(x)$ is the noiseless function that we assume positive in the sequel. We also assume that $1 + E(\mathcal{N}) > 0$ such that the argmin[3] of the expected value of $\mathcal{F}(x)$ is the argmin of $f(x)$. Often, the distribution of $\mathcal{N}$ is assumed symmetric, implying then that $1+E(\mathcal{N}) = 1 > 0$. Though one might think that this condition is sufficient such that minimizing $\mathcal{F}(x)$ amounts to minimizing $f(x)$, we sketch now, why divergence to ∞ of the distance to the optimum happens if $1 + \mathcal{N}$ can take negative values.

Assume that $f(x)$ converges to infinity when $\|x\|$ goes to ∞; typically $f(x)$ can be the famous sphere function $f(x) = \|x\|^2$ and assume that the random variable $\mathcal{N}$ admits a density function $p_{\mathcal{N}}(t), t \in \mathbb{R}$ whose support is an interval $[m_{\mathcal{N}}, M_{\mathcal{N}}[$, i.e. $\mathcal{N} \in [m_{\mathcal{N}}, M_{\mathcal{N}}[$ and the probability that $\mathcal{N} \in [a, b]$ for any $m_{\mathcal{N}} \leq a < b \leq M_{\mathcal{N}}$ is

[1] Their absolute value is smaller though. However, trying to minimize the absolute value of $\mathcal{F}$ instead is not a solution in general, consider for instance the function $f(x) = (\|x\|^2 + 1)(1 + \mathcal{N})$.

[2] We will say that a sequence $(d_n)_n$ diverges (resp. converges) log-linearly if there exists $c > 0$ (resp. $c < 0$) such that $\lim_n \frac{1}{n} \ln(d_n) = c$.

[3] The argmin of an objective function $x \mapsto h(x)$ are defined as $h(\arg\min_x h) = \min_x h(x)$

strictly positive. The function $g_{m_\mathcal{N}}(x) = f(x)(1 + m_\mathcal{N})$ gives a lower bound of the values that can be reached by the noisy fitness function for different instantiations of the random variable $\mathcal{N}$ (because f is positive). For a given x, $\mathcal{F}(x)$ can take values with positive probability in any open interval of $]g_{m_\mathcal{N}}(x), f(x)[$ [4].

In Fig. 1 are depicted a cut of $f(x) = \|x\|^2$ and $g_{m_\mathcal{N}}(x) = f(x)(1 + m_\mathcal{N})$ for $m_\mathcal{N}$ equals -0.5 and -1.5. The position of $m_\mathcal{N}$ with respect to -1 determines whether $g_{m_\mathcal{N}}(x)$ is convex or concave: for $m_\mathcal{N} > -1$, $g_{m_\mathcal{N}}(x)$ is convex, converging to infinity when $\|x\|$ goes to ∞ and for $m_\mathcal{N} < -1$, $g_{m_\mathcal{N}}(x)$ is concave, converging to minus infinity when $\|x\|$ goes to ∞. Minimizing $g_{m_\mathcal{N}}(x)$ in the case of $m_\mathcal{N} < -1$ means that $\|x\|$ is diverging to $+\infty$ and $g_{m_\mathcal{N}}(x)$ is diverging to $-\infty$ which is the opposite of the behavior one would like since we are aiming at minimizing the non-noisy function $f(x) = \|x\|^2$. Note that in the example sketched in the introduction with $\mathcal{N}$ taking

Fig. 1. [Dashed Line] One dimensional cut of $f(x) = \|x\|^2$ along one arbitrary unit vector. [Straight line] Left: One dimensional cut of $g_{-0.5}(x) = \|x\|^2(1 - 0.5)$. Right: One dimensional cut of $g_{-1.5}(x) = \|x\|^2(1 - 1.5)$. For a given x, the noisy-objective function can, in particular, take any value between the dashed curve and the straight curve.

the values γ, $-\gamma$ and 0, the plot of $\|x\|^2$ and $(1 - \gamma)\|x\|^2$ for $\gamma = 1.5$ are the curves represented in Fig 1 (right).

Experimental observations We investigate now numerically how the "shape" of the lower bound might affect the convergence. For this purpose we use a $(1, 5)$-ES and a $(1 + 1)$-ES using scale-invariant adaptation scheme for the step-size[5].

We investigate the function $\mathcal{F}_s(x) = \|x\|^2(1 + \mathcal{N})$ when the noise $\mathcal{N}$ is uniformly distributed in the ranges $[-0.5, 0.5]$ and $[-1.5, 1.5]$ respecitvely denoted $U_{[-0.5,0.5]}$ and $U_{[-1.5,1.5]}$. This latter noise corresponds to the concave lower bound $g_{-1.5}(x) = -0.5\|x\|^2$ plotted in Fig. 1. In Figure 2, the result of 10 independent runs of the $(1, 5)$-ES (10 upper curves of each graph) in dimension $d = 10$ are plotted for the non-noisy sphere (left), $f(x) = \|x\|^2(1 + U_{[-0.5,0.5]})$ (middle) and $f(x) = \|x\|^2(1 + U_{[-1.5,1.5]})$ (right). Not too surprisingly, we observe a drastic difference in the last two cases: the algorithm converges to the optimum for the noise $U_{[-0.5,0.5]}$ whereas the distance to the

[4] Note that $g_{m_\mathcal{N}}(x) < f(x)$ iff $m_\mathcal{N} < 0$.

