Log-linear convergence and divergence of the scale-invariant (1+1)-ES in noisy environments

Mohamed Jebalia, Anne Auger, Nikolaus Hansen
TAO Team - INRIA Saclay-Île-de-France,
LRI - Paris-Sud University
91405 Orsay Cedex, France
anne.auger@inria.fr
Fax: +33 1.69.15.42.40, Tel: +33 1.69.15.63.97

March 8, 2010

Abstract

Noise is present in many real-world continuous optimization problems. Stochastic search algorithms such as Evolution Strategies (ESs) have been proposed as effective search methods in such contexts. In this paper, we provide a mathematical analysis of the convergence of a (1+1)-ES on unimodal spherical objective functions in the presence of noise. We prove for a multiplicative noise model that for a positive expected value of the noisy objective function, convergence or divergence happens depending on the infimum of the support of the noise. Moreover, we investigate convergence rates and show that log-linear convergence is preserved in presence of noise. This result is a strong theoretical foundation of the robustness of ESs with respect to noise.

Keywords: Numerical optimization, Noisy optimization, Stochastic optimization algorithms, Evolution Strategies, Convergence, Convergence rates, Markov chains, Borel-Cantelli Lemma.

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 [1, 2, 3]. 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 fitness values. In this paper, we will investigate a certain class of noisy problems, which use the so-called *multiplicative noise*, where the noiseless objective function $f(x)$ is perturbed by the multiplication with a noise term independent of x and thus of $f(x)$. The plain multiplicative-noisy objective function $\mathcal{F}$ reads

$$\mathcal{F}(x) = f(x)\xi \ . \tag{1}$$

The noise random variable, ξ, is sampled independently at each new evaluation of a solution point. The multiplicative noise model is well established [4]. For the remainder we will assume $f \geq 0$ and minimization. We will conduct our subsequent analysis on an objective function $g \circ \mathcal{F}$, where g is a strictly increasing mapping which can take on any values in $\mathbb{R}$ (in particular also $g(0)$ can take on any value in $\mathbb{R}$). The noise model makes the assumption that the dispersion of the noise (e.g. the variance, given it exists) decreases when the optimum is approached, just as the gradient diminishes in many optimization problems when approaching the optimum. The assumption is realistic in particular as f might be stricly larger than zero, in which case the noise does not entirely vanish at the optimum (but the subsequently analyzed model is more restrictive as the minimal value of f is zero). Consequently, such noise models are also used to benchmark robustness of EAs with respect to noise [5, 6].

A typical goal in noisy optimization is to converge to the minimum *of the averaged value* of the observed random variable. If the expected value of the noise in Eq. 1 is positive, this means minimizing f. Using a different statistic than the expected value as minimization goal might lead to different results (see below).

The focus in this paper is continuous optimization (here minimization) where f maps a continuous search space, i.e., an open subset of $\mathbb{R}^d$ into $\mathbb{R}$. Evolution Strategies (ESs) are a class of EAs specifically designed for continuous optimization. In ESs a set of candidate solutions evolves by adding Gaussian perturbations (mutations) to the current, optionally recombined

solutions and selecting some of the new solutions. In state-of-the-art ESs, the mutation operator is adaptive and the parameters defining the underlying probability distribution are evolving. In the case of a Gaussian mutation operator those parameters are the step length (or step-size) giving the general scale of the search, and a covariance matrix whose eigenvectors correspond to the principal directions of the underlying ellipsoidal probability distribution. Adaptive step-size algorithms are described in [7, 8, 9], and a state-of-the art method for adapting the covariance matrix is presented in [10].

The fundamental characteristic of step-size adaptive ESs is that they can constantly increase the precision, i.e., they can approach an optimum with arbitrary precision and, ideally, the time needed to gain one order of magnitude is independent of the initial precision, i.e., the logarithm of the distance to the optimum decreases linearly. Formally, log-linear behavior (convergence or divergence) means that there exists a non-zero constant value c such that the distance to the optimum, d_n, at an iteration n satisfies

$$\lim_n \frac{1}{n} \ln(d_n) = c\,. \tag{2}$$

Convergence (resp. divergence) takes place if $c < 0$ (resp. $c > 0$). Log-linear convergence has been theoretically investigated in the case of non-noisy objective functions. The results are twofold. First, for a fixed dimension d, convergence of ESs cannot be faster than log-linear [11, 12]. Moreover, optimal convergence rates are reached on spherical functions $f(x) = g(\|x\|)$, where $g : [0, \infty[\mapsto \mathbb{R}$ is a strictly increasing function, $x \in \mathbb{R}^d$ and $\|.\|$ denotes the euclidean norm on $\mathbb{R}^d$, for an (artificial) adaptive step-size algorithm where the step-size is set at each iteration proportionally to the distance to optimum [11, 12]. An algorithm using this optimal adaptation rule is termed scale-invariant algorithm[1]. Second, log-linear convergence of a realistic adaptation scheme has been proven for self-adaptive ESs on the sphere function [13, 14]. Log-linear convergence can also be seen from another perspective and be formalized differently as "the expected time needed to halve the distance to the optimum is proportional to n". Log-linear convergence formulated in this manner has been proven for the (1+1)-ES using a 1/5-success rule on spherical functions and certain ellipsoidal functions [15, 16].

[1]Scale invariant algorithms have a central place in the theory of evolution strategies. They have been widely investigated in the case of progress rate theory that examines the one-step expected progress towards the optimum in the limit of the dimension to infinity.

ESs in noisy environments have been studied by Arnold and Beyer for multiplicative noise that writes

$$\mathcal{F}(x) = \|x\|^2 \left(1 + \frac{2\sigma_\epsilon^*}{d}\mathcal{N}(0,1)\right) , \tag{3}$$

where σ_ϵ^* characterizes the "normalized" standard deviation of the noise, d is the search space dimension and $\mathcal{N}(0,1)$ is the noise random variable which follows a Gaussian distribution with zero mean and standard deviation one [4, 17, 18, 19, 20]. Under the assumption that d goes to infinity, Arnold and Beyer derive a positive expected quality gain for the elitist (1+1)-ES [20]. This implies a decrease of the expectation of the square distance to the optimum.

However, given Eq. 1, where $f \geq 0$ and the noise random variable takes values smaller than zero, we find a simple counterexample to the convergence of the (1+1)-ES to the optimum of the noiseless part of the objective function. Let $f(x) = \|x\|^2$ and ξ take three distinct values: $1+a$, 1 and $1-a$ each with probability 1/3, where a satisfies $a > 1$. For a given $x \in \mathbb{R}^d$, the objective function $\mathcal{F}(x)$ takes 3 different values $(1+a)\|x\|^2$, $\|x\|^2$, $(1-a)\|x\|^2$ (each with probability 1/3). The last term is strictly negative if x is not zero. Therefore, once a negative objective function value is reached, the (1+1)-ES will never accept solutions closer to the optimum since they will always have a higher objective function value. Only solutions with noise values of $1-a$ and a larger value of $\|x\|^2$ have a lower objective function value $\mathcal{F}(x)$ and can therefore be accepted. These solutions are likely to occur and consequently, the (1+1)-ES will diverge log-linearly, i.e., the logarithm of the distance to the optimum will increase linearly. And yet, the expectation of $\mathcal{F}$ equals the noiseless part of the function $\|x\|^2$.

Starting from the observation that even if the expectation of the noisy function equals $\|x\|^2$ (where convergence is expected), divergence can be observed, the question arises whether and when this example generalizes to different settings. The objective of this paper is to address the question of how the properties of the noise distribution relate to convergence or divergence of the (1+1)-ES. The results will give particular insights into the multiplicative noise model. The results presented here are an extension of the results presented in [21].

The second question addressed, in case of convergence, is the actual convergence rate on a noisy objective function. Is the log-linear convergence rate of ESs preserved in the noisy case? A positive answer would be a strong

support for the robustness of ESs in noisy environments. Those results are not present in [21] and rely on the theory of Markov chains for continuous state space.

The algorithm investigated in this paper is the scale-invariant (1+1)-ES which is, in the case of the non-noisy spherical function, the ES allowing to have the fastest convergence rate per fitness evaluation [19] (excluding weighted recombination [22, 10, 23]). This paper is organized as follows: in Sections 2.1 and 2.2, we present the mathematical formulation of the objective function model and of the scale-invariant (1+1)-ES respectively. In Section 3, we put our counterexample informally into a broader context and present empirical results. In Section 4, we analyze the convergence of the scale invariant (1+1)-ES. In Section 5, we find that the behavior of the scale-invariant (1+1)-ES is log-linear when minimizing these noisy objective functions. In Section 6 the results are discussed. Proofs are presented in the appendix.

Preliminary notations

In this paper $\mathbb{Z}^+$ denotes the set of non-negative integers $\{0, 1, 2, \ldots\}$, $\mathbb{Z}_{-1}^+ = \mathbb{Z}^+ \cup \{-1\}$ and $\mathbb{N}$ denotes the set of positive integers $\{1, 2, \ldots\}$. The unit vector $(1, 0, \ldots, 0) \in \mathbb{R}^d$ is denoted as e_1. For a set A, $x \mapsto 1_A(x)$ denotes the indicator function that is equal to one if $x \in A$ and zero otherwise. $(\Omega, \mathcal{A}, P)$ is a probability space: Ω is a set, $\mathcal{A}$ a σ-algebra defined on this set and P a probability measure defined on $(\Omega, \mathcal{A})$. For $p \in \mathbb{N}$, $\mathbb{R}^p$ is equipped with the Borel σ-algebra denoted $\mathfrak{B}(\mathbb{R}^p)$. For a subset $S \subset \mathbb{R}^p$, $\mathfrak{B}(S)$ will denote the Borel σ-algebra on S. If X is a random variable defined on $(\Omega, \mathcal{A}, P)$, i.e. a measurable function from Ω to $\mathbb{R}$, then, for $B \subset \mathbb{R}$, $B \in \mathfrak{B}(\mathbb{R})$, the indicator function $1_{\{X \in B\}}$ maps Ω to $\{0, 1\}$ and equals one if and only if $X(\omega) \in B$ for $\omega \in \Omega$: $\omega \in \Omega \mapsto 1_{\{\omega : X(\omega) \in B\}}(\omega)$. $\mathcal{N}(a, b^2)$ denotes a normal distribution with mean a and variance b^2. $\mathcal{N}(0, I_d)$ is the multivariate normal distribution with mean $(0, \ldots, 0) \in \mathbb{R}^d$ and covariance matrix identity I_d.

2 Mathematical model for the fitness function and for the (1+1)-ES

2.1 Fitness function model

The objective function investigated in [21] represents a sphere function perturbed by a multiplicative lower bounded noise. The noisy objective function model investigated here generalizes this model in two respects. First, the noiseless part of the function, originally the sphere function $f(x) = \|x\|^2$, $x \in \mathbb{R}^d$, $d \in \mathbb{N}$, is replaced by a more general function $f(x) = \|x\|^\alpha$, where $\alpha > 0$. Second, the function f is composed with a strictly increasing function $g : \mathbb{R} \to \mathbb{R}$, for example $g(f) = \tanh(f) + f/10 - 20$. The investigated noisy function model reads

$$\mathcal{F}(x) = g\left(\|x\|^\alpha \xi\right) , \tag{4}$$

where ξ is a random variable modeling the noise, sampled independently at each new evaluation of a point. We assume that the law of ξ, denoted $\mathcal{L}_\xi$ has a probability density function denoted p_ξ. We also assume that the support of p_ξ is the range $]m_\xi, M_\xi[$ where $-\infty \leq m_\xi < M_\xi \leq +\infty$, and $m_\xi \neq 0$.

Remark 1 *In the case where g equals the identity and $\alpha = 2$, the model defined in Eq. 4 is the one used in [21] (with a change of notation for ξ). However, in [21] we made an additional assumption on the expectation of the noise stating that $0 < E(\xi) < +\infty$. This assumption guarantees a finite expectation of the noisy objective function and an agreement of the argmin[2] of f with the argmin of the expected value of the noisy objective function.*

2.2 The (1+1)-ES minimizing Eq. 4

Let $(\mathcal{N}_n)_{n\in\mathbb{Z}^+}$, be a sequence of random vectors defined on $(\Omega, \mathcal{A}, P)$, independent and identically distributed (i.i.d.) with common law the isotropic multivariate normal distribution on $\mathbb{R}^d$ denoted by $\mathcal{N}(0, I_d)$. The density function of $\mathcal{N}(0, I_d)$ is a d-dimensional function denoted $p_\mathcal{N}$. Let $(\xi_n)_{n\in\mathbb{Z}^+_{-1}}$ be a sequence of random variables defined on $(\Omega, \mathcal{A}, P)$ i.i.d. with common law $\mathcal{L}_\xi$ introduced in the previous section. We also assume that the sequences $(\mathcal{N}_n)_{n\in\mathbb{Z}^+}$ and $(\xi_n)_{n\in\mathbb{Z}^+_{-1}}$ are independent implying in particular that for each $n \in \mathbb{Z}^+$, $\mathcal{N}_n$ and ξ_n are independent. In the sequel we might omit to indicate

[2]The argmin of a function $x \mapsto h(x)$ is defined as $\{y \,|\, h(y) = \min_x h(x)\}$.

the definition domain of the indexes n that should be clear within the context and denote for instance $(\mathcal{N}_n)_n$ instead of $(\mathcal{N}_n)_{n\in\mathbb{Z}^+}$.

The (1+1)-ES is a simple algorithm evolving a unique solution that is a vector of $\mathbb{R}^d$ which is also called parent. The sequence of solutions (or parents) generated is a sequence of random vectors denoted $(\mathrm{X}_n)_n$ defined on $(\Omega, \mathcal{A}, P)$. The sequence of noise associated with the sequence of parents is denoted $(O_n)_n$. Both $(\mathrm{X}_n)_n$ and $(O_n)_n$ obey a recurrence relation that we describe first step-by-step.

Let $\mathrm{X}_0 \in \mathbb{R}^d$ be the first sampled parent. We assume that $\|\mathrm{X}_0\| > 0$ almost surely and that X_0, $(\mathcal{N}_n)_n$ and $(\xi_n)_n$ are independent. The objective function value associated with X_0 equals $g(\|\mathrm{X}_0\|^\alpha \xi_{-1})$ and the selected noise $O_0 = \xi_{-1}$. For this initialization step, the law of the random variable O_0 equals $\mathcal{L}_\xi$, however as it will become clear later, for the other iterations ($n \geq 1$), O_n has different laws.

We assume now that $(\mathrm{X}_k)_{0\leq k\leq n-1}$ and $(O_k)_{0\leq k\leq n-1}$ are given and we describe how the next iterates X_n and O_n are generated. First, the vector X_n is perturbed by the addition of $\mathcal{N}_n$ scaled by a strictly positive real number called step-size σ_n to create a new candidate solution called offspring that writes $\mathrm{X}_n + \sigma_n \mathcal{N}_n$. Because $\mathcal{N}(0, I_d)$ is a spherical distribution, the algorithm is called isotropic ES.

The efficiency of an isotropic ES is closely related to the adaptation scheme of the step-sizes mutation sequence $(\sigma_n)_n$. On an isotropic unimodal function the optimal adaptation scheme of the sequence $(\sigma_n)_n$ of an isotropic ES is given, according to [11, 12], by the (artificial) scale invariant adaptation rule in which the step-size is set proportionally to the distance to the optimum, i.e., $\sigma_n = \sigma\|\mathrm{X}_n\|$ (we assume here that the optimum is in $(0, \ldots, 0) \in \mathbb{R}^d$) where σ is a strictly positive constant. The simplicity of this step-size update rule renders the analysis easier to carry out than for real step-size adaptation schemes [14]. Therefore the scale-invariant (1+1)-ES is usually a good choice for the first theoretical investigations and is the algorithm we will investigate in the sequel.

The fitness function value associated with the offspring $\mathrm{X}_n + \sigma_n \mathcal{N}_n$ equals $g(\|\mathrm{X}_n + \sigma_n \mathcal{N}_n\|^\alpha \xi_n)$. This offspring will become the new parent if and only its fitness value is smaller than the one of its parent X_n.

Adding up the different steps, we can write the recurrence relations

obeyed by $(\mathrm{X}_n)_n$ and $(O_n)_n$. First X_{n+1} satisfies

$$\mathrm{X}_{n+1} = \begin{cases} \mathrm{X}_n + \sigma\|\mathrm{X}_n\|\,\mathcal{N}_n & \text{if } g\left(\left\|\mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathcal{N}_n\right\|^{\alpha}\xi_n\right) < g\left(\|\mathrm{X}_n\|^{\alpha}O_n\right) \\ \mathrm{X}_n & \text{otherwise}\,, \end{cases} \tag{5}$$

and the accepted noise O_{n+1} of the new parent X_{n+1} obeys:

$$O_{n+1} = \begin{cases} \xi_n & \text{if } g\left(\left\|\mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathcal{N}_n\right\|^{\alpha}\xi_n\right) < g\left(\|\mathrm{X}_n\|^{\alpha}O_n\right) \\ O_n & \text{otherwise}\ . \end{cases} \tag{6}$$

Since g preserves the ordering, it can be dropped in the acceptance condition in Eqs. 5 and 6 and we can write equivalently

$$\mathrm{X}_{n+1} = \begin{cases} \mathrm{X}_n + \sigma\|\mathrm{X}_n\|\,\mathcal{N}_n & \text{if } \left\|\mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathcal{N}_n\right\|^{\alpha}\xi_n < \|\mathrm{X}_n\|^{\alpha}O_n \\ \mathrm{X}_n & \text{otherwise}\,, \end{cases} \tag{7}$$

and for the accepted noise

$$O_{n+1} = \begin{cases} \xi_n & \text{if } \left\|\mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathcal{N}_n\right\|^{\alpha}\xi_n < \|\mathrm{X}_n\|^{\alpha}O_n \\ O_n & \text{otherwise}\ . \end{cases} \tag{8}$$

3 Motivation: convergence or divergence?

In this section, the counterexample informally described in the introduction is put into a broader context and sustained with some empirical results.