[5] In a scale-invariant ES, the step-size is set at each iteration as a (strictly positive) constant σ times the distance to the optimum. This artificial adaption scheme (since in practice one does not know the distance to the optimum!) allows to achieve optimal convergence rate for ES and is therefore very interesting from a theoretical point of view. The algorithm is mathematically defined in Section 3.

optimum increases (log)-linearly for the noise having a lower bound smaller than -1 [6]. Comparing the left and middle graphs we also observe, as expected, that the presence of noise slows down the convergence. On the same figure (lower curves of the graphs), the

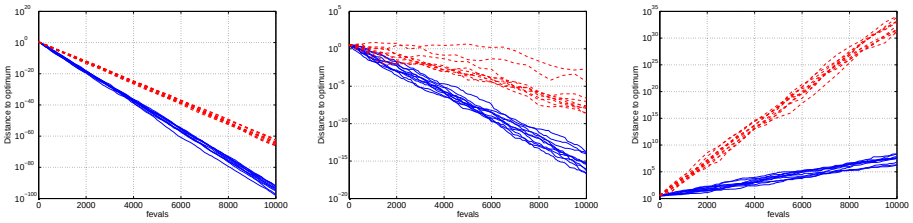

Fig. 2. Distance to the optimum (in log-scale) versus number of evaluations. Ten independent runs for the scale-invariant $(1, 5)$-ES (10 upper curves of each graph) and $(1 + 1)$-ES (10 lower curves of each graphs) with $d = 10$ and $\sigma = 1/d$. Left: $f(x) = \|x\|^2$. Middle: $f(x) = \|x\|^2(1+U_{[-0.5,0.5]})$. Right: $f(x) = \|x\|^2(1 + U_{[-1.5,1.5]})$.

results of 10 independent runs of the $(1+1)$-ES are plotted for the three same functions. As in the case of the comma strategy we observe that the $(1 + 1)$-ES diverges in the case of the noise $U_{[-1.5,1.5]}$ and that, when convergence occurs, the convergence rate is slower in presence of noise. Last, we investigate numerically the $(1 + 1)$-ES where $\mathcal{N}$ is normally distributed and in particular unbounded. This corresponds to the case investigated in [4]. We carry out tests for a standard deviation of the Gaussian noise equals 0.1, 2 and 10. Results are presented in Fig. 3. We observe convergence when the standard deviation of the noise equals 0.1 and divergence in the last two cases.

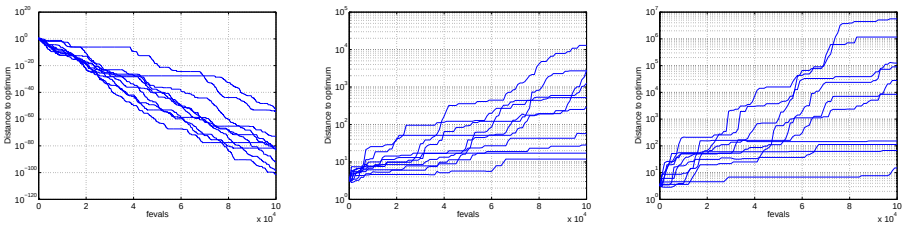

Fig. 3. Ten independent runs for the scale-invariant $(1 + 1)$-ES with a normally distributed noise: on $f(x) = \|x\|^2(1 + \sigma_\epsilon \mathcal{N}(0, 1))$ with σ_ϵ equals 0.1 (left), 2 (middle) and 10 (right) for $d = 10$ and $\sigma = 1/d$.

3 Convergence and divergence of the $(1 + 1)$-ES

In this section, we provide a simple mathematical analysis of the convergence and divergence of the $(1 + 1)$-ES experimentally observed in the previous section. We focus for

[6] However, contrary to what we will see for the $(1 + 1)$-ES, we do not state that "-1" is a limit value between convergence and divergence in the case of $(1, \lambda)$-ES. Indeed convergence and divergence depends on the intrinsic properties of the noise and on λ and σ as well (see [8]).

the sake of simplicity on lower bounded noise, i.e. the support of the noise is included in $[m_{\mathcal{N}}, +\infty[$. We prove that the $(1+1)$-ES minimizing the noisy sphere converges if $m_{\mathcal{N}} > -1$ and diverges if $m_{\mathcal{N}} < -1$. The proofs are rather simple and rely on the Borel-Cantelli Lemma. For the sake of readability we provide here a sketch of the demonstrations and send the proofs with the technical details in the Appendix of the paper.

Mathematical model for the $(1+1)$-ES The $(1+1)$-ES is a simple ES which evolves a single solution. At an iteration n, this solution denoted X_n, is called parent. The minimization of a given function f mapping $\mathbb{R}^d$ $(d \geq 1)$ into $\mathbb{R}$ using the $(1+1)$-ES algorithm is as follows: At every iteration n, the parent X_n is perturbated by a Gaussian random variable $\sigma_n \mathrm{N}_n$, where σ_n is a strictly positive value called step-size and $(\mathrm{N}_n)_n \in \mathbb{R}^d$ are independent realizations of a multivariate isotropic normal distribution on $\mathbb{R}^d$ denoted by $\mathrm{N}(0, I_d)$ [7]. The resulting offspring $\mathrm{X}_n + \sigma_n \mathrm{N}_n$ is accepted if and only if its fitness value is smaller than the one of its parent X_n. One of the key points in minimization using isotropic ES[8] is how to adapt the sequence of step-sizes (σ_n). Convergence of the $(1+1)$-ES is sub-log-linear bounded below by an explicit log-linear rate. This lower bound for the convergence rate is attained for the specific case of the sphere function and scale-invariant algorithm where the step-size is chosen proportional to the distance to the optimum, i.e. $\sigma_n = \sigma \|\mathrm{X}_n\|$ where σ is a strictly positive constant [6, 9]. The scale-invariant algorithm has a major place in the theory of ES since it corresponds to the dynamic algorithm implicitly studied in the one-step analysis computing progress rate or fitness gain [11, 8]. Using this adaptation scheme, the algorithm is referred to as the scale-invariant $(1+1)$-ES and the offspring writes as $\mathrm{X}_n + \sigma \|\mathrm{X}_n\| \mathrm{N}_n$. The noisy sphere function is denoted

$$\mathcal{F}_s(x) = \|x\|^2 (1 + \mathcal{N}) \tag{2}$$

where we assume that the random variable $\mathcal{N}$ has a finite expectation such that $E(\mathcal{N}) > -1$ and admits a density function $p_{\mathcal{N}}$ which lies in the range $[m_{\mathcal{N}}, M_{\mathcal{N}}[$ where $-\infty < m_{\mathcal{N}} < M_{\mathcal{N}} \leq +\infty$, $M_{\mathcal{N}} > -1$ and $m_{\mathcal{N}} \neq -1$. The normalized noisy part $\mathcal{N}$ of the noisy sphere function will be called normalized overvaluation of x. The term normalized overvaluation was already defined in [4] where it corresponds to the opposite of the quantity considered here up to a factor $d/2$. The minimization of this function using the scale-invariant $(1+1)$-ES is mathematically modeled by the sequence of parents (X_n) with their relative noisy fitnesses $(\mathcal{F}_s(\mathrm{X}_n))$ and normalized overvaluations (O_n). At an iteration n, the fitness of the parent is $\mathcal{F}_s(\mathrm{X}_n) = \|\mathrm{X}_n\|^2 (1 + O_n)$ and the fitness of an offspring equals $\|\mathrm{X}_n + \sigma \|\mathrm{X}_n\| \mathrm{N}_n\|^2 (1 + \mathcal{N}_n)$ where $(\mathcal{N}_n)_n$ is a sequence of independent random variables with $\mathcal{N}$ as a common law. Let $\mathrm{X}_0 \in \mathbb{R}^d$ be the first parent with a normalized overvaluation O_0 sampled from the distribution of $\mathcal{N}$. Then

[7] $\mathrm{N}(0, I_d)$ is the multivariate normal distribution with mean $(0, \ldots, 0) \in \mathbb{R}^d$ and covariance matrix the identity I_d.

[8] ES are called isotropic when the covariance matrix of the distribution of the random vectors $(\mathrm{N}_n)_n$ is I_d.

the update of X_n for $n \geq 0$ writes as:

$$\begin{aligned} \mathrm{X}_{n+1} &= \mathrm{X}_n + \sigma \|\mathrm{X}_n\| \mathrm{N}_n \text{ if } \|\mathrm{X}_n + \sigma \|\mathrm{X}_n\| \mathrm{N}_n\|^2 (1 + \mathcal{N}_n) < \|\mathrm{X}_n\|^2 (1 + O_n) \ , \\ &= \mathrm{X}_n \text{ otherwise} \, , \end{aligned} \tag{3}$$

and the new normalized overvaluation O_{n+1} is then:

$$\begin{aligned} O_{n+1} &= \mathcal{N}_n \text{ if } \|\mathrm{X}_n + \sigma \|\mathrm{X}_n\| \mathrm{N}_n\|^2 (1 + \mathcal{N}_n) < \|\mathrm{X}_n\|^2 (1 + O_n) \ , \\ &= O_n \text{ otherwise} \, . \end{aligned} \tag{4}$$

The $(1+1)$-ES algorithm ensures that the sequence relative to the function to minimize (which is $(\mathcal{F}_s(\mathrm{X}_n))$ in our case) decreases. This property makes the theoretical study of the $(1 + 1)$-ES easier than that of comma strategies. Our study shows that the behavior of the scale-invariant $(1 + 1)$-ES on the noisy sphere function (2) depends on the lower bound of the noise $m_{\mathcal{N}}$.