3.1 Elementary remarks on the noise model

We consider the noise model defined in Eq. 1, where $f \geq 0$. In this case, if $E(\xi) > 0$, the argmin of the expected value of $\mathcal{F}(x)$ equals the argmin of $f(x)$. The support of the noise random variable ξ admits m_ξ as infimum. Therefore, because f is non-negative, $m_\xi f(x)$ is the infimum of the values that can be reached by the noisy fitness function for different instantiations of the random variable ξ for a given x.

Fig. 1 depicts a cut of $f(x) = \|x\|^2$ and $m_\xi f(x)$ for m_ξ equals 0.5 and -0.5. The sign of m_ξ determines whether $m_\xi f(x)$ is decreasing or increasing:

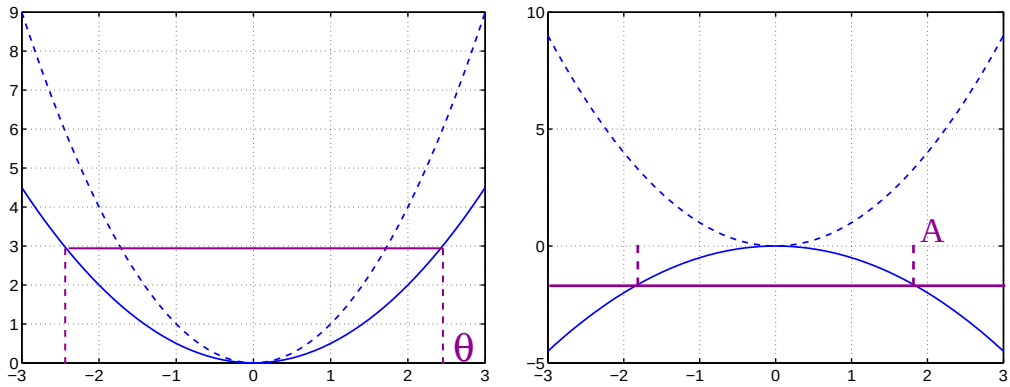

Figure 1: Dashed Lines: one dimensional cut of $f(x) = \|x\|^2$ along one arbitrary unit vector. Solid (bent) lines: greatest lower bound of $\mathcal{F}(x)$, i.e., $m_\xi f(x)$ for $m_\xi = 0.5$ (left) and $m_\xi = -0.5$ (right). For a given x, the noisy-objective function can, in particular, take any value between the dashed curve and the solid curve. Given a fitness according to the solid horizontal line, only values below this line are accepted as new parents. Thus, on the left graph, θ represents an upper bound for $\|\mathrm{X}_n\|$ and on the right graph, A represents a lower bound for $\|\mathrm{X}_n\|$, prohibiting convergence to $x = 0$.

for $m_\xi > 0$, $m_\xi f(x)$ is convex and converges to $+\infty$ for $\|x\| \to \infty$, and for $m_\xi < 0$, $m_\xi f(x)$ is concave and converges to $-\infty$ for $\|x\| \to \infty$. Minimizing $m_\xi f(x)$ in the case of $m_\xi < 0$ means that $\|x\|$ is diverging to $+\infty$ which is the opposite of the desired behavior when minimizing the non-noisy function $f(x) = \|x\|^2$. Referring to the example sketched in the introduction, $\|x\|^2$ and $(1-a)\|x\|^2$ for $a = 1.5$ are the curves represented in Fig. 1, right.

3.2 Experimental observations

We investigate now numerically how the infimum of the noise values might affect the convergence. For this purpose we use a (1+1)-ES and a (1,5)-ES both with scale-invariant adaptation scheme for the step-size.

We investigate the function $\mathcal{F}(x) = \|x\|^2\xi$ when the noise ξ is uniformly distributed in the ranges $[0.5, 1.5]$ and $[-0.5, 2.5]$ respectively denoted $U_{[0.5,1.5]}$ and $U_{[-0.5,2.5]}$. This latter noise corresponds to the concave lower bound $-0.5\|x\|^2$ plotted in Fig. 1. In Figure 2, the result of 10 independent runs of the (1+1)-ES (10 lower curves of each graph) in dimension $d = 10$ are plot-

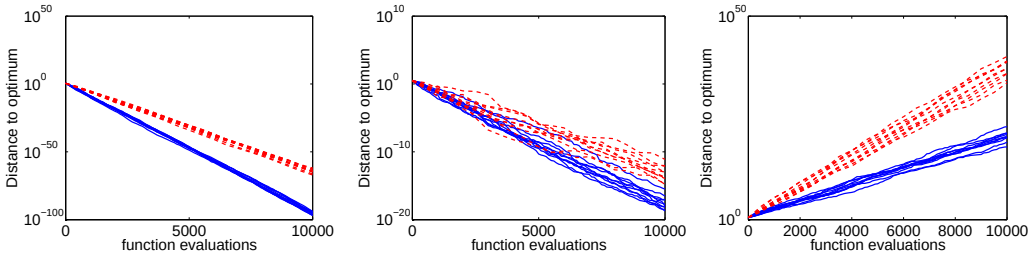

Figure 2: Distance to the optimum (in log-scale) versus number of evaluations. Ten independent runs for the scale-invariant (1+1)-ES (10 lower curves of each graphs) and (1,5)-ES (10 upper curves of each graph) with $d = 10$ and $\sigma = 1/d$. Left: $f(x) = \|x\|^2$. Middle: $f(x) = \|x\|^2 U_{[0.5,1.5]}$. Right: $f(x) = \|x\|^2 U_{[-0.5,2.5]}$.

ted for the non-noisy sphere (left), $\|x\|^2 U_{[0.5,1.5]}$ (middle) and $\|x\|^2 U_{[-0.5,2.5]}$ (right). In both noisy cases the expected value of the function equals $\|x\|^2$ since $E(U_{[0.5,1.5]}) = E(U_{[-0.5,2.5]}) = 1$, however, we observe a drastic difference between the two cases: the algorithm converges to the optimum for the noise $U_{[0.5,1.5]}$ whereas the distance to the optimum increases (log)-linearly for the noise with infimum -0.5. Comparing the left and middle figures, we can also conclude as expected, that the presence of noise slows down the convergence (here by a factor of about five). We conducted the same experiments for a (1,5)-ES (10 upper curves of each graph) and observe the same two behaviors, convergence on $\|x\|^2 U_{[0.5,1.5]}$ (middle) and divergence on $\|x\|^2 U_{[-0.5,2.5]}$ (right). However, contrary to what we will prove for the (1+1)-ES, we do not state that "$m_\xi = 0$" is a limit value between convergence and divergence in the case of a (1,λ)-ES. Indeed convergence and divergence depends on the intrinsic properties of the noise, on λ and σ as well [19].

Last, we investigate numerically the (1+1)-ES where ξ is normally distributed with expectation one and in particular unbounded. This corresponds to the case investigated in [20]. We display results for a standard deviation of the Gaussian noise of 0.1, 2 and 10 in Fig. 3. Within the given time horizon we observe convergence when the standard deviation of the noise equals 0.1 and divergence in the last two cases.

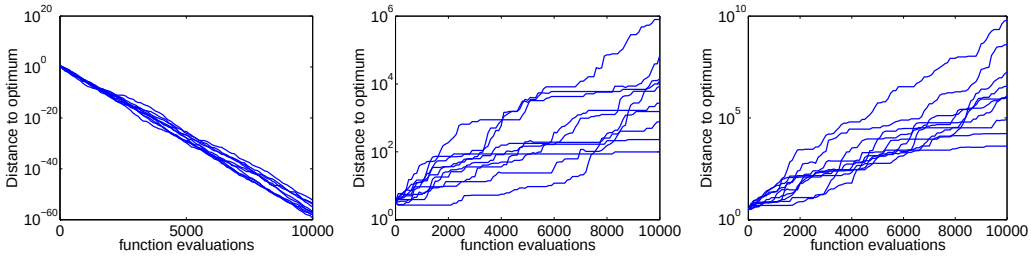

Figure 3: Ten independent runs for the scale-invariant (1+1)-ES with a normally distributed noise: on $f(x) = \|x\|^2 \mathcal{N}(1, \sigma_\epsilon^2)$ with σ_ϵ equals 0.1 (left), 2 (middle) and 10 (right) for $d = 10$ and $\sigma = 1/d$.

4 Theoretical analysis of convergence

In this section we provide a mathematical analysis of the results observed experimentally for the scale-invariant (1+1)-ES and for the fitness function model Eq. 4. We prove that convergence or divergence of the (1+1)-ES is *solely determined* by the infimum of the support of the noise m_ξ (given finite $m_\xi \neq 0$) and $m_\xi = 0$ is the limit case between convergence and divergence. One important consequence of this result is that for any positive expected value of the noisy function, for the same algorithm, divergence *and* convergence can be observed, depending on the infimum of the noise distribution.

4.1 Positive lower bound: the convergent case

In this section, the infimum of the noise m_ξ is strictly positive.

Theorem 1 (Almost sure convergence) *The (1+1)-ES minimizing the noisy sphere (Eq. 4) defined in Eq. 5 converges if $m_\xi > 0$, in the sense that, the sequences $(\mathcal{F}(\mathrm{X}_n))_n$ and $(\|\mathrm{X}_n\|)_n$ converge respectively to a random variable $l \geq g(0)$ and 0 almost surely. If in addition, the function g is continuous in 0, the limit l equals $g(0)$.*

Proof: see page 27

As for the proof of [21, Proposition 1], the proof of this theorem is based on the use of the Borel-Cantelli Lemma that we remind on page 27. The difference between the result in [21, Proposition 1] and the result of Theorem 1 is that the limit value of the sequence $(\mathcal{F}(\mathrm{X}_n))_n$ is zero in [21, Proposition 1],

however, it depends here on the properties of the function g in the neighborhood of zero.

4.2 Negative greatest lower bound: the non convergent cases

We investigate here the case where $m_\xi < 0$. Negative noise values can be sampled and will be sampled as stated in the following Lemma:

Lemma 1 *Consider the sequences* $(\mathrm{X}_n)_n$ *and* $(O_n)_n$ *defined respectively in Eq. 5 and Eq. 6. There exists* $n_1 \geq 0$ *such that, for all* $n \geq n_1$, $O_n < 0$. *Moreover, if* $m_\xi > -\infty$ *then* $\|\mathrm{X}_n\| \geq A := \|\mathrm{X}_{n_1}\| \left(\frac{|O_{n_1}|}{|m_\xi|}\right)^{\frac{1}{\alpha}} > 0$ *almost surely.*

Proof: see page 31

Note that the lower bound A given in this lemma is represented in Figure 1 for $\alpha = 2$. As a consequence of this lemma, for $n \geq n_1$, all accepted fitness values are smaller than $g(0)$ and associated with a negative noise value. This also means for a new solution, that, given a negative noise value and therefore the situation of its potential acceptance, solutions further away from the optimum produce better fitness values and are more likely to be accepted.

4.2.1 Divergence case

In the case where $m_\xi < 0$ is finite, we have the following result:

Theorem 2 (Almost sure divergence) *The (1+1)-ES minimizing the noisy sphere (Eq. 4) defined in Eq. 5 diverges if* $m_\xi < 0$ *and* $m_\xi > -\infty$, *in the sense that the sequence* $(\|\mathrm{X}_n\|)_n$ *diverges to* $+\infty$ *almost surely.*

Proof: see page 32

The proof of this theorem heavily relies on the fact that m_ξ is finite. In the case where $m_\xi = -\infty$, a weaker result can be derived though as presented in the next section.

4.2.2 Negative expected quality gain

We now investigate the case where the infimum of the support of the noise m_ξ can equal $-\infty$, i.e., $-\infty \leq m_\xi < 0$.

Theorem 3 *Consider the sequences $(O_n)_n$ and $(\mathrm{X}_n)_n$ defined by the recurrence relations Eq. 5 and Eq. 6 for the minimization of the objective function defined in Eq. 4. If $m_\xi < 0$ then, for $n \geq n_1$ (where n_1 is defined in Lemma 1), the sequence of the expectations of the distances to the optimum of the non noisy objective function is increasing in the sense that $E\left(\frac{\|\mathrm{X}_{n+1}\|^2}{\|\mathrm{X}_n\|^2} \mid \mathrm{X}_n, O_n, \xi_n\right) \geq 1$. Therefore, for $n \geq n_1$, $E(\|\mathrm{X}_n\|^2) \geq E(\|X_{n_1}\|^2) > 0$, and the sequence $(E(\|\mathrm{X}_n\|)^2)_n$ cannot converge to zero.*

Proof: see page 33

Comparison with previous results Theorem 3 includes the particular case of Gaussian noise where $m_\xi = -\infty$. Therefore, the algorithm cannot converge in the L^2 norm in the case of Gaussian noise. This result implies a negative expected quality gain $E\left(1 - \frac{\|\mathrm{X}_{n+1}\|^2}{\|\mathrm{X}_n\|^2} | \mathrm{X}_n\right)$ (see definition in [4]), opposed to the result of Arnold and Beyer [20] that derived a positive expected quality gain for the objective function model defined in Eq. 3. Theorem 3 proves that the approximations used in [20] do not hold in finite dimensional search spaces.

The results in [20] are presented with numerical experiments that seem to backup the positive expected quality gain. But, for $m_\xi < 0$, a run of the (1+1)-ES exhibits two entirely different phases: an initial transient phase (before n_1 in Lemma 1) and a final stationary phase. During the transient phase all noise value realisations are positive and convergence can be observed (as in Figure 3, left). The final stationary phase begins after the first negative noise value has been sampled. In the stationary phase the strategy diverges[3] (as in Figure 3, right) for any $\sigma > 0$. In case of normalized Gaussian noise [4], the behavior switches from the transient to the stationary phase when $\xi = \frac{2\sigma_\epsilon^*}{d}\mathcal{N}(0,1) + 1$ is negative for the first time. The length of the transient phase follows a geometric distribution with mean

[3]More precisely, the strategy diverges, if the sequence O_n admits a stable distribution (see below), which needs yet to be shown. If, in contrast, O_n diverges to infinity, the update of X_n will stall and neither divergence nor convergence will occur.

$P(\xi < 0)^{-1} = CDF_{\mathcal{N}}(-d/(2\sigma_\epsilon^*))^{-1} \in]2, \infty[$, where $CDF_{\mathcal{N}}$ denotes the cumulative distribution function of the standard normal distribution. Most of the experiments in [20] have been conducted with $\sigma_\epsilon^* \leq 2$ and $d \geq 40$. For such a configuration, the probability that a negative fitness value is sampled is smaller than 10^{-23}. Therefore, the probability to leave the transient phase before n_1 time steps is smaller than $1 - (1 - 10^{-23})^{n_1} \approx n_1 10^{-23}$, i.e. even after 10^{10} time steps the probability is as low as 10^{-13}. Consequently, the transient convergence is observed in such a setting. In contrast, for $\sigma_\epsilon^* = 2$ and $d = 10$, the probability to remain in the transient phase even for only 1000 time steps is smaller than 1%.

5 Log-linear behavior for lower bounded noise

It is generally observed in the case of optimization with Evolution Strategies (ES) and theoretically proven in the case of minimization of non-noisy sphere functions, using either the artificial scale-invariant adaptation rule [13, 11, 12] or the real self-adaptation rule [13, 14], that ESs converge (or diverge) log-linearly in the sense of Eq. 2. In Fig. 2, we have observed log-linear behavior for the (1+1)-ES minimizing a noisy spherical function. The goal of this section is to prove this log-linear behavior when the noise is lower bounded, i.e., $m_\xi > -\infty$.

The main ingredient used in previous studies to prove log-linear behavior is the law of large numbers (LLN): the LLN for independent random variables in [13, 11], the LLN for orthogonal random variables in [12] and the LLN for Markov chains in [14]. In our case, the correlation between the variables in Eqs. 5 and 6 suggests the use of the LLN for Markov chains.

5.1 Motivations

Log-linear behavior means that, after an adaptation time, the sequence $(\ln \|\mathrm{X}_n\|)_n$, where $(\|\mathrm{X}_n\|)_n$ is defined in Eq. 5, increases or decreases linearly with the number of iterations. This suggests that one has to investigate the sequence $(\ln (\|\mathrm{X}_n\|))_n$. The following proposition is the first step for proving log-linear behavior. It expresses the term $\frac{1}{n} \ln (\|X_n\|/\|X_0\|)$ as the sum of n random variables divided by n. The same idea has been previously used in [13, 14, 11, 12].

Lemma 2 *Let $(\mathrm{X}_n)_n$ be the sequence of random vectors valued in $\mathbb{R}^d$ satisfying the recurrence relation (5). Then, for all $n \geq 1$, the equality*

$$\frac{1}{n}\ln\left(\frac{\|\mathrm{X}_n\|}{\|\mathrm{X}_0\|}\right) = \frac{1}{n}\sum_{k=0}^{n-1}\ln\left(\left\|\frac{\mathrm{X}_k}{\|\mathrm{X}_k\|} + \sigma\mathcal{N}_k 1_{\left\{\left\|\frac{\mathrm{X}_k}{\|\mathrm{X}_k\|}+\sigma\mathcal{N}_k\right\|^{\alpha}\xi_k<O_k\right\}}\right\|\right) \tag{9}$$

holds almost surely.

Proof: see page 35

The previous lemma suggests the use of a LLN for proving the convergence of the right hand side of Eq. 9. However, we are not going to apply directly the LLN to the right hand side but will first exploit the invariance under rotation of the multivariate normal distribution to simplify the right hand side. This will be done at the price of losing the almost sure equality.

As we will see below, invariance under rotation implies that the sequence $(O_n)_n$ is a Markov chain. A Markov chain $(\Phi_n)_n$ taking values in $\mathbb{R}^p$, $p \in \mathbb{N}$, is entirely characterized by its initial law, i.e., the law of Φ_0 and its *transition kernel* $P_\Phi(.,.)$ where for all $x \in \mathbb{R}^p$, $P_\Phi(x,.)$ is a probability measure and for all $A \in \mathfrak{B}(\mathbb{R}^p)$, $P_\Phi(.,A)$ is a non-negative measurable function on $\mathbb{R}^p$ and for all $x \in \mathbb{R}^p$, for all $A \in \mathfrak{B}(\mathbb{R}^p)$

$$P_\Phi(x,A) = P(\Phi_1 \in A | \Phi_0 = x) \ .$$

The sequence of random variables $(O_n)_n$ defined in Eq. 8 can be written in a more compact manner as

$$O_{n+1} = O_n + (\xi_n - O_n)1_{\{\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathcal{N}_n\|^\alpha \xi_n < O_n\}} \tag{10}$$

where O_0 follows the law $\mathcal{L}_\xi$. The following proposition states that O_n is a Markov chain, derives its transition kernel and shows a more convenient way to generate a sequence following the same distribution as O_n.

Proposition 1 *The sequence $(O_n)_n$ is a Markov chain with the same initial law and transition kernel as the Markov chain $(Z_n)_n$ defined as*

$$Z_{n+1} = Z_n + (\xi_n - Z_n)1_{\{\|\mathrm{e}_1+\sigma\mathcal{N}_n\|^\alpha \xi_n < Z_n\}} \tag{11}$$

where Z_0 is distributed according to $\mathcal{L}_\xi$.

For all $n \in \mathbb{Z}^+$, O_n and $Z_n \in]m_\xi, M_\xi[$, their common initial law is $\mathcal{L}_\xi$ and their transition kernel $P(.,.)$ satisfies for all $z \in]m_\xi, M_\xi[$, for all $A \in \mathfrak{B}(]m_\xi, M_\xi[)$,

$$P(z, A) = P_1(z, A) + \delta_z(A)P_2(z) \tag{12}$$

where $P_1(z, A)$ equals $P\left(\{\xi_0 \in A\} \cap \{\|\mathrm{e}_1 + \sigma\mathcal{N}_0\|^\alpha \xi_0 < z\}\right)$, δ_z is the Dirac measure centered at z and $P_2(z) = P\left(\|\mathrm{e}_1 + \sigma\mathcal{N}_0\|^\alpha \xi_0 \geq z\right)$.

Proof: see page 36

We now define the function $F(Z_n, \mathcal{N}_n, \xi_n)$ as $\ln(\|\mathrm{e}_1 + \sigma\mathcal{N}_n 1_{\{\|\mathrm{e}_1+\sigma\mathcal{N}_n\|^\alpha \xi_n < Z_n\}}\|)$ corresponding to the inner part of the right hand side of Eq. 9 where steps are sampled from e_1 and O_n is replaced by Z_n. The following lemma makes the connection between the term whose limit we want to investigate $(\frac{1}{n} \ln \|\mathrm{X}_n\|)$ and the sample average $\frac{1}{n}\sum_{k=0}^{n-1} F(Z_k, \mathcal{N}_k, \xi_k)$:

Lemma 3 *For $n \geq 1$, the following equation holds in distribution*

$$\frac{1}{n} \ln\left(\frac{\|X_n\|}{\|X_0\|}\right) = \frac{1}{n}\sum_{k=0}^{n-1} F(Z_k, \mathcal{N}_k, \xi_k) \ , \tag{13}$$

where F is defined as

$$F(Z_n, \mathcal{N}_n, \xi_n) = \ln\left(\|\mathrm{e}_1 + \sigma\mathcal{N}_n 1_{\{\|\mathrm{e}_1+\sigma\mathcal{N}_n\|^\alpha \xi_n < Z_n\}}\|\right) \tag{14}$$

Proof: see page 37

Consequently, if the right-hand side of Eq. 13 converges almost surely to a finite value γ, then $\frac{1}{n} \ln \|\mathrm{X}_n\|$ will also converge (in probability) to γ. In Proposition 1, we have established that $(Z_n)_n$ is a Markov chain. Besides, Eq. 13 expresses Z_{n+1} as a function of $(Z_n, \mathcal{N}_n, \xi_n)$ where $(\mathcal{N}_n)_n$ and $(\xi_n)_n$ are independent sequences. Therefore, $(Z_n, \mathcal{N}_n, \xi_n)$ is also a Markov chain. The sample average $\frac{1}{n}\sum_{k=0}^{n-1} F(Z_k, \mathcal{N}_k, \xi_k)$ converges to a constant γ if the law of large numbers (LLN) holds for the Markov chain $(Z_n, \mathcal{N}_n, \xi_n)_n$. If in addition $\gamma \neq 0$, then log-linear behavior holds in probability for the sequence $(\|\mathrm{X}_n\|)_n$ given in Eq. 5. We now understand that we need to establish a LLN for the Markov chain $(Z_n, \mathcal{N}_n, \xi_n)_n$ in order to conclude the log-linear convergence of $(\|\mathrm{X}_n\|)_n$. This is what we will do in the next section.

5.2 Stability

We have argued that log-linear behavior will follow from establishing a LLN for $(Z_n, \mathcal{N}_n, \xi_n)_n$, however since $(\mathcal{N}_n, \xi_n)$ is independent of Z_n, we will see that the properties needed to establish a LLN for the Markov chain $(Z_n, \mathcal{N}_n, \xi_n)_n$ follow from the properties needed to establish a LLN for the chain $(Z_n)_n$. The chain $(Z_n)_n$ satisfies a LLN if certain so-called *stability criteria* can be proven. Before investigating stability criteria for the chain $(Z_n)_n$, we recall some definitions and results about φ-irreducible Markov Chains that are used in the sequel. We refer to Meyn and Tweedie [24] for a complete presentation of this theory.

5.2.1 Basics about Markov chains and definitions

Given a Markov chain $(\Phi_n)_n \subset \mathbb{R}^p$, with transition kernel $P_\Phi(.,.)$, the weakest stability criterion is the so-called φ-irreducibility: a chain $(\Phi_n)_n$ is *irreducible with respect to a measure* φ if:

$$\forall (x, A) \in \mathbb{R}^p \times \mathfrak{B}(\mathbb{R}^p), \varphi(A) > 0, \exists\ n_0 \geq 0 \text{ such that } P_\Phi^{n_0}(x, A) > 0\,, \quad (15)$$

where $P_\Phi^{n_0}(x, A)$ equals $P(\Phi_{n_0} \in A | \Phi_0 = x)$. Another equivalent definition for the φ-irreducibility of the Markov chain $(\Phi_n)_n$ is: $\forall x \in \mathbb{R}^p, \forall A \in \mathfrak{B}(\mathbb{R}^p)$ such that $\varphi(A) > 0$, $P(\tau_A < +\infty | \Phi_0 = x) > 0$ where, τ_A is the hitting time of Φ_n on A, i.e.,

$$\tau_A = \min\{n \geq 1 \text{ such that } \Phi_n \in A\}.$$

If the last term of Eq. 15 is equal to one, the chain is *recurrent*. A φ-irreducible chain $(\Phi_n)_n$ is *Harris recurrent* if:

$$\forall A \in \mathfrak{B}(\mathbb{R}^p) \text{ such that } \varphi(A) > 0; P_x(\eta_A = \infty) = 1, x \in \mathbb{R}^p\,,$$

where η_A is the occupation time of A defined as $\eta_A = \sum_{n=1}^{\infty} 1_{\{\Phi_n \in A\}}$.
A chain $(\Phi_n)_n$ which is Harris-recurrent admits an *invariant measure*, i.e., a measure π on $\mathfrak{B}(\mathbb{R}^p)$ satisfying:

$$\pi(A) = \int_{\mathbb{R}^p} P_\Phi(x, A) \mathrm{d}\pi(x), A \in \mathfrak{B}(\mathbb{R}^p)\,.$$

If in addition this measure is a probability measure, the chain is called *positive*. Positive, Harris-recurrent chains satisfy the Strong law of large numbers (LLN) as stated in [24, Theorem 17.0.1] and recalled here.

Theorem 4 (LLN for Harris positive chains) *Suppose that $(\Phi_n)_n$ is a positive Harris chain with invariant probability measure π, then the LLN holds for any function G satisfying $\pi(|G|) = \int |G(x)| \mathrm{d}\pi(x) < \infty$, i.e.,*

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} G(\Phi_k) = \pi(G)\,. \tag{16}$$

To show Harris-recurrence or positivity, it is possible to make use of practical drift conditions. However, here, those stability criteria will be implied from a stronger property called *uniform ergodicity*. A Markov chain $(\Phi_n)_n$ is said to be uniformly ergodic if it is positive Harris-recurrent and

$$\lim_{n\to\infty} \sup_{x\in\mathbb{R}^p} \|P_\Phi^n(x,.) - \pi(.)\| = 0\,,$$

where π is the invariant probability measure and $\|\nu\| = \sup_{g:|g|\leq 1} |\int g(x) \mathrm{d}\nu(x)|$ is the so-called total variation norm. Uniform ergodicity can be shown using the following theorem which is derived from a specific case of [24, Theorem 16.2.1, Theorem 16.2.4].

Theorem 5 (Condition for uniform ergodicity) *Suppose that there exists a finite non-trivial measure ν on $\mathfrak{B}(\mathbb{R}^p)$ such that a Markov chain $(\Phi_n)_n$ satisfies $P_\Phi(x, A) \geq \nu(A)$ for all $x \in \mathbb{R}^p$ and $A \in \mathfrak{B}(\mathbb{R}^p)$. Then $(\Phi_n)_n$ is uniformly ergodic and thus positive and Harris-recurrent.*

5.2.2 Stability

In the following, we study the Markov chain $(Z_n, \mathcal{N}_n, \xi_n)_n$. Its stability will follow from the use of Theorem 5 and consequently the LLN given in Theorem 4 will hold for $(Z_n, \mathcal{N}_n, \xi_n)_n$. Since $(\mathcal{N}_n, \xi_n)$ is independent of Z_n, the stability of $(Z_n, \mathcal{N}_n, \xi_n)_n$ will follow almost immediately from the stability of Z_n that we will investigate separately. The transition kernel of $(Z_n)_n$ verifies the following minorization condition:

Proposition 2 (Doeblin or minorization condition) *Let $m_\xi \neq 0$, and the non-trivial measure ν be defined as*

$$\nu(A) = \int_{\mathbb{R}^d} \int_{m_\xi}^{M_\xi} 1_A(u) 1_{\{\|\mathrm{e}_1 + \sigma t\|^\alpha u < m_\xi\}}(u,t) p_{\mathcal{N}}(t) p_\xi(u) \mathrm{d}u\mathrm{d}t\ .$$

Then, $\forall z \in]m_\xi, M_\xi[, \forall A \in \mathfrak{B}(]m_\xi, M_\xi[)$ we have $P_1(z, A) \geq \nu(A)$.

Proof: see page 41

The following corollary holds as a direct consequence of the application of Theorem 5 using the result of Proposition 2.

Corollary 1 *If $m_\xi \neq 0$, the chain $(Z_n)_n$ is positive Harris recurrent.*

Proof: see page 42

From the previous corollary, we know that $(Z_n)_n$ admits an invariant probability measure. We will denote this measure as μ. Let $\nu_\mathcal{N}$ be the probability measure defined on $\mathfrak{B}(\mathbb{R}^d)$ associated with a multivariate normal distribution, i.e. for all $A \in \mathfrak{B}(\mathbb{R}^d)$, $\nu_\mathcal{N}(A) = \int_A p_\mathcal{N}(x)dx$. Then for all $n \in \mathbb{Z}^+$, $P(\mathcal{N}_n \in A) = \nu_\mathcal{N}(A)$ for $A \in \mathfrak{B}(\mathbb{R}^d)$. In the same way let ν_ξ be the probability measure defined on $\mathfrak{B}(\mathbb{R})$ associated with the noise distribution, i.e. for all $A \in \mathfrak{B}(\mathbb{R})$, $\nu_\xi(A) = \int_A p_\xi(x)dx$.

Corollary 2 *If $m_\xi \neq 0$, the chain $(Z_n, \mathcal{N}_n, \xi_n)_n$ is positive Harris recurrent admitting the product measure $\mu \otimes \nu_\mathcal{N} \otimes \nu_\xi$ as invariant measure.*

Proof: see page 42

We are now ready to state the main result of this section.

Theorem 6 *The (1+1)-ES defined in Eq. 5 (and Eq. 6) minimizing the noisy sphere (Eq. 4) converges almost surely to zero if $m_\xi > 0$ and diverges almost surely to infinity when $-\infty < m_\xi < 0$. For $m_\xi \neq 0$, let γ be defined as*

$$\gamma := \int E\left(\ln \|\mathrm{e}_1 + \sigma\mathcal{N}_0 1_{\{\|\mathrm{e}_1+\sigma\mathcal{N}_0\|^\alpha \xi_0 \leq z\}}\|\right) d\mu(z) \tag{17}$$

where μ is the invariant probability measure of the Markov chain $(Z_n)_n$ (Eq. 11). Then γ is well defined, finite and the algorithm converges (or diverges) log-linearly in the sense that:

$$\frac{1}{n} \ln \|\mathrm{X}_n\| \to \gamma \tag{18}$$

holds in probability. Moreover, the convergence (or divergence) rate γ is strictly negative if $m_\xi > 0$ and strictly positive if $m_\xi < 0$.

Proof: see page 42

The convergence rate γ depends on σ. The following proposition establishes that the mapping $\sigma \to \gamma(\sigma)$ is continuous.

Proposition 3 *The mapping $\sigma \mapsto \gamma(\sigma)$ is continuous on $]0, +\infty[$.*

Proof: see page 46

6 Discussion

In this paper, we have developed a rigorous theory of convergence for the (1+1)-ES together with convergence rates in noisy environments. Note that other rigorous theoretical studies exist, however restricted to a discrete and finite search space [25, 26]. We have analyzed the (1+1)-ES on a class of unimodal and spherical noisy fitness functions. For this class, where the noise is multiplicative, we have proven rigorously that even when the expected fitness value is a positive function with a unique minimum, convergence *and* divergence can happen. The result is largely independent of the type of noise distribution and the limit between convergence and divergence is *only* determined by the sign of the greatest lower bound of the support of the noise distribution—after a first negative noise value is observed, the (1+1)-ES cannot converge to the optimum. Previously obtained approximative results that suggest convergence with Gaussian noise do *not* hold in finite dimensional search spaces.

Though the proofs were carried out for the (1+1)-ES, i.e., an elitist evolution strategy, the underlying mechanism suggests that similar results hold for other algorithms. For a negative greatest lower bound of the noise distribution, the same divergent behavior is foreseeable for elitist algorithms in general. We have seen with numerical simulations that the qualitative result was the same also for non elitist selection (comma selection). However, the limit case between convergence and divergence will be different, less easy to specify and it will depend on strategy parameters. In particular, divergence with comma selection can happen even when the lower bound of the noise support is positive, and convergence can happen even when it is negative.

Another variant of the (1+1)-ES for noisy optimization is the (1+1)-ES with reevaluation of the fitness of the parent, investigated in [27]. Arnold

[4] conjectures that the best policy is to reevaluate the parent not in each iteration. In any case, the (1+1)-ES with reevaluation loses its monotonicity of the fitness value of the parent. Behavior and analysis of the (1+1)-ES with reevaluation is closer to the (1,λ)-ES than to the (1+1)-ES without reevaluation and less intricate, in particular if the parent is evaluated in each iteration.

The divergent behavior of the (1+1)-ES for example on the classical noise model

$$\mathcal{F}(x) = f(x)(1 + \sigma_\epsilon \mathcal{N}(0, 1)) \tag{19}$$

comes as a surprise. Whether the divergent behavior can be easily observed in a simulation depends on the setting of the parameter σ_ϵ. For $\sigma_\epsilon \geq 0.5$ the observation can be made easily (see Section 4.2.2 and Figure 3). We believe that this observation must have been made before by other researchers, but that it has been ignored due to the lack of a reasonable explanation and in view of the fact that for smaller σ_ϵ it could not be reproduced. This paper explains the observation of divergence in a rigorous way.

The practical relevance of our analysis is limited for the following reasons.

1. The investigated (1+1)-ES lacks a realistic step-size adaptation procedure. From the view point of choosing such a procedure the present analysis investigates the best case scenario. Our justification for this limitation is twofold. First, our knowledge on the functioning of step-size adaptation procedures for the (1+1)-ES in the presence of noise is rather limited and would deserve a considerable amount of empirical or theoretical research on its own. Second, the best case scenario provides a useful comparison for the evaluation of step-size adaptation procedures to be developed. The rigorous analysis of the (1+1)-ES with step-size adaptation will be more involved and therefore left to future work.

2. Only spherical functions have been considered and for any sequence converging to the optimum the noise level converges to zero. This scale-invariance property is essential for proving log-linear convergence, but not necessarily realistic in practice.

3. The (1+1)-ES is not the most promising strategy to be applied in noisy environments. Non-elitist comma strategies are more advisable [4] and also covariance matrix adaptation (CMA) is most likely advantageous even under noisy conditions [28, 29].