Theorem 1. *The $(1 + 1)$-ES minimizing the noisy sphere (Eq. 2) defined in Eq. 3 converges to zero if $m_{\mathcal{N}} > -1$ and diverges to infinity when $m_{\mathcal{N}} < -1$.*

Proof. The proof of this theorem is split in two cases $m_{\mathcal{N}} > -1$ and $m_{\mathcal{N}} < -1$ respectively investigated in Proposition 1 and Proposition 2. □

The proofs heavily rely on the second Borel-Cantelli Lemma that we recall below. But first, we need a formal definition of 'infinitely often (i.o.)': Let q_n be some statement, eg. $|a_n - a| > \epsilon$. We say (q_n i.o.) if for all n, $\exists\ m \geq n$ such that q_m is true. Similarly, for a sequence of events A_n in a probability space, (A_n i.o.) equals $\{w | w \in A_n \text{ i.o.}\} = \cap_{n \geq 0} \cup_{m \geq n} A_m := \overline{\lim}\ A_n$. The second Borel-Cantelli Lemma (BCL) states that:

Lemma 1. *Let $(A_n)_{n \geq 0}$ be a sequence of events in some probability space. If the events A_n are independent and verify $\sum_{n \geq 0} P(A_n) = +\infty$ then $P(\overline{\lim}\ A_n) = 1$.*

Proposition 1 (Convergence for $m_{\mathcal{N}} > -1$). *If $m_{\mathcal{N}} > -1$, the sequences $(\mathcal{F}_s(\mathrm{X}_n))$ and $(\|\mathrm{X}_n\|)$ converge to zero almost surely.*

Sketch of the proof (see detailed proof in Appendix) The condition $m_{\mathcal{N}} > -1$ ensures that the decreasing sequence $(\mathcal{F}_s(\mathrm{X}_n))$ is positive. Therefore it converges. Besides the sequence $(\|\mathrm{X}_n\|)$ is upper bounded by $\theta := \mathcal{F}_s(X_0)/(1 + m_{\mathcal{N}})$ as shown in Fig. 1 (left). Consequently, the probability to hit, at each iteration n, a fixed neighborhood of 0 is lower bounded by a strictly positive constant. Applying BCL we deduce the convergence of the sequence $(\mathcal{F}_s(\mathrm{X}_n))$ (and then that of $(\|\mathrm{X}_n\|)$) to zero. □

Proposition 2 (Divergence for $m_{\mathcal{N}} < -1$). *If $m_{\mathcal{N}} < -1$, the sequence $(\mathcal{F}_s(\mathrm{X}_n))$ diverges to $-\infty$ almost surely and the sequence $(\|\mathrm{X}_n\|)$ diverges to $+\infty$ almost surely.*

Sketch of the proof (see detailed proof in Appendix) As $1 + m_{\mathcal{N}} < 0$, the probability to sample a noise $\mathcal{N}_n$ such that $1 + \mathcal{N}_n < 0$ is striclty positive. Therefore there exists an integer n_1 such that for all $n \geq n_1$, $\mathcal{F}_s(\mathrm{X}_n) < 0$. Consequently $(\|\mathrm{X}_n\|)$ is lower bounded by A as illustrated in Fig. 1 (right) where the straight horizontal line represents

the slope $y = \mathcal{F}_s(X_{n_1})$. Besides, the probability to have $\mathcal{F}_s(\mathrm{X}_n)$ as small as we want is lower bounded by a strictly positive constant which gives with BCL the divergence of the sequence $(\mathcal{F}_s(\mathrm{X}_n))$ to $-\infty$, i.e. the sequence $(\|\mathrm{X}_n\|)$ diverges to $+\infty$. □
Remark that for the example sketched in the introduction where $\mathcal{N}$ takes the 3 different values γ, 0 and $-\gamma$ and under the condition $\gamma > 1$ the proof of divergence will follow the same lines.

4 Discussion and conclusion

We conclude from Theorem 1 that what matters for convergence or divergence of the $(1 + 1)$-ES in the case of noisy objective function with positive noiseless part is the position of the lower bound $m_{\mathcal{N}}$ of the noise distribution $\mathcal{N}$ with respect to -1 or in other words the existence or not of possible negative fitness values. This result applies in particular when $\mathcal{N}$ equals a truncated normal distribution, i.e. $\mathcal{N} = \sigma_\epsilon \mathcal{N}(0, 1)1_{[-a,a]}$[9] for any a and σ_ϵ positive. Whenever $\sigma_\epsilon a > 1$, Proposition 2 applies and the $(1 + 1)$-ES diverges.

Those results might appear in contradiction with those of Arnold and Beyer [4] proving that the expected fitness gain is positive−and therefore convergence in mean holds for the scale-invariant ES−for a noise distributed according to a normal distribution. In their model, Arnold and Beyer scale the standard deviation of the noise σ_ϵ with $1/d$, i.e. when $d \to \infty$, σ_ϵ converges to 0. The largest value for the normalized σ^*_ϵ in [4, Fig 5, 6, 8], for $d = 80$ corresponds to a standard deviation of 0.05 for which the probability to have $(1 + 0.05\,\mathcal{N}) < 0$ is upper bounded by $10^{-88\ (10)}$, i.e. relatively unlikely! Therefore though they consider some unbounded noise having a support in $\mathbb{R}$, the normalization of the standard deviation of the noise implies a so small probability to sample $1+\mathcal{N}$ below -1 that the unbounded noise reduces to the case of convergence where $m_{\mathcal{N}} > -1$. The same conclusion holds for the numerical example given in Section 2, Fig. 3 (left) where the standard deviation of 0.1 corresponds to a probability to have $(1 + 0.1\,\mathcal{N}) < 0$ lower bounded by 10^{-23}. Therefore though the theory predicts divergence as soon as $m_{\mathcal{N}} < -1$, what matters in practice is how likely the probability to sample $\mathcal{N} < -1$ is.

In conclusion, we have illustrated that convergence but also divergence can happen for the multiplicative noise model. Those results are due to the probability to sample $1 + \mathcal{N}$ smaller than 0 and are therefore intrinsic to the noise model and not to the '+' strategy. The probability that $1+\mathcal{N}$ can be very small, in which case theory predicts divergence that will not be observed in simulations. We decided to present simple proofs relying on Borel-Cantelli Lemma. As a consequence, those proofs do not show the log-linear convergence and divergence observed in Section 2. Obtaining the log-linear behavior can be achieved using the theory of Markov chain on continuous state space. Last, we did not include results concerning a translated sphere $f(x) = \|x\|^2 + \alpha$ with $\alpha \geq 0$ for which our proofs of convergence can be extended but where linear convergence does not hold anymore due to the fact that the variance of the noise distribution does not reduce to zero close to the optimum.

[9] The indicator function $1_{[-a,a]}(x)$ equals 1 if $x \in [-a, a]$ and 0 otherwise.

[10] For computing the lower bound we use the fact that $P(\mathcal{N}(0, 1) < x) \leq \exp(-x^2/2)/|x|\sqrt{(2\pi)}$ for $x < 0$.

Acknowledgments

The authors would like to thank Nikolaus Hansen for many valuable discussions. This work receives partial supports from the ANR/RNTL project Optimisation Multidisciplinaire (OMD).

References

1. D. V. Arnold. *Noisy Optimization with Evolution Strategies*. GENA. Kluwer Academic Publishers, 2002.
2. D. V. Arnold and H.-G. Beye. Efficiency and mutation strength adaptation of the $(\mu/\mu_{\mathrm{i}}, \lambda)$-es in a noisy environment. In M. Schoenauer and al, editors, *Proceedings of Parallel Problem Solving from Nature - PPSN VI*, volume 1917 of *LNCS*, pages 39–48. Springer, 2000.
3. D. V. Arnold and H.-G. Beyer. Investigation of the (μ, λ)-ES in the presence of noise. In *Proceedings of 2001 IEEE Congress on Evolutionary Computation*, pages 332–339. IEEE Press, 2001.
4. D. V. Arnold and H.-G. Beyer. Local performance of the (1+1)-ES in a noisy environment. *IEEE Transactions on Evolutionary Computation*, 6(1):30–41, 2002.
5. D. V. Arnold and H.-G. Beyer. A comparison of evolution strategies with other direct search methods in the presence of noise. *Computational Optimization and Applications*, 24:135–159, 2003.
6. A. Auger and N. Hansen. Reconsidering the progress rate theory for evolution strategies in finite dimensions. In A. Press, editor, *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2006)*, pages 445–452, 2006.
7. 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.
8. H.-G. Beyer. *The Theory of Evolution Strategies*. Natural Computing Series. Springer-Verlag, 2001.
9. M. Jebalia, A. Auger, and P. Liardet. Log-linear convergence and optimal bounds for the $(1+1)$-ES. In N. Monmarché and al., editors, *Proceedings of Evolution Artificielle (EA'07)*, volume 4926 of *LNCS*, pages 207–218. Springer, 2008.
10. Y. Jin and J. Branke. Evolutionary Optimization in Uncertain Environments-A Survey. *IEEE Transactions on Evolutionary Computation*, 9(3):303–317, June 2005.
11. I. Rechenberg. *Evolutionsstrategie*. Friedrich Frommann Verlag (Günther Holzboog KG), Stuttgart, 1973.
12. P. Suganthan, N. Hansen, J. Liang, K. Deb, Y. P. Chen, A. Auger, and S. Tiwari. Problem definitions and evaluation criteria for the CEC 2005 special session on real-parameter optimization. Technical report, Nanyang Technological University, Singapore and KanGAL Report Number 2005005 (Kanpur Genetic Algorithms Laboratory, IIT Kanpur), May 2005.

Appendix

Proof of Proposition 1 The sequence $(\mathcal{F}_s(\mathrm{X}_n))$ is decreasing and is lower bounded by 0 as $\mathcal{F}_s(\mathrm{X}_n) \geq \|\mathrm{X}_n\|^2 (1 + m_{\mathcal{N}}) \geq 0$. Therefore it converges to a limit $l \geq 0$. Let us show that $l = 0$. Let $\epsilon > 0$, we have to show that $\exists\, n_0 \geq 0$ such that $\mathcal{F}_s(\mathrm{X}_n) \leq \epsilon$ for $n \geq n_0$. Since the sequence $(\mathcal{F}_s(\mathrm{X}_n))$ is decreasing, we only have to show that $\exists\, n_0 \geq 0$ such that $\mathcal{F}_s(X_{n_0}) \leq \epsilon$. Let $\beta > 1$ and such that $[1+m_{\mathcal{N}}, \beta(1+m_{\mathcal{N}})[\subset \mathrm{supp}(1+\mathcal{N})$.