Our results reveal a potential limitation of the formulation of noisy objective functions. We feel that the observed divergence should primarily be interpreted as deficiency of a multiplicative noise model with negative greatest lower bound and give two reasons. First, we do not believe that in practice worse solutions (in terms of their *expected* fitness value) tend to produce exceptionally good fitness values with a higher probability than better solutions. Second, the underlying mechanism of divergence is not limited to the (1+1)-ES but applies to any elitist algorithm.

Consequently, our result suggests an implication for the construction of benchmark functions as for example the typical noisy benchmark Eq. 19. We suggest to replace the typical model by

$$\mathcal{F}(x) = f(x) \exp(\sigma_\epsilon \mathcal{N}(0,1)) \ . \tag{20}$$

For small (positive) values of σ_ϵ, both models are very similar and they align for $\sigma_\epsilon \to 0$, e.g. in the common analytical approach with $\sigma_\epsilon \propto 1/d$ and $d \to \infty$. For larger σ_ϵ the new model seems more realistic because better solutions also potentially deliver the best fitness values.

One important reason for the success of stochastic search algorithms like evolution strategies is their fast convergence rate to the optimum (log-linear convergence) empirically observed even on non-convex and rugged fitness functions. Log-linear convergence is the lower bound for rank-based search algorithms like ESs even in the non-noisy case [30]. In this paper, we have proven that log-linear convergence (and divergence) for the (1+1)-ES are preserved in the presence of noise. The class of functions considered includes non-convex, non-differentiable and non-smooth functions even when the noise is set to a non-zero constant. This remarkable result shows the robustness of ESs in the presence of noise for the first time rigorously.

Acknowledgments

This work was supported by the French National Research Agency (ANR) grant No. ANR-08-COSI-007-12.

References

[1] Volker Nissen and Jörn Propach. On the robustness of population-based versus point-based optimization in the presence of noise. *IEEE Trans. Evolutionary Computation*, 2(3):107–119, 1998.

[2] 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.

[3] Y. Jin and J. Branke. Evolutionary Optimization in Uncertain Environments-A Survey. *IEEE Transactions on Evolutionary Computation*, 9(3):303–317, June 2005.

[4] D. V. Arnold. *Noisy Optimization with Evolution Strategies*. GENA. Kluwer Academic Publishers, 2002.

[5] P.N. Suganthan, N. Hansen, J.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.

[6] N. Hansen, S. Finck, R. Ros, and A. Auger. Real-parameter black-box optimization benchmarking 2009: Noisy functions definitions. Technical Report RR-6869, INRIA, 2009.

[7] M. Schumer and K. Steiglitz. Adaptive step size random search. *Automatic Control, IEEE Transactions on*, 13:270–276, 1968.

[8] I. Rechenberg. *Evolutionstrategie: Optimierung Technisher Systeme nach Prinzipien des Biologischen Evolution*. Fromman-Hozlboog Verlag, Stuttgart, 1973.

[9] Hans-Paul Schwefel. Collective phenomena in evolutionary systems. In P. Checkland and I. Kiss, editors, *Problems of Constancy and Change – The Complementarity of Systems Approaches to Complexity, Proc. 31st Annual Meeting*, volume 2, pages 1025–1033, Budapest, 1987. Int'l Soc. for General System Research.

[10] Nikolaus Hansen and Andreas Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

[11] A. Auger and N. Hansen. Reconsidering the progress rate theory for evolution strategies in finite dimensions. In ACM Press, editor, *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2006)*, pages 445–452, 2006.

[12] 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.

[13] Alexis Bienvenüe and Olivier François. Global convergence for evolution strategies in spherical problems: some simple proofs and difficulties. *Theoretical Computer Science*, 306(1-3):269–289, 2003.

[14] Anne Auger. Convergence results for (1,λ)-SA-ES using the theory of φ-irreducible markov chains. *Theoretical Computer Science*, 334(1–3):35–69, 2005.

[15] Jens Jägersküpper. Analysis of a simple evolutionary algorithm for minimization in euclidean spaces. *Theoretical Computer Science*, 379(3):329–347, 2007.

[16] Jens Jägersküpper. Rigorous runtime analysis of the (1+1)-ES: 1/5-rule and ellipsoidal fitness landscapes. In Springer LNCS, editor, *Foundations of Genetic Algorithms: 8th International Workshop, FoGA 2005*, volume 3469, pages 260–281, 2005.

[17] D. V. Arnold and H.-G. Beyer. Efficiency and mutation strength adaptation of the $(\mu/\mu_I, \lambda)$-ES in a noisy environment. In Marc Schoenauer et al, editor, *Proceedings of Parallel Problem Solving from Nature - PPSN VI*, volume 1917 of *LNCS*, pages 39–48. Springer, 2000.

[18] 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.

[19] H.-G. Beyer. *The Theory of Evolution Strategies*. Natural Computing Series. Springer-Verlag, 2001.

[20] 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.

[21] M. Jebalia and A. Auger. On multiplicative noise models for stochastic search. In Günter Rudolph, Thomas Jansen, Simon Lucas, Carlo Polini, and Nicola Beume, editors, *Proceedings of Parallel Problem Solving from Nature (PPSN X)*, volume 5199 of *Lecture Notes in Computer Science*, pages 52–61. Springer Verlag, 2008.

[22] Günter Rudolph. *Convergence Properties of Evolutionary algorithms*. Verlag Dr. Kovac, Hamburg, 1997.

[23] Dirk V. Arnold. Optimal weighted recombination. In *Foundations of Genetic Algorithms 8*, pages 215–237. Springer Verlag, 2005.

[24] S.P. Meyn and R.L. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, New York, 1993.

[25] T. Nakama. Theoretical analysis of genetic algorithms in noisy environments based on markov model. In ACM Press, editor, *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2008)*, pages 1001–1008, 2008.

[26] T. Nakama. Markov chain analysis of genetic algorithms in a wide variety of noisy environments. In ACM Press, editor, *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2009)*, pages 827–834, 2009.

[27] Hans-Georg Beyer. Toward a theory of evolution strategies: Some asymptotical results from the (1,+ λ)-theory. *Evolutionary Computation*, 1(2):165–188, 1993.

[28] N. Hansen, A. Niederberger, L. Guzzella, and P. Koumoutsakos. A method for handling uncertainty in evolutionary optimization with an application to feedback control of combustion. *IEEE Transactions on Evolutionary Computation*, 13(1):180–197, 2009.

[29] N. Hansen. Benchmarking a BI-population CMA-ES on the BBOB-2009 noisy testbed. In G. Raidl et al., editor, *Workshop Proceedings of*

the GECCO Genetic and Evolutionary Computation Conference, pages 2397–2402. ACM, July 2009.

[30] O. Teytaud and S. Gelly. General lower bounds for evolutionary algorithms. In 10^{th} *International Conference on Parallel Problem Solving from Nature (PPSN 2006)*, volume 4193, pages 21–31. Springer Berlin, 2006.

[31] D. Williams. *Probability with Martingales*. Cambridge University Press, Cambrdige, 1991.

Appendix

We provide in the appendix the proofs of the theorems stated in the core of the paper. The proofs often require intermediate results, organized in lemmas and propositions that are stated and proven before to tackle the proofs of the main results.

Further notations The set of strictly negative real numbers is denoted $\mathbb{R}_-^*$ and the set of strictly positive real numbers $\mathbb{R}_+^*$. The offspring sampled at iteration n is denoted $\tilde{\mathrm{X}}_n$, i.e., $\tilde{\mathrm{X}}_n := \mathrm{X}_n + \sigma_n \mathcal{N}_n$. The smallest σ-algebra on Ω such that a random variable X defined on $(\Omega, \mathcal{A}, P)$ is measurable with respect to this σ-algebra is denoted $\boldsymbol{\sigma}(X)$, of course $\boldsymbol{\sigma}(X) \subset \mathcal{A}$. In a similar way, the smallest σ-algebra such that $X_1, \ldots, X_n$ are measurable with respect to the σ-algebra is denoted $\boldsymbol{\sigma}(X_1, \ldots, X_n)$. In the sequel we will sometimes abbreviate "almost surely" by "a.s.".

The following technical lemma will be useful for several proofs.

Lemma 4 *The sequence* $(\mathrm{X}_n)_n$ *introduced in Eq. 5 satisfies: for every* $n \geq 0$, $\|\mathrm{X}_n\| \neq 0$ *almost surely.*

Proof The result is proved by induction. The first parent is chosen randomly with $P(\|\mathrm{X}_0\| = 0) = 0$. Suppose that $P(\|\mathrm{X}_n\| = 0) = 0$. As the offspring $\tilde{\mathrm{X}}_n$ is obtained by adding to X_n a random vector admitting an absolutely continuous distribution with respect to the Lebesgue measure then $P(\|\tilde{\mathrm{X}}_n\| = 0) = 0$. By induction hypothesis, $P(\|\mathrm{X}_n\| \neq 0) = 1$, $P(\|\mathrm{X}_{n+1}\| = 0) = P(\|\mathrm{X}_{n+1}\| = 0 \cap \|\mathrm{X}_n\| \neq 0)$. Besides, the probability

$P(\|\mathrm{X}_{n+1}\| = 0 \cap \|\mathrm{X}_n\| \neq 0)$ equals the sum of the probability of the event $A := (\|\mathrm{X}_{n+1}\| = 0) \cap (\|\mathrm{X}_n\| \neq 0) \cap (\tilde{\mathrm{X}}_n$ is accepted) and the probability of the event $B := (\|\mathrm{X}_{n+1}\| = 0) \cap (\|\mathrm{X}_n\| \neq 0) \cap (\tilde{\mathrm{X}}_n$ is not accepted). Moreover, $A = (\|\tilde{\mathrm{X}}_n\| = 0) \cap (\|\mathrm{X}_n\| \neq 0)$ and thus $P(A) = 0$. Also, the event $(\|\mathrm{X}_{n+1}\| = 0) \cap (\tilde{\mathrm{X}}_n$ is not accepted) equals the event $(\|\mathrm{X}_n\| = 0) \cap (\tilde{\mathrm{X}}_n$ is not accepted) which implies that $B = (\|\mathrm{X}_n\| = 0) \cap (\|\mathrm{X}_n\| \neq 0) \cap (\tilde{\mathrm{X}}_n$ is not accepted) is the empty set and thus $P(B) = 0$. Since we have seen that $P(\|\mathrm{X}_{n+1}\| = 0) = P(A) + P(B)$, we thus obtain that $P(\|\mathrm{X}_{n+1}\| = 0) = 0$. □

Proofs of Theorem 1 and Theorem 2 heavily rely on the second Borel-Cantelli Lemma that we recall below. But first, we need to introduce the following formal definition of 'infinitely often (i.o.)':

Definition 1 *Let q_n be some statement, e.g. $|a_n - a| > \epsilon$. 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 E_n in a probability space, $(E_n$ i.o.$)$ equals $\{w | w \in E_n$ i.o.$\} = \cap_{n\geq 0} \cup_{m \geq n} E_m =: \overline{\lim}\, E_n$.*

Given this definition, the second Borel-Cantelli Lemma (BCL) states that:

Lemma 5 *Let $(E_n)_{n\geq 0}$ be a sequence of events in some probability space. If the events E_n are independent and verify $\sum_{n\geq 0} P(E_n) = +\infty$ then $P(\overline{\lim}\, E_n) = 1$.*

Proof of Theorem 1 (stated page 11)

Before to be able to prove Theorem 1, we need to establish three technical lemmas.

Lemma 6 *Let $t \in \mathbb{R}$ and let h_t be the mapping from $\mathbb{R}^d$ to $\mathbb{R}$ defined for all $x \in \mathbb{R}^d$ as $h_t(x) = E(e^{it\|x+\sigma\mathcal{N}\|})$ where $\mathcal{N}$ is distributed as $\mathcal{N}(0, I_d)$. Then, for all vectors u_1, u_2 with $\|u_1\| = \|u_2\| = 1$, $h_t(u_1) = h_t(u_2)$ and thus without loss of generality, for all u, $\|u\| = 1$,*

$$h_t(u) = E(e^{it\|\mathrm{e}_1+\sigma\mathcal{N}\|}) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{it\|\mathrm{e}_1+\sigma x\|} e^{-\frac{\|x\|^2}{2}} dx$$

where e_1 is the vector $(1, 0, \ldots, 0)$.

Proof This result is a consequence of the fact that the standard d-dimensional normal distribution $\mathcal{N}$ is spherical (isotropic). Let $x \in \mathbb{R}^d$ and $\|x\| = 1$, $h_t(x) = E(e^{it\|x+\sigma\mathcal{N}\|})$. Let R be an orthogonal matrix such that $Rx = e_1$. Since R is an orthogonal matrix, $\|Ry\| = \|y\|$ for all $y \in \mathbb{R}^d$ and therefore $\|x+\sigma\mathcal{N}\| = \|R(x+\sigma\mathcal{N})\|$ almost surely. Besides $\|R(x+\sigma\mathcal{N})\| = \|e_1+\sigma R\mathcal{N}\|$ but since $\mathcal{N}$ is spherical, $R\mathcal{N}$ has the same law as $\mathcal{N}$ and thus $\|x + \sigma\mathcal{N}\|$ and $\|e_1 + \sigma\mathcal{N}\|$ have the same distribution. Therefore they admit the same characteristic function, i.e. $E(e^{it\|x+\sigma\mathcal{N}\|}) = E(e^{it\|e_1+\sigma\mathcal{N}\|})$. □

Lemma 7 *Let $(\mathrm{X}_n)_n$ be the sequence of random vectors in $\mathbb{R}^d$ defined in Eq. 5 and $(\mathcal{N}_n)_n$ be the relative sequence of independent random vectors following the same distribution $\mathcal{N}(0, I_d)$ used to define the sequence $(\mathrm{X}_n)_n$ as shown in Eq. 5. Then the variables $Y_n := \left\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathcal{N}_n \right\|$, for $n \geq 0$, are independent and follow the same distribution as $Y := \|\mathrm{e}_1 + \sigma\mathcal{N}(0, I_d)\|$, where e_1 is the vector $(1, 0, \ldots, 0)$.*

Proof For showing the independence of $(Y_n)_{n\in\mathbb{Z}^+}$, we will prove that for all n, for all $t_0 \in \mathbb{R}, \ldots, t_n \in \mathbb{R}$, $E(e^{it_0Y_0} \ldots e^{it_nY_n}) = E(e^{it_0Y_0}) \ldots E(e^{it_nY_n})$. We will proceed by induction and suppose that for all $t_0 \in \mathbb{R}, \ldots, t_{n-1} \in \mathbb{R}$ $E(e^{it_0Y_0} \ldots e^{it_{n-1}Y_{n-1}}) = E(e^{it_0Y_0}) \ldots E(e^{it_{n-1}Y_{n-1}})$ and prove that for all $t_0 \in \mathbb{R}, \ldots, t_n \in \mathbb{R}$

$$E(e^{it_0Y_0} \ldots e^{it_nY_n}) = E(e^{it_0Y_0}) \ldots E(e^{it_nY_n}).$$

Let ζ_n be the σ-algebra $\boldsymbol{\sigma}(X_0, \mathcal{N}_0, X_1, \mathcal{N}_1, \ldots, X_{n-1}, \mathcal{N}_{n-1}, X_n)$, let $t_0, \ldots, t_n \in \mathbb{R}^{n+1}$, then the following holds $E(e^{it_0Y_0} \ldots e^{it_nY_n}) = E(E(e^{it_0Y_0} \ldots e^{it_nY_n}|\zeta_n))$. Since $e^{it_0Y_0} \ldots e^{it_{n-1}Y_{n-1}}$ is bounded and ζ_n-measurable [31, p88, j]

$$E(e^{it_0Y_0} \ldots e^{it_nY_n}|\zeta_n) = e^{it_0Y_0} \ldots e^{it_{n-1}Y_{n-1}} E(e^{it_nY_n}|\zeta_n) \ . \tag{21}$$

Besides, $E(e^{it_nY_n}|\zeta_n) = E\left(e^{it_n\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathcal{N}_n\right\|}|\zeta_n\right)$. By independence of $\mathcal{N}_n$, $E\left(e^{it_n\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathcal{N}_n\right\|}|\zeta_n\right) = h_{t_n}(X_n/\|X_n\|)$ where h_{t_n} is defined in Lemma 6. Since the norm of the vector $\mathrm{X}_n/\|\mathrm{X}_n\|$ is 1 we know from Lemma 6 that $E\left(e^{it_n\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathcal{N}_n\right\|}|\zeta_n\right) = E(e^{it_n\|\mathrm{e}_1+\sigma\mathcal{N}\|})$. Injecting this in Eq. 21, we obtain

$$E(e^{it_0Y_0} \ldots e^{it_nY_n}|\zeta_n) = e^{it_0Y_0} \ldots e^{it_{n-1}Y_{n-1}} E(e^{it_n\|\mathrm{e}_1+\sigma\mathcal{N}\|}) \tag{22}$$

We will now use again Lemma 6 to prove that $E(e^{it_n\|\mathrm{e}_1+\sigma\mathcal{N}\|}) = E(e^{it_nY_n})$: We start by the right hand side of the equation and decompose it using the conditional expectation with respect to $\boldsymbol{\sigma}(\mathrm{X}_n)$ and obtain that

$$E(e^{it_nY_n}) = E\left[E\left(e^{it_n\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathcal{N}_n\right\|}|\boldsymbol{\sigma}(\mathrm{X}_n)\right)\right] \ . \tag{23}$$

Moreover, by independence of $\mathcal{N}_n$, $E\left(e^{it_n\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathcal{N}_n\right\|}|\boldsymbol{\sigma}(\mathrm{X}_n)\right) = h_{t_n}(\mathrm{X}_n/\|\mathrm{X}_n\|)$. Using again Lemma 6 we have that $E(e^{it_nY_n}) = E(e^{it_n\|\mathrm{e}_1+\sigma\mathcal{N}\|})$. Injecting this result in Eq. 22, we obtain the following equation

$$E(e^{it_0Y_0}\dots e^{it_nY_n}|\zeta_n) = e^{it_0Y_0}\dots e^{it_{n-1}Y_{n-1}}E(e^{it_nY_n}) \ . \tag{24}$$

We take now the expectation of both sides of the previous equation and obtain $E(e^{it_0Y_0}\dots e^{it_nY_n}) = E(e^{it_0Y_0}\dots e^{it_{n-1}Y_{n-1}})E(e^{it_nY_n})$. Moreover by induction hypothesis we know that $E(e^{it_0Y_0}\dots e^{it_{n-1}Y_{n-1}}) = E(e^{it_0Y_0})\dots E(e^{it_{n-1}Y_{n-1}})$ which thus imply that

$$E(e^{it_0Y_0}\dots e^{it_nY_n}) = E(e^{it_0Y_0})\dots E(e^{it_{n-1}Y_{n-1}})E(e^{it_nY_n}) \ ,$$

which achieves to prove the independence of $(Y_n)_{n\in\mathbb{Z}^+}$. □

Lemma 8 *If $m_\xi > 0$, the following points hold:*

1. *The sequence $(\|\mathrm{X}_n\|)_n$ is upper bounded by $\theta := \|\mathrm{X}_0\|\left(\frac{O_0}{m_\xi}\right)^{\frac{1}{\alpha}} > 0$.*
2. *Let $\epsilon > 0$ and $\beta > 1$ such that $\beta m_\xi \in supp(\xi)$. For $n \geq 0$, the event*

$$E_n := \left(\left\{\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathcal{N}_n\right\|^\alpha \leq \frac{\epsilon}{2\beta\theta^\alpha m_\xi}\right\}\cap\{\xi_n \leq \beta m_\xi\}\right)$$

   verifies $E_n \subset \{\|\mathrm{X}_{n+1}\|^\alpha O_{n+1} \leq \epsilon\}$.
3. *For $n \geq 1$, the events E_n are independent.*

Note that the upper bound θ given in this lemma is represented in Figure 1 for $\alpha = 2$.

Proof 1. For $n \geq 0$, $\mathcal{F}(\mathrm{X}_n) \leq \mathcal{F}(\mathrm{X}_0)$, i.e., $g(\|\mathrm{X}_n\|^\alpha O_n) \leq g(\|\mathrm{X}_0\|^\alpha O_0)$ which implies that $\|\mathrm{X}_n\|^\alpha O_n \leq \|\mathrm{X}_0\|^\alpha O_0$ as g is increasing. Since $O_n \geq m_\xi$, $\|\mathrm{X}_n\|^\alpha m_\xi \leq \|\mathrm{X}_n\|^\alpha O_n \leq \|\mathrm{X}_0\|^\alpha O_0$ which gives that $\|\mathrm{X}_n\| \leq \|\mathrm{X}_0\| \left(\frac{O_0}{m_\xi}\right)^{\frac{1}{\alpha}}$.

2. First, the event E_n is well defined as we can divide by $\|\mathrm{X}_n\|$ thanks to Lemma 4 stating that $\|\mathrm{X}_n\| \neq 0$ almost surely. Let $\epsilon > 0$ and $\beta > 1$ such that $]m_\xi, \beta m_\xi] \subset \mathrm{supp}(\xi)$ (with $\beta m_\xi < M_\xi$ if $M_\xi < +\infty$). For $n \geq 0$, the event E_n implies for the offspring $\tilde{\mathrm{X}}_n = \mathrm{X}_n + \sigma\|\mathrm{X}_n\|\mathcal{N}_n$ created at iteration n,

$$\mathcal{F}(\tilde{\mathrm{X}}_n) = g\left(\|\mathrm{X}_n\|^\alpha \left\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathcal{N}_n \right\|^\alpha \xi_n\right) \leq g\left(\theta^\alpha \frac{\epsilon}{2\beta m_\xi \theta^\alpha} \beta m_\xi\right) .$$

The right hand side of this last term equals $g(\frac{\epsilon}{2})$ such that $\mathcal{F}(\tilde{\mathrm{X}}_n) \leq g\left(\frac{\epsilon}{2}\right) < g(\epsilon)$. If this offspring is accepted then $\mathcal{F}(\mathrm{X}_{n+1}) < g(\epsilon)$, otherwise the fitness is already smaller than $g(\epsilon)$ and we have also $\mathcal{F}(\mathrm{X}_{n+1}) < g(\epsilon)$ which implies that $\|\mathrm{X}_{n+1}\|^\alpha O_{n+1} \leq \epsilon$.

3. For $n \geq 1$, the event E_n is a function of the random variables Y_n, ξ_n, $\|X_0\|, \xi_0$. We have seen in Lemma 7 that $(Y_n)_{n\geq 0}$ are independent and a similar proof leads to the conclusion that $(Y_n, \xi_n)_n$ are independent. Also for $n \geq 1$, Y_n, ξ_n are independent of X_0, ξ_0. Therefore $(E_n)_{n\geq 1}$ are independent. □

Proof of Theorem 1: Let $(U_n)_{n\in\mathbb{Z}^+}$ be the sequence defined for $n \geq 0$ as $U_n := \|\mathrm{X}_n\|^\alpha O_n$. Then, for $n \geq 0$, $\mathcal{F}(\mathrm{X}_n) = g(U_n)$. The sequence $(g(U_n))_{n\in\mathbb{Z}^+}$ is decreasing and lower bounded by $g(0)$ as, for $n \geq 0$, $g(U_n) \geq g(\|\mathrm{X}_n\|^\alpha m_\xi) \geq g(0)$. Therefore, it converges almost surely to a random variable that we denote l and which verifies $l(w) \geq g(0)\ \forall w \in \Omega$. Moreover, the sequence $(U_n)_{n\in\mathbb{Z}^+}$ is positive and is decreasing as the sequence $(g(U_n))_{n\in\mathbb{Z}^+}$ is decreasing and g is an increasing map. Therefore it converges to a positive random variable almost surely. Let us show that the limit of the sequence $(U_n)_{n\in\mathbb{Z}^+}$ is zero. Let $\epsilon > 0$, we have to show that $\exists\ n_0 \geq 0$ such that $U_n \leq \epsilon$ for $n \geq n_0$. Since the sequence $(U_n)_{n\in\mathbb{Z}^+}$ is decreasing, we only have to show that $\exists\ n_0 \geq 0$ such that $U_{n_0} \leq \epsilon$. Let $\beta > 1$ and such that $]m_\xi, \beta m_\xi] \subset \mathrm{supp}(\xi)$. In Lemma 8, we have defined the event E_n, shown that it is included in the event $\{U_{n+1} \leq \epsilon\}$ and proved that the events $(E_n)_{n\geq 1}$ are independent. Moreover, $P(E_n) = P(\|\mathrm{e}_1 + \sigma\mathcal{N}_n\|^\alpha \leq \frac{\epsilon}{2\beta\theta^\alpha m_\xi}) P(\xi_n \leq \beta m_\xi)$ (where θ is the random variable defined in Lemma 8) is a strictly positive con-

stant for all $n \geq 1$. Then $\sum_{n=0}^{+\infty} P(E_n) = +\infty$. This gives by BCL (Lemma 5) that $P(\overline{\lim}\ E_n) = 1$. Therefore $P(\overline{\lim}\ \{U_{n+1} \leq \epsilon\}) = 1$. Since $(U_n)_n$ is decreasing, $\exists n_0$ such that $\forall n \geq n_0,\ U_{n+1} \leq \epsilon$. Therefore the sequence $(U_n)_n$ converges to zero. Consequently, the sequence $(\|\mathrm{X}_n\|)_n$ converges to zero as we have

$$0 \leq m_\xi \|\mathrm{X}_n\|^\alpha \leq U_n = O_n \|\mathrm{X}_n\|^\alpha .$$

Finally, if g is continuous then the limit l of the sequence $g(U_n)$ equals $g(0)$. □

Proof of Lemma 1 (stated page 12)

Let us show that $\exists\ p_0 \geq 0$ such that $\xi_{p_0} < 0$ almost surely. We are going to show this statement by contradiction. Suppose that $\forall p \geq 0, \xi_p \geq 0$. Then, we have $M_n := \frac{1}{n} \sum_{p=0}^{n-1} 1_{\{\xi_p < 0\}} = 0$, for all $n \geq 0$. Therefore M_n goes to zero when n goes to infinity. However, by the law of large numbers (LLN) for independent random variables, M_n converges to $P(\xi_0 < 0)$ when n goes to infinity. As the limit of M_n is unique, this implies that $P(\xi_0 < 0) = 0$ which is not true since $m_\xi < 0$. Consequently, there exits $p_0 \geq 0$ such that $\xi_{p_0} < 0$. Now, we define n_1 as $n_1 := 1 + \min\{p \in \mathbb{Z}^+ \text{ such that } \xi_p < 0\}$. The offspring where the first negative noise has been sampled will be selected since it has a smaller fitness value than all the other individuals (that have a positive fitness value). Therefore $O_{n_1} < 0$ which implies that $\mathcal{F}(\mathrm{X}_{n_1}) = g\left(\|\mathrm{X}_{n_1}\|^\alpha O_{n_1}\right) < g(0)$. Let $n \geq n_1$. The sequence $(\mathcal{F}(\mathrm{X}_n))_n$ is decreasing such that $\mathcal{F}(\mathrm{X}_n) \leq \mathcal{F}(\mathrm{X}_{n_1}) < g(0)$. This implies that $O_n < 0$, otherwise $\mathcal{F}(\mathrm{X}_n) = g(\|\mathrm{X}_n\|^\alpha O_n) \geq g(0)$ which can not hold, i.e. we have shown that after the first individual with a negative fitness has been accepted, all other accepted individuals have a negative fitness.

Let us show now that the sequence $(\|\mathrm{X}_n\|)_{n \geq n_1}$ is lower bounded. Because of the '+' selection, we have:

$$\forall n \geq n_1, \mathcal{F}(\mathrm{X}_n) = g\left(\|\mathrm{X}_n\|^\alpha O_n\right) \leq \mathcal{F}(\mathrm{X}_{n_1}) = g\left(\|\mathrm{X}_{n_1}\|^\alpha O_{n_1}\right) .$$

As the map g is increasing and the noise is lower bounded by m_ξ, we have:

$$\forall n \geq n_1, \|\mathrm{X}_n\|^\alpha m_\xi \leq \|\mathrm{X}_n\|^\alpha O_n \leq \|\mathrm{X}_{n_1}\|^\alpha O_{n_1} ,$$

which gives $\|\mathrm{X}_n\| \geq \|\mathrm{X}_{n_1}\| \left(\frac{|O_{n_1}|}{|m_\xi|}\right)^{\frac{1}{\alpha}}$ for all $n \geq n_1$. □

Proof of Theorem 2 (stated page 12)

The proof of Theorem 2 requires the following lemma.

Lemma 9 *Assume that $m_\xi < 0$. Consider the random variable n_1 and the quantity A defined in Lemma 1. Let $m < \|\mathrm{X}_{n_1}\|^\alpha O_{n_1} < 0$ and $\beta > 1$ such that $\frac{m_\xi}{\beta} \in supp(\xi) \cap \mathbb{R}_-^*$. For $n \geq n_1$, the event F_n defined by*

$$F_n := \left(\left\{ |1 - \sigma\|\mathcal{N}_n\||^\alpha \geq \frac{|m|}{|m_\xi|} \frac{\beta+1}{A^\alpha} \right\} \cap \left\{ \xi_n \leq \frac{m_\xi}{\beta} \right\} \right)$$

verifies $F_n \subset (\|\mathrm{X}_{n+1}\|^\alpha O_{n+1} \leq m)$.

Proof Let (U_n) be the sequence defined as $U_n := \|\mathrm{X}_n\|^\alpha O_n$. By Lemma 1, $\exists\ n_1 \geq 0$, $A > 0$ such that $U_n < 0$ and $\|\mathrm{X}_n\| \geq A\ \forall n \geq n_1$. We consider $n \geq n_1$, then $\|\mathrm{X}_n\| > A$. For all $y \in \mathbb{R}^d, \|y\| \neq 0$, the reverse triangle inequality implies[4]

$$\left\| \frac{y}{\|y\|} + \sigma\mathcal{N}_n \right\| \geq |1 - \sigma\|\mathcal{N}_n\|| \ . \tag{25}$$

Let $\beta > 1$ such that $\frac{m_\xi}{\beta} \in \mathrm{supp}(\xi) \cap \mathbb{R}_-^*$. Suppose that we have $|1-\sigma\|\mathcal{N}_n\||^\alpha \geq \frac{(\beta+1)|m|}{A^\alpha |m_\xi|}$ and $\xi_n \leq \frac{m_\xi}{\beta} < 0$, then with Eq. 25 we have

$$\|\mathrm{X}_n\|^\alpha \left\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathcal{N}_n \right\|^\alpha |\xi_n| \geq \|\mathrm{X}_n\|^\alpha |1-\sigma\|\mathcal{N}_n\||^\alpha |\xi_n| \geq A^\alpha \frac{(\beta+1)|m|}{A^\alpha |m_\xi|} \frac{|m_\xi|}{\beta} > |m| \, .$$

Therefore, the offspring $\tilde{\mathrm{X}}_n$ is such that

$$\mathcal{F}(\tilde{\mathrm{X}}_n) = g\left(\|\mathrm{X}_n\|^\alpha \left\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma\mathcal{N}_n \right\|^\alpha \xi_n \right) \leq g(m)$$

and thus $\mathcal{F}(\mathrm{X}_{n+1}) \leq \mathcal{F}(\tilde{\mathrm{X}}_n) < g(m)$. This implies that $U_{n+1} = \|\mathrm{X}_{n+1}\|^\alpha O_{n+1} \leq m$. Consequently, we have shown that for $n \geq n_0$, the event F_n is included in the event $\{U_{n+1} \leq m\}$. □

[4]The reverse triangle inequality states that for all x, y in $\mathbb{R}^d$, $|\|x\| - \|y\|| \leq \|x - y\|$.

Proof of Theorem 2: Let $n \geq n_1$ (n_1 defined in Lemma 1). We will show that the sequence $(\|\mathrm{X}_n\|)_n$ diverges to $+\infty$. First, we show that $U_n = \|\mathrm{X}_n\|^\alpha O_n$ diverges to $-\infty$. This is equivalent to show that, for any $m < 0$, $\exists\, n \geq n_1$ such that $U_n \leq m$. Similarly to the proof of Theorem 1, the Borel-Cantelli Lemma implies that we have $P(F_n \text{ i.o.}) = 1$ (the event F_n being defined in Lemma 9) therefore Lemma 9 gives that $P(U_{n+1} \leq m \text{ i.o.}) = 1$. Since U_n is decreasing, U_n converges to $-\infty$. For all $n \geq n_1$, $0 \geq O_n \geq m_\xi$, then $m_\xi \|\mathrm{X}_n\|^\alpha \leq U_n = \|\mathrm{X}_n\|^\alpha O_n$ for $n \geq n_1$. Consequently $(\|\mathrm{X}_n\|)_n$ converges to $+\infty$ almost surely. □

Proof of Theorem 3 (stated page 13)

Note that the case $-\infty < m_\xi < 0$ leads to a divergence of the algorithm as already stated in Theorem 2. Now we investigate the more general result where $-\infty \leq m_\xi < 0$. By Lemma 1, $\exists\ n_1 \geq 0$ such that $O_n < 0$ for all $n \geq n_1$ almost surely. For $n \geq 0$, and thanks to Lemma 4 which allows to divide by $\|\mathrm{X}_n\|^\alpha$ almost surely, the acceptance event (see Eq. 7 and Eq. 8) writes as

$$\|\mathrm{X}_n\|^\alpha \left\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathcal{N}_n \right\|^\alpha \xi_n < \|\mathrm{X}_n\|^\alpha O_n \,\text{a.s.}\,.$$

that can be simplified into

$$\left\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathcal{N}_n \right\|^\alpha \xi_n < O_n\,.$$

For $n \geq 0$, we have:

$$\begin{aligned} E\left(\frac{\|\mathrm{X}_{n+1}\|^2}{\|\mathrm{X}_n\|^2} \,|\, \mathrm{X}_n, O_n, \xi_n \right) &= E\left(1_{\{\left\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathcal{N}_n \right\|^\alpha \xi_n > O_n\}} \,|\, \mathrm{X}_n, O_n, \xi_n \right) \\ &+ E\left(\frac{\|\mathrm{X}_n\|^2 \left\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathcal{N}_n \right\|^2}{\|\mathrm{X}_n\|^2} 1_{\{\left\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathcal{N}_n \right\|^\alpha \xi_n < O_n\}} \,|\, \mathrm{X}_n, O_n, \xi_n \right) \end{aligned}$$

As the multivariate normal distribution is isotropic, we get

$$\begin{aligned} E\left(\frac{\|\mathrm{X}_{n+1}\|^2}{\|\mathrm{X}_n\|^2} \,|\, \mathrm{X}_n, O_n, \xi_n \right) &= E\left(1_{\{\|\mathrm{e}_1 + \sigma \mathcal{N}_n\|^\alpha \xi_n > O_n\}} \,|\, O_n, \xi_n \right) \\ &+ E\left(\|\mathrm{e}_1 + \sigma \mathcal{N}_n\|^2 1_{\{\|\mathrm{e}_1 + \sigma \mathcal{N}_n\|^\alpha \xi_n < O_n\}} \,|\, O_n, \xi_n \right) \end{aligned}$$