In Lemma 2, we have defined the event $A_{n,\epsilon,\beta}$, shown that it is included in the event $\{\mathcal{F}_s(\mathrm{X}_{n+1}) \leq \epsilon\}$ and proved that the events $(A_{n,\epsilon,\beta})_n$ are independent. Moreover, $P(A_{n,\epsilon,\beta}) = P(\|\mathrm{e}_1+\sigma\mathrm{N}\|^2 \leq \frac{\epsilon}{(1+\beta)\theta^2(1+m_{\mathcal{N}})})P(1+\mathcal{N} \leq \beta(1+m_{\mathcal{N}}))$ (where θ is defined in Lemma 2) is a strictly positive constant for all n. Then $\sum_{n=0}^{+\infty} P(A_n) = +\infty$. This gives by BCL that $P(\overline{\lim}\ An) = 1$. Therefore $P(\overline{\lim}\ \{\mathcal{F}_s(\mathrm{X}_{n+1}) \leq \epsilon\}) = 1$, i.e. $\exists n_0$ such that $\forall n \geq n_0,\ \mathcal{F}_s(\mathrm{X}_n) \leq \epsilon$. Therefore $\mathcal{F}_s(\mathrm{X}_n)$ converges to 0. The sequence $(\|\mathrm{X}_n\|)$ converges also to 0 as $\|\mathrm{X}_n\|^2 \leq \frac{\mathcal{F}_s(\mathrm{X}_n)}{1+m_{\mathcal{N}}}$. □

Lemma 2. *If $m_{\mathcal{N}} + 1 > 0$, the following points hold:*

1. *The sequence $(\|\mathrm{X}_n\|)$ is upper bounded by $\theta := \sqrt{\frac{\mathcal{F}_s(\mathrm{X}_0)}{1+m_{\mathcal{N}}}} > 0$.*
2. *Let $\epsilon > 0$ and $\beta > 1$ such that $\beta(1 + m_{\mathcal{N}}) \in supp(1 + \mathcal{N})$. For $n \geq 0$, the event $A_{n,\epsilon,\beta} := \left(\left\{\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathrm{N}_n\right\|^2 \leq \frac{\epsilon}{(1+\beta)\theta^2(1+m_{\mathcal{N}})}\right\} \cap \{1 + \mathcal{N}_n \leq \beta(1 + m_{\mathcal{N}})\}\right)$ [11] verifies $A_{n,\epsilon,\beta} \subset \{\mathcal{F}_s(\mathrm{X}_{n+1}) \leq \epsilon\}$. Moreover, the events $(A_{n,\epsilon,\beta})_n$ are independent.*

Proof. 1. For $n \geq 0$, $\mathcal{F}_s(\mathrm{X}_n) = \|\mathrm{X}_n\|^2 (1 + O_n) = \|\mathrm{X}_n\|^2 \left(1 + \mathcal{N}_{\phi(n)}\right)$ where $\phi(n)$ is the index of the last acceptance (obviously $\phi(n) \leq n$). Then, for $n \geq 0$
$\mathcal{F}_s(\mathrm{X}_n) \geq \|\mathrm{X}_n\|^2 (1 + m_{\mathcal{N}}) \geq 0$ and consequently $\|\mathrm{X}_n\|^2 \leq \frac{\mathcal{F}_s(\mathrm{X}_n)}{1+m_{\mathcal{N}}} \leq \frac{\mathcal{F}_s(\mathrm{X}_0)}{1+m_{\mathcal{N}}}$.
2. Let $\epsilon > 0$ and $\beta > 1$ such that $[1 + m_{\mathcal{N}}, \beta(1 + m_{\mathcal{N}})[\subset \text{ supp}(1 + \mathcal{N})$ (with $\beta m_{\mathcal{N}} < M_{\mathcal{N}}$ if $M_{\mathcal{N}} < +\infty$). For $n \geq 0$, the event
$\left\{\left(\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathrm{N}_n\right\|^2 < \frac{\epsilon}{(1+\beta)\theta^2(1+m_{\mathcal{N}})}\right) \cap (1 + \mathcal{N}_n < \beta(1 + m_{\mathcal{N}}))\right\}$ implies for the offspring $\tilde{\mathrm{X}}_n := \mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}_n$ created at the iteration n that
$\mathcal{F}_s(\tilde{\mathrm{X}}_n) = \|\mathrm{X}_n\|^2\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathrm{N}_n\right\|^2(1 + \mathcal{N}_n) \leq \theta^2 \frac{\epsilon}{(1+\beta)(1+m_{\mathcal{N}})\theta^2}\beta(1 + m_{\mathcal{N}})$.
Then $\mathcal{F}_s(\tilde{\mathrm{X}}_n) \leq \frac{\beta}{\beta+1}\epsilon < \epsilon$. If this offspring is accepted then $\mathcal{F}_s(\mathrm{X}_{n+1}) < \epsilon$, otherwise the fitness is already less than ϵ and we have also $\mathcal{F}_s(\mathrm{X}_{n+1}) < \epsilon$. Finally, the independency of the events $(A_{n,\epsilon,\beta})_n$ result from Lemma 3 applied to the sequence (X_n). □