Let $\mathcal{N}_{n,1}$ denote the first coordinate of the variable $\mathcal{N}_n$. The quantity $\|\mathrm{e}_1 + \sigma\mathcal{N}_n\|^2$ equals $1 + 2\sigma\mathcal{N}_{n,1} + \sigma^2\|\mathcal{N}_n\|^2$ and we have

$$E\left(\frac{\|\mathrm{X}_{n+1}\|^2}{\|\mathrm{X}_n\|^2} \,|\, \mathrm{X}_n, O_n, \xi_n\right) = 1 + \sigma^2 E\left(\|\mathcal{N}_n\|^2 1_{\{\|\mathrm{e}_1+\sigma\mathcal{N}_n\|^\alpha \xi_n < O_n\}} \,|\, O_n, \xi_n\right) + E\left(2\sigma\mathcal{N}_{n,1} 1_{\{\|\mathrm{e}_1+\sigma\mathcal{N}_n\|^\alpha \xi_n < O_n\}} \,|\, O_n, \xi_n\right)$$

In the sequel, we suppose that $n \geq n_1$. Therefore, we have $O_n < 0$. Thus, in the last equation, the event $(\|\mathrm{e}_1 + \sigma\mathcal{N}_n\|^\alpha \xi_n < O_n)$ is equivalent to the event $(\xi_n < 0 \cap \|\mathrm{e}_1 + \sigma\mathcal{N}_n\|^2 > A(O_n, \xi_n))$ where $A(O_n, \xi_n)$ is defined as $A(O_n, \xi_n) := \left(\frac{|O_n|}{|\xi_n|}\right)^{\frac{2}{\alpha}}$. Therefore, we get

$$E\left(\frac{\|\mathrm{X}_{n+1}\|^2}{\|\mathrm{X}_n\|^2} \,|\, \mathrm{X}_n, O_n, \xi_n\right) = 1 + \sigma^2 1_{\{\xi_n<0\}} E\left(\|\mathcal{N}_n\|^2 1_{\{1+2\sigma\mathcal{N}_{n,1}+\sigma^2\|\mathcal{N}_n\|^2 > A(O_n,\xi_n)\}} \,|\, O_n, \xi_n\right) + 2\sigma 1_{\{\xi_n<0\}} E\left(\mathcal{N}_{n,1} 1_{\{1+2\sigma\mathcal{N}_{n,1}+\sigma^2\|\mathcal{N}_n\|^2 > A(O_n,\xi_n)\}} \,|\, O_n, \xi_n\right)$$

Now, we will show that $M(O_n, \xi_n) := E\left(\mathcal{N}_{n,1} 1_{\{\|\mathrm{e}_1+\sigma\mathcal{N}_n\|^2 > A(O_n,\xi_n)\}} \,|\, O_n, \xi_n\right) \geq 0$. The quantity $M(O_n, \xi_n)$ can be rewritten as

$$M(O_n, \xi_n) = \int_{\mathbb{R}^d} x_1 1_{\{\|\mathrm{e}_1+\sigma x\|^2 > A(O_n,\xi_n)\}}(x) p_{\mathcal{N}}(x) dx \,. \tag{26}$$

Let O_n and ξ_n be fixed and let $(x_1, \ldots, x_d) \in \mathbb{R}^d$. If x_1 is such that

$$x_1 < 0 \text{ and } 1 + 2\sigma x_1 + \sigma^2\|x\|^2 > A(O_n, \xi_n)$$

then

$$1 + 2\sigma(-x_1) + \sigma^2\left((x_1)^2 + \sum_{i=2}^{d}(x_i)^2\right) \geq 1 + 2\sigma x_1 + \sigma^2\|x\|^2 > A(O_n, \xi_n)$$

Let $B(O_n, \xi_n, x)$ denote the quantity $\frac{A(O_n,\xi_n)-1-\sigma^2\|x\|^2}{2\sigma}$. Then

$$B(O_n, \xi_n, (x_1, x_2 \ldots, x_d)) = B(O_n, \xi_n, (-x_1, x_2 \ldots, x_d)) \,, \tag{27}$$

and we have

$$\text{if } x_1 < 0 \,\text{then } 1_{\{x_1 > B(O_n,\xi_n,(x_1,x_2,\ldots,x_d))\}} \leq 1_{\{-x_1 > B(O_n,\xi_n,(-x_1,x_2,\ldots,x_d))\}} \,. \tag{28}$$

The quantity $M(O_n, \xi_n)$ can be rewritten as

$$\int_{\mathbb{R}^{d-1}} \left[\int_{\mathbb{R}} x_1 1_{\{x_1 \leq 0\}} 1_{\{\|e_1 + \sigma x\|^2 > A(O_n, \xi_n)\}}(x) p(x_1) dx_1 \right] p(x_2) \dots p(x_d) dx_2 \dots dx_d$$
$$+ \int_{\mathbb{R}^{d-1}} \left[\int_{\mathbb{R}} x_1 1_{\{x_1 \geq 0\}} 1_{\{\|e_1 + \sigma x\|^2 > A(O_n, \xi_n)\}}(x) p(x_1) dx_1 \right] p(x_2) \dots p(x_d) dx_2 \dots dx_d.$$

where $p(y) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{y^2}{2})$ is the density of a normal distribution with mean zero and standard deviation 1 (we have $p_{\mathcal{N}}(x) = p(x_1) \dots p(x_d)$). Applying a change of variables in the second term ($u_1 = -x_1, u_2 = x_2, \dots, u_d = x_d$), and using Eq. 27, one gets that $M(O_n, \xi_n)$ equals

$$\int_{\mathbb{R}^{d-1}} \left[\int_{\mathbb{R}} x_1 1_{\{x_1 \leq 0\}} 1_{\{x_1 > B(O_n, \xi_n, x)\}}(x) p(x_1) dx_1 \right] p(x_2) \dots p(x_d) dx_2 \dots dx_d$$
$$+ \int_{\mathbb{R}^{d-1}} \left[\int_{\mathbb{R}} -u_1 1_{\{u_1 \leq 0\}} 1_{\{-u_1 > B(O_n, \xi_n, u)\}}(u) p(u_1) du_1 \right] p(u_2) \dots p(u_d) du_2 \dots du_d \,.$$

This gives $M(O_n, \xi_n) =$

$$\int_{\mathbb{R}^{d-1}} \left[\int_{\mathbb{R}} x_1 1_{\{x_1 \leq 0\}} \left(1_{\{x_1 > B(O_n, \xi_n, x)\}}(x) - 1_{\{-x_1 > B(O_n, \xi_n, x)\}}(x) \right) p(x_1) dx_1 \right]$$
$$p(x_2) \dots p(x_d) dx_2 \dots dx_d \,.$$

By Eq. 28, one has $x_1 1_{\{x_1 \leq 0\}} \left(1_{\{x_1 > B(O_n, \xi_n, x)\}}(x) - 1_{\{-x_1 > B(O_n, \xi_n, x)\}}(x) \right) \geq 0$ for all $x \in \mathbb{R}^d$. Consequently $M(O_n, \xi_n) \geq 0$ which implies that

$$E\left(\frac{\|\mathrm{X}_{n+1}\|^2}{\|\mathrm{X}_n\|^2} \,|\, \mathrm{X}_n, O_n, \xi_n \right) \geq 1 \;\forall\; n \geq n_1 \,.$$

□

Proof of Lemma 2 (stated page 14)

Taking the norm in Eq. 7, we have for $k \geq 0$

$$\|\mathrm{X}_{k+1}\| = \|\mathrm{X}_k + \sigma \|\mathrm{X}_k\| \mathcal{N}_n 1_{\{\|\mathrm{X}_k + \sigma \|\mathrm{X}_k\| \mathcal{N}_k\|^\alpha \xi_k < O_n \|\mathrm{X}_k\|^\alpha\}}\|$$

Lemma 4 states that $k \geq 0$, $\|\mathrm{X}_k\| \neq 0$ almost surely. Then the previous equation can be rewritten as

$$\|\mathrm{X}_{k+1}\| = \|\mathrm{X}_k\| \left\| \frac{\mathrm{X}_k}{\|\mathrm{X}_k\|} + \sigma \mathcal{N}_k 1_{\left\{ \left\| \frac{\mathrm{X}_k}{\|\mathrm{X}_k\|} + \sigma \mathcal{N}_k \right\|^{\alpha} \xi_k < O_k \right\}} \right\| \text{ a.s.}$$

Taking the logarithm of the previous equation, one has for $k \geq 0$

$$\ln(\|\mathrm{X}_{k+1}\|) = \ln(\|\mathrm{X}_k\|) + \ln \left(\left\| \frac{\mathrm{X}_k}{\|\mathrm{X}_k\|} + \sigma \mathcal{N}_k 1_{\left\{ \left\| \frac{\mathrm{X}_k}{\|\mathrm{X}_k\|} + \sigma \mathcal{N}_k \right\|^{\alpha} \xi_k < O_k \right\}} \right\| \right) \text{ a.s.} \tag{29}$$

For $n \geq 1$, we sum the equations (29) from 0 to $n-1$ and divide by n, one gets

$$\begin{aligned} \frac{1}{n} \ln \left(\frac{\|\mathrm{X}_n\|}{\|\mathrm{X}_0\|} \right) &= \frac{1}{n} \sum_{k=0}^{n-1} \ln \left(\frac{\|\mathrm{X}_{k+1}\|}{\|\mathrm{X}_k\|} \right) \\ &= \frac{1}{n} \sum_{k=0}^{n-1} \ln \left(\left\| \frac{\mathrm{X}_k}{\|\mathrm{X}_k\|} + \sigma \mathcal{N}_k 1_{\left\{ \left\| \frac{\mathrm{X}_k}{\|\mathrm{X}_k\|} + \sigma \mathcal{N}_k \right\|^{\alpha} \xi_k < O_k \right\}} \right\| \right) . \end{aligned}$$

□

Proof of Proposition 1 (stated page 15)

Let $(\Phi_n)_n$ be a sequence of random variables and $(t_n)_n$ a filtration adapted to the sequence[5]. Let us remind that $(\Phi_n)_n$ is a Markov chain with transition kernel Q and initial law $\mathcal{L}$ if (1) for all n, Φ_n is t_n measurable; (2) the random variable Φ_0 follows the law $\mathcal{L}$; (3) for all measurable and bounded function f, $E(f(\Phi_{n+1})|t_n) = Qf(\Phi_n)$.

Let us take the filtration $t_n = \boldsymbol{\sigma}(X_0, \mathcal{N}_0, \xi_0, \ldots, \mathcal{N}_{n-1}, \xi_{n-1})$. Then both O_n and Z_n are t_n-measurable. Also O_0 and Z_0 are distributed as $\mathcal{L}_\xi$ such that (1) and (2) are satisfied.

Let f be a measurable and bounded function on $\mathbb{R}$ we want to show that $E(f(O_{n+1})|t_n) = Pf(O_n)$ and $E(f(Z_{n+1})|t_n) = Pf(Z_n)$ where P is the transition kernel given in Proposition 1. Let f be a measurable function, then

$$E(f(O_{n+1})|t_n) = E(f((\xi_n - O_n)1_{\{\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma \mathcal{N}_n \|^{\alpha} \xi_n < O_n\}} + O_n)|t_n)$$

[5] t_n are σ-algebra such that $t_n \subset t_{n+1}$ and ϕ_n is t_n-measurable.

Since $\mathcal{N}_n$ and ξ_n are independent, the previous equation writes

$$E(f(O_{n+1})|t_n) = S(\mathrm{X}_n/\|\mathrm{X}_n\|, O_n)$$

where $S(u,v) = E(f((\xi_n - v)1_{\{\|u+\sigma\mathcal{N}_n\|^\alpha \xi_n < v\}} + v))$. Using now the isotropy of $\mathcal{N}_n$, we have $S(u,v) = S(\mathrm{e}_1, v)$ for all v (same proof as Lemma 6). Let define $Q(v) = S(\mathrm{e}_1, v) = E(f((\xi_n - v)1_{\{\|\mathrm{e}_1+\sigma\mathcal{N}_n\|^\alpha \xi_n < v\}} + v))$. We then have $E(f(O_{n+1})|t_n) = Q(O_n)$. Using the same argument we immediately obtain $E(f(Z_{n+1})|t_n) = Q(Z_n)$. The function Q can be simplified into

$$Q(v) = E(f(\xi_n)1_{\{\|\mathrm{e}_1+\sigma\mathcal{N}_n\|^\alpha \xi_n < v\}}) + f(v)E(1_{\{\|\mathrm{e}_1+\sigma\mathcal{N}_n\|^\alpha \xi_n \geq v\}})$$

It remains now to compute $Pf(x)$ where P is given in Proposition 1. By definition, $Pf(x) = \int f(x')P(x,dx') = \int f(x')P_1(x,dx') + f(x)E(1_{\|\mathrm{e}_1+\sigma\mathcal{N}_0\|^\alpha \xi_0 \geq x})$. Besides, $P_1(x,A) = E(1_A(\xi_0)1_{\{\|\mathrm{e}_1+\sigma\mathcal{N}_0\|^\alpha \xi_0 < x\}})$ and thus $\int f(x')P_1(x,dx') = E(f(\xi_0)1_{\{\|\mathrm{e}_1+\sigma\mathcal{N}_0\|^\alpha \xi_0 < x\}})$. Thus $Q(v) = Pf(v)$ which achieves the proof of (3). □

Proof of Lemma 3 (stated page 16)

Step 1: Let us define the sequences $(\widetilde{\mathrm{X}_n})_n$ and $(\widetilde{O_n})_n$ in the following way:

$$\widetilde{\mathrm{X}_{n+1}} = \widetilde{\mathrm{X}_n} + \sigma\|\widetilde{\mathrm{X}_n}\| M_n(\widetilde{\mathrm{X}_n})\mathcal{N}_n 1_{\{\|\frac{\widetilde{\mathrm{X}_n}}{\|\widetilde{\mathrm{X}_n}\|} + \sigma M(\widetilde{\mathrm{X}_n})\mathcal{N}_n\|^\alpha \xi_n < \widetilde{O_n}\}} \tag{30}$$

with $\widetilde{\mathrm{X}_0} = \mathrm{X}_0$ and where $M_n(\widetilde{\mathrm{X}_n})$ is an orthogonal matrix that sends e_1 on $\frac{\widetilde{\mathrm{X}_n}}{\|\widetilde{\mathrm{X}_n}\|}$. In the same way,

$$\widetilde{O_{n+1}} := (\xi_n - \widetilde{O_n})1_{\{\|\frac{\widetilde{\mathrm{X}_n}}{\|\widetilde{\mathrm{X}_n}\|} + \sigma M(\widetilde{\mathrm{X}_n})\mathcal{N}_n\|^\alpha \xi_n < \widetilde{O_n}\}} + \widetilde{O_n} \tag{31}$$

with $\widetilde{O_0} = O_0$. We are now going to prove

(i) for all n, $(\widetilde{\mathrm{X}_n}, \widetilde{O_n})$ and (X_n, O_n) have the same law;

(ii) for all n, $\widetilde{\mathrm{X}_n}$ and X_n have the same law;

(iii) for all n, $Z_n = \widetilde{O_n}$ almost surely.

Proof of (i): We will proceed by induction. Since $(\widetilde{\mathrm{X}_0}, \widetilde{O_0}) = (\mathrm{X}_0, O_0)$, (i) is true for $n = 0$. Assume now that $(\widetilde{\mathrm{X}_n}, \widetilde{O_n})$ and (X_n, O_n) have the same distribution. Let us prove that $(\widetilde{\mathrm{X}_{n+1}}, \widetilde{O_{n+1}})$ and $(\mathrm{X}_{n+1}, O_{n+1})$ have the same distribution, i.e. let us prove that for $t \in \mathbb{R}^d, t' \in \mathbb{R}$, $E(e^{it.\widetilde{\mathrm{X}_{n+1}}} e^{it'\widetilde{O_{n+1}}}) = E(e^{it.\mathrm{X}_{n+1}} e^{it'O_{n+1}})$ where the . denotes the usual scalar product in $\mathbb{R}^d$. According to Eq. 30, and Eq. 31

$$E(e^{it.\widetilde{\mathrm{X}_{n+1}}} e^{it'\widetilde{O_{n+1}}}) = E\left(e^{it.(\widetilde{\mathrm{X}_n} + \sigma \|\widetilde{\mathrm{X}_n}\| M_n(\widetilde{\mathrm{X}_n}) \mathcal{N}_n 1\{\|\frac{\widetilde{\mathrm{X}_n}}{\|\widetilde{\mathrm{X}_n}\|} + \sigma M(\widetilde{\mathrm{X}_n})\mathcal{N}_n\|^\alpha \xi_n < \widetilde{O_n}\})} \right.$$
$$\left. e^{it'\widetilde{O_n} + (\xi_n - \widetilde{O_n}) 1\{\|\frac{\widetilde{\mathrm{X}_n}}{\|\widetilde{\mathrm{X}_n}\|} + \sigma M(\widetilde{\mathrm{X}_n})\mathcal{N}_n\|^\alpha \xi_n < \widetilde{O_n}\}} \right) \quad (32)$$

Using $t_n = \boldsymbol{\sigma}(X_0, \mathcal{N}_0, \xi_0, \ldots, \mathcal{N}_{n-1}, \xi_{n-1})$,

$$E(e^{it.\widetilde{\mathrm{X}_{n+1}}} e^{it'\widetilde{O_{n+1}}}) = E\left(E\left(e^{it.(\widetilde{\mathrm{X}_n} + \sigma \|\widetilde{\mathrm{X}_n}\| M_n(\widetilde{\mathrm{X}_n}) \mathcal{N}_n 1\{\|\frac{\widetilde{\mathrm{X}_n}}{\|\widetilde{\mathrm{X}_n}\|} + \sigma M(\widetilde{\mathrm{X}_n})\mathcal{N}_n\|^\alpha \xi_n < \widetilde{O_n}\})} \right.\right.$$
$$\left.\left. e^{it'\widetilde{O_n} + (\xi_n - \widetilde{O_n}) 1\{\|\frac{\widetilde{\mathrm{X}_n}}{\|\widetilde{\mathrm{X}_n}\|} + \sigma M(\widetilde{\mathrm{X}_n})\mathcal{N}_n\|^\alpha \xi_n < \widetilde{O_n}\}} | t_n \right) \right) \quad (33)$$

Since $(\widetilde{\mathrm{X}_n}, \widetilde{O_n})$ is t_n-measurable, the right-hand side of the previous equation equals

$$E\left(e^{it.\widetilde{\mathrm{X}_n}} e^{it'\widetilde{O_n}} E\left(e^{it.(\sigma \|\widetilde{\mathrm{X}_n}\| M_n(\widetilde{\mathrm{X}_n}) \mathcal{N}_n 1\{\|\frac{\widetilde{\mathrm{X}_n}}{\|\widetilde{\mathrm{X}_n}\|} + \sigma M(\widetilde{\mathrm{X}_n})\mathcal{N}_n\|^\alpha \xi_n < \widetilde{O_n}\})} \right.\right.$$
$$\left.\left. e^{it'(\xi_n - \widetilde{O_n}) 1\{\|\frac{\widetilde{\mathrm{X}_n}}{\|\widetilde{\mathrm{X}_n}\|} + \sigma M(\widetilde{\mathrm{X}_n})\mathcal{N}_n\|^\alpha \xi_n < \widetilde{O_n}\}} | t_n \right) \right) \quad (34)$$

Since ξ_n and $\mathcal{N}_n$ are independent of t_n, we have that

$$E\left(e^{it.(\sigma \|\widetilde{\mathrm{X}_n}\| M_n(\widetilde{\mathrm{X}_n}) \mathcal{N}_n 1\{\|\frac{\widetilde{\mathrm{X}_n}}{\|\widetilde{\mathrm{X}_n}\|} + \sigma M(\widetilde{\mathrm{X}_n})\mathcal{N}_n\|^\alpha \xi_n < \widetilde{O_n}\})} \right.$$
$$\left. e^{it'(\xi_n - \widetilde{O_n}) 1\{\|\frac{\widetilde{\mathrm{X}_n}}{\|\widetilde{\mathrm{X}_n}\|} + \sigma M(\widetilde{\mathrm{X}_n})\mathcal{N}_n\|^\alpha \xi_n < \widetilde{O_n}\}} | t_n \right) = \gamma_0(\widetilde{\mathrm{X}_n}, M(\widetilde{\mathrm{X}_n}), \widetilde{O_n}) \quad (35)$$

with $\gamma_0(u, A, o) = E\left(e^{it.(\sigma \|u\| A\mathcal{N}_n 1\{\|\frac{u}{\|u\|} + \sigma A\mathcal{N}_n\|^\alpha \xi_n < o\})} e^{it'(\xi_n - o) 1\{\|\frac{u}{\|u\|} + \sigma A\mathcal{N}_n\|^\alpha \xi_n < o\}} \right)$. By the isotropy of the distribution of $\mathcal{N}_n$, $\gamma_0(u, A, o) = \gamma_0(u, I_d, o)$ for all

$u \in \mathbb{R}^d$, $o \in \mathbb{R}$ and any orthogonal matrix A and thus $\gamma_0(\widetilde{\mathrm{X}_n}, M(\widetilde{\mathrm{X}_n}), \widetilde{O_n}) = \gamma_0(\widetilde{\mathrm{X}_n}, I_d, \widetilde{O_n})$. We have thus proven that

$$E(e^{it.\widetilde{\mathrm{X}_{n+1}}} e^{it'\widetilde{O_{n+1}}}) = E(e^{it.\widetilde{\mathrm{X}_n}} e^{it'\widetilde{O_n}} \gamma_0(\widetilde{\mathrm{X}_n}, I_d, \widetilde{O_n}))$$

Using the same fastidious technique we can show that

$$E(e^{it.\mathrm{X}_{n+1}} e^{it'O_{n+1}}) = E(e^{it.\mathrm{X}_n} e^{it'O_n} \gamma_0(\mathrm{X}_n, I_d, O_n)) \ .$$

By induction hypothesis, (X_n, O_n) and $(\widetilde{\mathrm{X}_n}, \widetilde{O_n})$ have the same law and therefore $E(e^{it.\mathrm{X}_n} e^{it'O_n} \gamma_0(\mathrm{X}_n, I_d, O_n)) = E(e^{it.\widetilde{\mathrm{X}_n}} e^{it'\widetilde{O_n}} \gamma_0(\widetilde{\mathrm{X}_n}, I_d, \widetilde{O_n}))$ which in turn implies that

$$E(e^{it.\widetilde{\mathrm{X}_{n+1}}} e^{it'\widetilde{O_{n+1}}}) = E(e^{it.\mathrm{X}_{n+1}} e^{it'O_{n+1}}) \ .$$

Proof of (ii): We will show that $E(e^{it.\mathrm{X}_{n+1}}) = E(e^{it.\widetilde{\mathrm{X}_{n+1}}})$.

$$E(e^{it.\widetilde{\mathrm{X}_{n+1}}}) = E(E(e^{it(\widetilde{\mathrm{X}_n} + \sigma\|\widetilde{\mathrm{X}_n}\| M_n(\widetilde{\mathrm{X}_n})\mathcal{N}_n 1\{\|\frac{\widetilde{\mathrm{X}_n}}{\|\widetilde{\mathrm{X}_n}\|} + \sigma M(\widetilde{\mathrm{X}_n})\mathcal{N}_n\|^\alpha \xi_n < \widetilde{O_n}\})} | t_n))$$

Since $\widetilde{\mathrm{X}_n}$ is t_n-measurable,

$$E(e^{it.\widetilde{\mathrm{X}_{n+1}}}) = E(e^{it.\widetilde{\mathrm{X}_n}} E(e^{it.(\sigma\|\widetilde{\mathrm{X}_n}\| M_n(\widetilde{\mathrm{X}_n})\mathcal{N}_n 1\{\|\frac{\widetilde{\mathrm{X}_n}}{\|\widetilde{\mathrm{X}_n}\|} + \sigma M(\widetilde{\mathrm{X}_n})\mathcal{N}_n\|^\alpha \xi_n < \widetilde{O_n}\})} | t_n)) \tag{36}$$

Since ξ_n and $\mathcal{N}_n$ are independent of t_n, we have that

$$E(e^{it.(\sigma\|\widetilde{\mathrm{X}_n}\| M_n(\widetilde{\mathrm{X}_n})\mathcal{N}_n 1\{\|\frac{\widetilde{\mathrm{X}_n}}{\|\widetilde{\mathrm{X}_n}\|} + \sigma M(\widetilde{\mathrm{X}_n})\mathcal{N}_n\|^\alpha \xi_n < \widetilde{O_n}\})} | t_n) = \gamma(\widetilde{\mathrm{X}_n}, M(\widetilde{\mathrm{X}_n}), \widetilde{O_n})$$

where $\gamma(u, A, o) = E(e^{it.(\sigma\|u\| A\mathcal{N}_n 1\{\|\frac{u}{\|u\|} + \sigma A\mathcal{N}_n\|^\alpha \xi_n < o\})})$. By the isotropy of the distribution of $\mathcal{N}_n$, $\gamma(u, A, o) = \gamma(u, I_d, o)$ for all $u \in \mathbb{R}^d$, $o \in \mathbb{R}$ and any orthogonal matrix A and thus $\gamma(\widetilde{\mathrm{X}_n}, M(\widetilde{\mathrm{X}_n}), \widetilde{O_n}) = \gamma(\widetilde{\mathrm{X}_n}, I_d, \widetilde{O_n})$. Injecting this in Eq. 36 we obtain that

$$E(e^{it.\widetilde{\mathrm{X}_{n+1}}}) = E(e^{it.\widetilde{\mathrm{X}_n}} \gamma(\widetilde{\mathrm{X}_n}, I_d, \widetilde{O_n})) \ . \tag{37}$$

Using (i) in the previous equation we obtain

$$E(e^{it.\widetilde{\mathrm{X}_{n+1}}}) = E(e^{it.\mathrm{X}_n} \gamma(\mathrm{X}_n, I_d, O_n)) \ . \tag{38}$$

We will now compute $E(e^{it.\mathrm{X}_{n+1}})$. From the definition of X_n we have

$$E(e^{it.\mathrm{X}_{n+1}}) = E(e^{it.(\mathrm{X}_n+\sigma\|\mathrm{X}_n\|\mathcal{N}_n 1\{\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathcal{N}_n\|^\alpha\xi_n<O_n\})})$$

and conditioning with respect to t_n we obtain

$$E(e^{it.\mathrm{X}_{n+1}}) = E(E(e^{it.(\mathrm{X}_n+\sigma\|\mathrm{X}_n\|\mathcal{N}_n 1\{\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathcal{N}_n\|^\alpha\xi_n<O_n\})}|t_n)) \quad (39)$$

$$= E(e^{it.\mathrm{X}_n}E(e^{it.(\sigma\|\mathrm{X}_n\|\mathcal{N}_n 1\{\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathcal{N}_n\|^\alpha\xi_n<O_n\})}|t_n)) \quad (40)$$

Moreover, since $\mathcal{N}_n$ and ξ_n are independent of t_n we obtain

$$E(e^{it.(\sigma\|\mathrm{X}_n\|\mathcal{N}_n 1\{\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma\mathcal{N}_n\|^\alpha\xi_n<O_n\})}|t_n)) = \gamma(\mathrm{X}_n, I_d, O_n)$$

and thus

$$E(e^{it.\mathrm{X}_{n+1}}) = E(e^{it.\mathrm{X}_n}\gamma(\mathrm{X}_n, I_d, O_n)) \quad (41)$$

From Eq. 41 and Eq. 38, we thus have that $E(e^{it.\widetilde{\mathrm{X}_{n+1}}}) = E(e^{it.\mathrm{X}_{n+1}})$.
Proof of (iii): We will show by recurrence that $Z_n = \widetilde{O_n}$ almost surely. Since $\widetilde{O_0} = O_0 = Z_0$ (iii) is true for $n = 0$. We assume now that $Z_n = \widetilde{O_n}$ almost surely. Let us simplify the notation for $M(\widetilde{\mathrm{X}_n})$ that we now write M_n. Since M_n is an orthogonal matrix[6] $\|\frac{\widetilde{\mathrm{X}_n}}{\|\widetilde{\mathrm{X}_n}\|} + \sigma M_n\mathcal{N}_n\| = \|M_n^T(\frac{\widetilde{\mathrm{X}_n}}{\|\widetilde{\mathrm{X}_n}\|} + \sigma M_n\mathcal{N}_n)\|$ $= \|M_n^T\frac{\widetilde{\mathrm{X}_n}}{\|\widetilde{\mathrm{X}_n}\|} + \sigma M_n^T M_n\mathcal{N}_n\| = \|M_n^T\frac{\widetilde{\mathrm{X}_n}}{\|\widetilde{\mathrm{X}_n}\|} + \sigma\mathcal{N}_n\|$. Since M_n is orthogonal $M_n^T = M_n^{-1}$ and since $M_n\mathrm{e}_1 = \widetilde{\mathrm{X}_n}/\|\widetilde{\mathrm{X}_n}\|$, we have that $M_n^T\frac{\widetilde{\mathrm{X}_n}}{\|\widetilde{\mathrm{X}_n}\|} = \mathrm{e}_1$. We can thus simplify $\|M_n^T\frac{\widetilde{\mathrm{X}_n}}{\|\widetilde{\mathrm{X}_n}\|} + \sigma\mathcal{N}_n\|$ into $\|\mathrm{e}_1 + \sigma\mathcal{N}_n\|$. Therefore

$$\widetilde{O_{n+1}} = (\xi_n - \widetilde{O_n})1_{\{\|\mathrm{e}_1+\sigma\mathcal{N}_n\|^\alpha\xi_n<\widetilde{O_n}\}} + \widetilde{O_n}$$

and thus by induction hypothesis

$$\widetilde{O_{n+1}} = (\xi_n - Z_n)1_{\{\|\mathrm{e}_1+\sigma\mathcal{N}_n\|^\alpha\xi_n<Z_n\}} + Z_n$$

which in turn imply that $\widetilde{O_{n+1}} = Z_{n+1}$ almost surely.
Step 2: Applying Lemma 2 to the sequence $(\widetilde{\mathrm{X}_n})_n$ we obtain

$$\frac{1}{n}\ln\frac{\|\widetilde{\mathrm{X}_n}\|}{\|\widetilde{\mathrm{X}_0}\|} = \frac{1}{n}\sum_{k=0}^{n-1}\ln\left(\left\|\frac{\widetilde{\mathrm{X}_k}}{\|\widetilde{\mathrm{X}_k}\|} + \sigma M_k\mathcal{N}_k 1_{\left\{\left\|\frac{\mathrm{X}_k}{\|\mathrm{X}_k\|}+\sigma M_k\mathcal{N}_k\right\|^\alpha\xi_k<\widetilde{O_k}\right\}}\right\|\right) \quad (42)$$

[6] An orthogonal matrix M satisfies $M^TM = I_d$ and thus $M^{-1} = M^T$. Moreover for all $x \in \mathbb{R}^d$, $\|Mx\| = \|x\|$.

where we have dropped the dependence in $\widetilde{\mathrm{X}_k}$ in the matrix M_k for notation convenience. As for (iii) we have that $\|\frac{\widetilde{\mathrm{X}_k}}{\|\widetilde{\mathrm{X}_k}\|} + \sigma M_k \mathcal{N}_k\| = \|M_k^T(\frac{\widetilde{\mathrm{X}_k}}{\|\widetilde{\mathrm{X}_k}\|} + \sigma M_k \mathcal{N}_k)\| = \|\mathrm{e}_1 + \sigma \mathcal{N}_k\|$, and thus Eq. 42 becomes

$$\frac{1}{n} \ln \frac{\|\widetilde{\mathrm{X}_n}\|}{\|\widetilde{\mathrm{X}_0}\|} = \frac{1}{n} \sum_{k=0}^{n-1} \ln \|\mathrm{e}_1 + \sigma \mathcal{N}_k 1_{\left\{\|\mathrm{e}_1 + \sigma \mathcal{N}_k\|^\alpha \xi_k < \widetilde{O_k}\right\}}\|$$

almost surely. Since by Step 1 (iii), we have that $\widetilde{O_k} = Z_k$ almost surely, we obtain that

$$\frac{1}{n} \ln \frac{\|\widetilde{\mathrm{X}_n}\|}{\|\widetilde{\mathrm{X}_0}\|} = \frac{1}{n} \sum_{k=0}^{n-1} \ln \|\mathrm{e}_1 + \sigma \mathcal{N}_k 1_{\{\|\mathrm{e}_1 + \sigma \mathcal{N}_k\|^\alpha \xi_k < Z_k\}}\| \tag{43}$$

almost surely. By Step 1 (ii), we have that $\|\widetilde{\mathrm{X}_n}\|$ and $\|\mathrm{X}_n\|$ follow the same law and thus

$$\frac{1}{n} \ln \frac{\|\mathrm{X}_n\|}{\|\mathrm{X}_0\|} = \frac{1}{n} \sum_{k=0}^{n-1} \ln \|\mathrm{e}_1 + \sigma \mathcal{N}_k 1_{\{\|\mathrm{e}_1 + \sigma \mathcal{N}_k\|^\alpha \xi_k < Z_k\}}\| \tag{44}$$

holds in distribution. □

Proof of Proposition 2 (stated page 18)

Let us show that $\nu : \mathfrak{B}(]m_\xi, M_\xi[) \mapsto \mathbb{R}^+ \cup \{+\infty\}$ defined as

$$\nu(A) = \int_{\mathbb{R}^d} \int_{m_\xi}^{M_\xi} 1_A(u) 1_{\{\|\mathrm{e}_1 + \sigma t\|^\alpha u < m_\xi\}}(u,t) p_{\mathcal{N}}(t) p_\xi(u) \mathrm{d}u \mathrm{d}t$$

is a finite measure. First, we have $\nu(\emptyset) = 0$. Second, if E_1 and E_2 are two disjoint sets then $\nu(E_1 \cup E_2) = \nu(E_1) + \nu(E_2)$ as the function $1_{E_1 \cup E_2}$ is identically equal to $1_{E_1} + 1_{E_2}$ when $E_1 \cap E_2 = \emptyset$. Third,

$$\nu(]m_\xi, M_\xi[) = \int_{\mathbb{R}^d} \int_{m_\xi}^{M_\xi} 1_{\{\|\mathrm{e}_1 + \sigma t\|^\alpha u < m_\xi\}}(u,t) p_{\mathcal{N}}(t) p_\xi(u) \mathrm{d}u \mathrm{d}t \leq 1 \,.$$

Now, if $m_\xi = 0$, the indicator function $1_{\{\|\mathrm{e}_1 + \sigma t\|^\alpha u < m_\xi\}}(u,t)$ equals zero for any $t \in \mathbb{R}^d$ and $u \in]0, M_\xi[$ almost surely. Therefore, ν is identically equal to

zero. However, if $m_\xi \neq 0$, then, for $A \in \mathfrak{B}(]m_\xi, M_\xi[)$ with a strictly positive Lebesgue measure, the set

$$\mathcal{A} := \{(u,t) \in (]m_\xi, M_\xi[\cap A) \times \mathbb{R}^d \text{ such that } \|\mathrm{e}_1 + \sigma t\|^\alpha u < m_\xi\}$$

has a strictly positive measure with respect to the Lebesgue measure defined on $\mathfrak{B}(\mathbb{R}^d \times]m_\xi, M_\xi[)$. This implies that ν is not identically equal to zero if and only if $m_\xi \neq 0$. Moreover, for $t \in \mathbb{R}^d$, $(u,z) \in]m_\xi, M_\xi[^2$

$$\|\mathrm{e}_1 + \sigma t\|^\alpha u < m_\xi \Rightarrow \|\mathrm{e}_1 + \sigma t\|^\alpha u < z$$

which gives that $\forall z \in]m_\xi, M_\xi[, \forall A \in \mathfrak{B}(]m_\xi, M_\xi[), P_1(z,A) \geq \nu(A)$. □

Proof of Corollary 1 (stated page 19)

By Proposition 2, the condition of Theorem 5 is satisfied for $(Z_n)_n$ and thus it is positive and Harris recurrent. □

Proof of Corollary 2 (stated page 19)

The product measure $\nu \otimes \nu_{\mathcal{N}} \otimes \nu_\xi$ where ν is given in Proposition 2 is a minorization measure for the Markov chain $(Z_n, \mathcal{N}_n, \xi_n)_n$ and the product measure $\mu \otimes \nu_{\mathcal{N}} \otimes \nu_\xi$ is an invariant probability measure for $(Z_n, \mathcal{N}_n, \xi_n)_n$. Therefore $(Z_n, \mathcal{N}_n, \xi_n)_n$ is positive Harris recurrent. □

Proof of Theorem 6 (stated page 19)

To establish the proof of Theorem 6, we need the following two lemma.