Lemma 3. *Let (U_n) be a sequence of random vectors in $\mathbb{R}^d$ such that $P(\|U_n\| = 0) = 0$ and N_n independent random vectors distributed as $\mathrm{N}(0, I_d)$. Then the variables $Y_n := \left\|\frac{U_n}{\|U_n\|} + \sigma\mathrm{N}_n\right\|$ are independent.*

Proof. The independance of the random variables Y_n is due to the fact that the multivariate Gaussian variable $\mathrm{N}(0, I_d)$ is isotropic and is therefore invariant by rotation. The length of the vector $\frac{U_n}{\|U_n\|} + \sigma\mathrm{N}_n$ will therefore be independent of where we start on the unit hypersphere, i.e., independent of the vector $\frac{U_n}{\|U_n\|}$. □

Proof of Proposition 2 Let $n \geq n_1$ (n_1 defined in Lemma 4). We have to show that for any $m < \mathcal{F}_s(X_{n_1}) < 0$, $\exists\ n \geq n_1$ such that $\mathcal{F}_s(\mathrm{X}_n) \leq m$, or equivalently $|\mathcal{F}_s(\mathrm{X}_n)| \geq$

[11] The multivariate Gaussian distribution is absolutely continuous with respect to the Lebesgue measure such that $P(\|\mathrm{X}_n\| = 0) = 0$ and then we can divide by $\|\mathrm{X}_n\|$ almost surely.

$|m|$. Similarly to the proof of Proposition 1, by BCL we have ($B_{n,m,\beta}$ i.o.) (($B_{n,m,\beta}$ being defined in Lemma 4) therefore Lemma 4 gives that ($\mathcal{F}_s(\mathrm{X}_{n+1}) \leq m$ i.o.). Then $\mathcal{F}_s(\mathrm{X}_n) = \|\mathrm{X}_n\|^2 (1 + O_n)$ tends to $-\infty$. For all $n \geq n_1$, $0 \geq 1 + O_n \geq 1 + m_{\mathcal{N}}$, then $\frac{|\mathcal{F}_s(\mathrm{X}_n)|}{|1+m_{\mathcal{N}}|} \leq \|\mathrm{X}_n\|^2$ for $n \geq n_1$. Consequently ($\|\mathrm{X}_n\|$) converges to $+\infty$ almost surely. □

Lemma 4. *Assume that $m_{\mathcal{N}} + 1 < 0$. The following points hold:*

1. *There exists $n_1 \geq 0$ and $A := \sqrt{\frac{|\mathcal{F}_s(X_{n_1})|}{|1+m_{\mathcal{N}}|}} > 0$ such that $\mathcal{F}_s(\mathrm{X}_n) < 0$ and $\|\mathrm{X}_n\| \geq A$ for $n \geq n_1$ almost surely.*
2. *Let $m < \mathcal{F}_s(X_{n_1}) < 0$ and $\beta > 1$. For $n \geq n_1$, the event $B_{n,m,\beta}$ defined by $B_{n,m,\beta} := \left(\left\{|1 - \sigma\|\mathrm{N}_n\||^2 \geq \frac{|m|}{|m_{\mathcal{N}}+1|}\frac{\beta+1}{A^2}\right\} \cap \left\{1 + \mathcal{N}_n \leq \frac{1+m_{\mathcal{N}}}{\beta}\right\}\right)$ verifies $B_{n,\epsilon,\beta} \subset (\mathcal{F}_s(\mathrm{X}_{n+1}) \leq m)$.*

Proof. 1. We first prove that the event $\mathcal{A} := \{\, \exists\, n_1 \geq 0$ such that $\forall\, n \geq n_1$, $\mathcal{F}_s(\mathrm{X}_n) < 0\}$ is equivalent to the event $\mathcal{B} := \{\, \exists\, p_0 \geq 0$ such that $\mathcal{N}_{p_0} < -1\,\}$. Proving that $\mathcal{A} \subset \mathcal{B}$ is equivalent to show that $\mathcal{B}^c \subset \mathcal{A}^c$. Suppose that $\forall p \geq 0,\ \mathcal{N}_p \geq -1$. Then $\forall p \geq 0,\ O_p \geq -1$. Therefore $\forall p \geq 0,\ \mathcal{F}_s(X_p) = \|X_p\|^2 (1 + O_p) \geq 0$. Now we have to show that $\mathcal{B} \subset \mathcal{A}$: Suppose that $\exists\ p_0 \geq 0$ such that $\mathcal{N}_{p_0} < -1$. We denote $p_1 \geq 0$ the integer defined by $p_1 = \min\{p \in \mathbb{N}$ such that $\mathcal{N}_p < -1\}$. Then $\mathcal{F}_s\,(X_{p_1}) < 0$ and $\mathcal{F}_s\,(X_p) \geq 0$ for all $0 \leq p \leq p_1 - 1$. Since $(\mathcal{F}_s\,(\mathrm{X}_n))$ is a decreasing sequence, $\mathcal{F}_s(\mathrm{X}_n) < 0\ \forall\ n \geq p_1$. This implies that $P(\mathcal{A}) = P(\mathcal{B})$. Now, we have for all $n \geq 0$, $P(\mathcal{B}^c) = P(\cap_{p=0}^{+\infty}(\mathcal{N}_p \geq -1)) \leq \Pi_{p=0}^{n} P\,(\mathcal{N}_p \geq -1) = (P\,(\mathcal{N} \geq -1))^n$. Let $a := P\,(\mathcal{N} \geq -1)$[12]. As $m_{\mathcal{N}} < -1$, then $a < 1$ which gives $P(\mathcal{B}^c) = 0$ and therefore $P(\mathcal{A}) = 1$. Then $\exists\ n_1 \geq 0$ such that $\mathcal{F}_s(\mathrm{X}_n) < 0$ for $n \geq n_1$ almost surely. The sequence $(\mathcal{F}_s(\mathrm{X}_n))_n$ is decreasing (because of the elitist selection). Then for $n \geq n_1$, $\mathcal{F}_s(\mathrm{X}_n) \leq \mathcal{F}_s(X_{n_1}) < 0$. This gives $|\mathcal{F}_s(\mathrm{X}_n)| \geq |\mathcal{F}_s(X_{n_1})| > 0$. It is easy to see (from Eq. 4) that for all $n \in \mathbb{N}$, $O_n = \mathcal{N}_{\psi(n)}$ where $\psi(n)$ is the last acceptance index before the iteration n. Combining this with the fact if $1 + m_{\mathcal{N}} \leq 1 + \mathcal{N}_{\psi(n)} < 0$ one gets $0 < |\mathcal{F}_s(X_{n_1})| \leq |\mathcal{F}_s(\mathrm{X}_n)| = \|\mathrm{X}_n\|^2 |1 + \mathcal{N}_{\psi(n)}| \leq \|\mathrm{X}_n\|^2 |1 + m_{\mathcal{N}}|$. Then $\|\mathrm{X}_n\|^2 \geq \frac{|\mathcal{F}_s(X_{n_1})|}{|1+m_{\mathcal{N}}|} > 0$.
2. By the first result of the Lemma, $\exists\, n_1 \geq 0,\ A > 0$ such that $\mathcal{F}_s(\mathrm{X}_n) < 0$ and $\|\mathrm{X}_n\| \geq A\ \forall n \geq n_1$. We consider $n \geq n_1$, then $\|\mathrm{X}_n\| > A$. We notice that $\forall\ y \in \mathbb{R}^d \backslash \{(0,0)\}$, $\left\| \frac{y}{\|y\|} + \sigma \mathrm{N} \right\| \geq |1 - \sigma\|\mathrm{N}\||$. Let $\beta > 1$. As the upper bound $M_{\mathcal{N}}$ verifies $1 + M_{\mathcal{N}} > 0$, $\frac{1+m_{\mathcal{N}}}{\beta} \in \text{supp}(1 + \mathcal{N}) \cap \mathbb{R}^-$. Suppose that we have $|1 - \sigma\|\mathrm{N}_n\||^2 \geq \frac{(\beta+1)|m|}{A^2|1+m_{\mathcal{N}}|}$ and $|1 + \mathcal{N}_n| \geq \frac{|1+m_{\mathcal{N}}|}{\beta}$, then the offspring $\tilde{\mathrm{X}}_n := \mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathrm{N}_n$ is such that $|\mathcal{F}_s(\tilde{\mathrm{X}}_n)| = \|\mathrm{X}_n\|^2 \left\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathrm{N}_n \right\|^2 |1 + \mathcal{N}_n| \geq \|\mathrm{X}_n\|^2 |1 - \sigma\|\mathrm{N}_n\||^2 |1 + \mathcal{N}_n|$. Then $|\mathcal{F}_s(\tilde{\mathrm{X}}_n)| \geq \frac{\beta+1}{\beta}|m| > |m|$ which gives $\mathcal{F}_s(\mathrm{X}_{n+1}) \leq \mathcal{F}_s(\tilde{\mathrm{X}}_n) \leq m$. Consequently, for $n \geq n_0$, the event $B_{n,m,\beta} := \left\{|1 - \sigma\|\mathrm{N}_n\|^2| \geq \frac{(\beta+1)|m|}{A^2|1+m_{\mathcal{N}}|}\right\} \cap \left\{|1 + \mathcal{N}_n| \geq \frac{|1+m_{\mathcal{N}}|}{\beta}\right\}$ is included in $\{\mathcal{F}_s(\mathrm{X}_{n+1}) \leq m\}$. □

[12] We apply the same reasoning with $a = 2/3$ for the example given in the introduction where $\mathcal{N}$ take values in $\{-\gamma, 0, \gamma\}$ (with $\gamma > 1$).