Lemma 10 *Suppose that $m_\xi \neq 0$. The quantity γ defined as*

$$\gamma := \int E\left(\ln \|\mathrm{e}_1 + \sigma \mathcal{N}_0 1_{\{\|\mathrm{e}_1 + \sigma \mathcal{N}_0\|^\alpha \xi_0 \leq z\}}\|\right) d\mu(z) \tag{45}$$

where μ is the invariant probability measure of $(Z_n)_n$ is well defined and finite.

Proof Let $g : \mathbb{R}^d \times \mathbb{R}_+^* \times \mathbb{R} \times \mathbb{R} \mapsto \mathbb{R}$ be defined for (x, σ, y, z) in $\mathbb{R}^d \times \mathbb{R}_+^* \times \mathbb{R} \times \mathbb{R}$ by

$$g(x, \sigma, y, z) = \|\mathrm{e}_1 + 1_{\{\|\mathrm{e}_1 + \sigma x\|^\alpha y < z\}}(x, y, z)\sigma x\| .$$

Then the quantity γ corresponds to integrating the function $\ln(g)$ with respect to the variables x, y and z. We notice that

$$g((x_1, x_2, \ldots, x_d), \sigma, y, z) = g((x_1, \epsilon_2 x_2, \ldots, \epsilon_d x_d), \sigma, y, z)$$

for all $(\epsilon_2, \ldots, \epsilon_d)$ in $\{-1, +1\}^{d-1}$ and $(x_1, x_2, \ldots, x_d)$ in $\mathbb{R}^d$. Therefore, we can restrict the integration with respect to the variable x to the domain $\mathcal{D} := \mathbb{R}^* \times]0, +\infty[^{d-1}$, more precisely the quantity γ can be rewritten as

$$\gamma = \frac{1}{(2\pi)^{d/2}} \int_{\mathcal{D}} \int_{m_\xi}^{M_\xi} \int_{m_\xi}^{M_\xi} \ln\left(g(x, \sigma, y, z)\right) e^{-\frac{\|x\|^2}{2}} p_\xi(y) dx dy d\mu(z) .$$

where μ is the invariant probability measure of the Markov chain $(Z_n)_n$. We introduce γ^+ as:

$$\gamma^+ = \frac{1}{(2\pi)^{d/2}} \int_{\mathcal{D}} \int_{m_\xi}^{M_\xi} \int_{m_\xi}^{M_\xi} \ln^+\left[g(x, \sigma, y, z)\right] e^{-\frac{\|x\|^2}{2}} p_\xi(y) dx dy d\mu(z) \tag{46}$$

and γ^- as:

$$\gamma^- = \frac{1}{(2\pi)^{d/2}} \int_{\mathcal{D}} \int_{m_\xi}^{M_\xi} \int_{m_\xi}^{M_\xi} \ln^-\left[g(x, \sigma, y, z)\right] e^{-\frac{\|x\|^2}{2}} p_\xi(y) dx dy d\mu(z) \tag{47}$$

such that $\gamma = \gamma^+ - \gamma^-$. The quantities γ^+ and γ^- are well defined but could be infinite. Using spherical coordinates (with $d \geq 2$) we obtain after partial integration

$$\gamma^- = \left(\frac{1}{2}\right)^{\frac{d}{2}} \frac{1}{W_{d-2}\Gamma\left(\frac{d}{2}\right)} \int_0^{+\infty} \int_0^{\frac{\pi}{2}} \int_{m_\xi}^{M_\xi} \int_{m_\xi}^{M_\xi} \ln^-\left[h(r, \theta, \sigma, y, z)\right] r^{d-1} e^{-\frac{r^2}{2}} \sin^{d-2}(\theta) p_\xi(y)\, dr\, d\theta dy d\mu(z) , \tag{48}$$

and

$$\gamma^+ = \left(\frac{1}{2}\right)^{\frac{d}{2}} \frac{1}{W_{d-2}\Gamma\left(\frac{d}{2}\right)} \int_0^{+\infty} \int_0^{\pi} \int_{m_\xi}^{M_\xi} \int_{m_\xi}^{M_\xi} \ln^+\left[h(r, \theta, \sigma, y, z)\right] r^{d-1} e^{-\frac{r^2}{2}} \sin^{d-2}(\theta) p_\xi(y)\, dr\, d\theta dy d\mu(z) , \tag{49}$$

where we have used the classical Wallis integral $W_{d-2} = \int_0^{\pi/2} \sin^{d-2}\theta \, d\theta$ and the surface area of the d-dimensional unit ball $S_d = 2\pi^{d/2}/\Gamma(\frac{d}{2})$ where $\Gamma()$ denotes the Gamma function defined as $\Gamma(z) = \int_0^\infty t^{z-1}e^{-t}\,dt$ and where h is the positive function defined on $\mathbb{R}^+ \times [0,\pi] \times \mathbb{R}_+^* \times \mathbb{R} \times \mathbb{R}$ by

$$h(r,\theta,\sigma,y,z) = \|1_{\{\|\sigma r - e^{i\theta}\|^\alpha y < z\}}(r,\theta,y,z)\sigma r - e^{i\theta}\| \,.$$

For (r,θ,σ,y,z) in $\mathbb{R}^+ \times [0,\pi] \times \mathbb{R}_+^* \times \mathbb{R} \times \mathbb{R}$, we have

$$\ln^+(h(r,\theta,\sigma,y,z)) \le \ln^+(1+\sigma r) \le \sigma r \tag{50}$$

and

$$\ln^-(h(r,\theta,\sigma,y,z)) \le \ln^-(\sin(\theta)) \,. \tag{51}$$

This gives

$$\gamma^+ \le \left(\frac{1}{2}\right)^{\frac{d}{2}} \frac{\sigma\pi}{W_{d-2}\Gamma\left(\frac{d}{2}\right)} \int_0^{+\infty} r^d e^{-\frac{r^2}{2}}\,dr < +\infty \,,$$

and

$$\begin{aligned}
\gamma^- &\le \left(\frac{1}{2}\right)^{\frac{d}{2}} \frac{1}{W_{d-2}\Gamma\left(\frac{d}{2}\right)} \int_0^{+\infty} \int_0^{\frac{\pi}{2}} \ln^-\left(sin(\theta)\right) r^{d-1} e^{-\frac{r^2}{2}} \sin^{d-2}(\theta)\, dr\, d\theta \\
&\le \left(\frac{1}{2}\right)^{\frac{d}{2}} \frac{2}{W_{d-2}\Gamma\left(\frac{d}{2}\right)} \int_0^{+\infty} r^{d-1} e^{-\frac{r^2}{2}}\,dr \int_0^{\frac{\pi}{2}} \sin^{d-\frac{5}{2}}(\theta)\,d\theta < +\infty \,.
\end{aligned}$$

For the remaining case $d = 1$, we have

$$\gamma^+ \le \frac{\sigma}{\sqrt{2\pi}} \int_{\mathbb{R}} |x| e^{-\frac{x^2}{2}} dx = \frac{2\sigma}{\sqrt{2\pi}} \int_{\mathbb{R}^+} x e^{-\frac{x^2}{2}} < +\infty \,,$$

For γ^-, after a change of variables ($v = \sigma x$), we get

$$\begin{aligned}
\gamma^- &\le \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}} \int_{m_\xi}^{M_\xi} \int_{-2}^{0} \int_{m_\xi}^{M_\xi} \frac{\ln\left(|1 + 1_{\{|1+v|^\alpha y < z\}}(v,y,z)v|\right)}{v} p_\xi(y) dv dy d\mu(z) \\
&= \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}} \int_{m_\xi}^{M_\xi} \int_{-2}^{0} \int_{m_\xi}^{M_\xi} \frac{\ln\left(|1+v|\right)}{v} 1_{\{|1+v|^\alpha y < z\}}(v,y,z) p_\xi(y) dv dy d\mu(z) \\
&\le \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}} \int_{m_\xi}^{M_\xi} \int_{-2}^{0} \int_{m_\xi}^{M_\xi} \frac{\ln\left(|1+v|\right)}{v} p_\xi(y) dv dy d\mu(z) \\
&= \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}} \int_{-2}^{0} \frac{\ln\left(|1+v|\right)}{v} dv < +\infty \,.
\end{aligned}$$

□

Lemma 11 *Suppose that $m_\xi \neq 0$. Then the quantity γ defined in Lemma 10 is such that $\gamma < 0$ if $m_\xi > 0$ and $\gamma > 0$ if $m_\xi < 0$.*

Proof Let $(\mathrm{a}_n)_{n\geq 0}$ be a sequence of i.i.d. random vectors in $\mathbb{R}^d$ distributed as $\mathcal{N}(0, I_d)$. Let also $(\mathrm{b}_n)_{n\geq 0}$ be a sequence of i.i.d. random variables with law $\mathcal{L}_\xi$. We consider $(T_n)_{n\in\mathbb{Z}^+}$ a Markov chain with transition kernel P (the transition kernel of $(Z_n)_n$) and initial distribution μ, the invariant probability measure of $(Z_n)_n$. By definition of the invariant measure, for all n, the law of T_n is μ. As for Z_n, the chain $(T_n)_n$ obeys the following induction relation:

$$T_{n+1} = \begin{cases} \mathrm{b}_n & \text{if } \|\mathrm{e}_1 + \sigma \mathrm{a}_n\|^\alpha \mathrm{b}_n < T_n \\ T_n & \text{otherwise}\,, \end{cases} \tag{52}$$

and T_0 distributed as μ. We also construct a sequence of random variables $(M_n)_{n\in\mathbb{N}}$, in the following manner

$$M_{n+1} = \begin{cases} \|\mathrm{e}_1 + \sigma \mathrm{a}_n\|^\alpha & \text{if } \|\mathrm{e}_1 + \sigma \mathrm{a}_n\|^\alpha \mathrm{b}_n < T_n\,, \\ 1 & \text{otherwise}\,, \end{cases} \tag{53}$$

The sequence $(M_n)_n$ satisfies $E_{\mu,\nu_\mathcal{N},\nu_\xi}\left[\ln(M_{n+1})\right] = \gamma$, where the notation $E_{\mu,\nu_\mathcal{N},\nu_\xi}$ reminds the different distribution of the random variables defining M_n. We also define the sequence of random variables $(W_n)_{n\in\mathbb{N}}$ as $W_{n+1} := M_{n+1}\frac{T_{n+1}}{T_n}$. Then, for $n \geq 1$, we have

$$W_{n+1} = \begin{cases} \|\mathrm{e}_1 + \sigma \mathrm{a}_n\|^\alpha \frac{\mathrm{b}_n}{T_n} & \text{if } \|\mathrm{e}_1 + \sigma \mathrm{a}_n\|^\alpha \mathrm{b}_n < T_n\,, \\ 1 & \text{otherwise}\,. \end{cases} \tag{54}$$

Besides, if $m_\xi > 0$, $\mathrm{b}_n > 0$, and then $T_n > 0$ for all $n \geq 0$. Consequently, $W_n \leq 1$ for all $n \in \mathbb{N}$. Since $W_{n+1} = M_{n+1}\frac{T_{n+1}}{T_n}$, we have that $\ln(W_{n+1}) = \ln(M_{n+1}) + \ln(T_{n+1}) - \ln(T_n)$. Therefore

$$E_{\mu,\nu_\mathcal{N},\nu_\xi}\left[\ln(W_{n+1})\right] = E_{\mu,\nu_\mathcal{N},\nu_\xi}\left[\ln(M_{n+1})\right] + E_\mu[\ln(T_{n+1})] - E_\mu[\ln(T_n)]\,.$$

Since $E_\mu[\ln(T_{n+1})] = E_\mu[\ln(T_n)]$, we have

$$\begin{aligned} \gamma = E_{\mu,\nu_\mathcal{N},\nu_\xi}\left[\ln(M_{n+1})\right] &= E_{\mu,\nu_\mathcal{N},\nu_\xi}\left[\ln(W_{n+1})\right] \\ &= E_{\mu,\nu_\mathcal{N},\nu_\xi}\left[\ln\left(\|\mathrm{e}_1 + \sigma \mathrm{a}_n\|^\alpha \frac{\mathrm{b}_n}{T_n}\right) 1_{\{\|\mathrm{e}_1+\sigma \mathrm{a}_n\|^\alpha \frac{\mathrm{b}_n}{T_n}<1\}}\right] \end{aligned}$$

The right hand-side of the last equation is strictly negative as the events $\left(\|\mathrm{e}_1 + \sigma \mathrm{a}_n\|^\alpha \frac{\mathrm{b}_n}{T_n} < 1\right)$ have a strictly positive measure and thus $\gamma < 0$ if $m_\xi > 0$.

Consider now the case where $m_\xi < 0$. Similarly to the proof of Lemma 1, there exists $n_0 \geq 0$ such that $T_n < 0$ for all $n \geq n_0$. This implies in particular that the support of the μ is embedded in $]m_\xi, 0[$ and thus that for all $n \geq 0$, $T_n < 0$. As in the previous case we have

$$E_{\mu,\nu_{\mathcal{N}},\nu_\xi}\left[\ln(W_{n+1})\right] = E_{\mu,\nu_{\mathcal{N}},\nu_\xi}\left[\ln(M_{n+1})\right] + E_\mu[\ln(-T_{n+1})] - E_\mu[\ln(-T_n)]\,.$$

and thus $\gamma = E_{\mu,\nu_{\mathcal{N}},\nu_\xi}\left[\ln(M_{n+1})\right] = E_{\mu,\nu_{\mathcal{N}},\nu_\xi}\left[\ln(W_{n+1})\right]$. Therefore

$$\gamma = E_{\mu,\nu_{\mathcal{N}},\nu_\xi}\left[\ln\left(\|\mathrm{e}_1 + \sigma \mathrm{a}_n\|^\alpha \frac{\mathrm{b}_n}{T_n}\right) 1_{\{\|\mathrm{e}_1+\sigma \mathrm{a}_n\|^\alpha \frac{\mathrm{b}_n}{T_n} > 1\}}\right]\,. \tag{55}$$

which is strictly positive since the events $\left(\|\mathrm{e}_1 + \sigma \mathrm{a}_n\|^\alpha \frac{\mathrm{b}_n}{T_n} > 1\right)$ have a strictly positive measure. □

Proof of Theorem 6: The almost sure convergence or divergence was already given in Theorem 1 and Theorem 2. Now, we investigate the convergence (or divergence) rate.

Corollary 2 states that $(Z_n, \mathcal{N}_n, \xi_n)_n$ is positive and Harris recurrent, moreover, Lemma 10 states that γ is well defined and finite. Therefore, we can apply the LLN for Markov chains (Theorem 4), in the sense that the right hand side of Eq. 13 converges almost surely to γ. Consequently the left-hand side of Eq. 13, the sequence $(\frac{1}{n}\ln\|\mathrm{X}_n\|)_n$, converges in distribution to γ. As γ is a constant, the convergence of the sequence $\frac{1}{n}\ln(\|\mathrm{X}_n\|)_n$ to γ holds also in probability. Finally, by Lemma 11, we have $\gamma < 0$ if $m_\xi > 0$ and $\gamma > 0$ if $m_\xi < 0$. □

Proof of Proposition 3 (stated page 20)

The convergence (or divergence) rate γ defined in Lemma 10 can be rewritten, according to the proof of Lemma 10, as $\gamma = \gamma^+ - \gamma^-$ where γ^+ and γ^- are positive finite quantities respectively defined in Equations 46, 49 and Equations 47, 48 which have been given in the proof of Lemma 10. The

continuity with respect to σ is shown using the Lebesgue dominated convergence theorem (for continuity) on every range $]0, M[$ and then for the whole $]0, +\infty[$ thanks to the inequalities given in Eq. 50 and Eq. 51. This gives the result for $d > 1$.
For the case $d = 1$, the integrand in γ^+ is continuous with respect to σ for almost all (x, y, z) in $\mathbb{R}\times]m_\xi, M_\xi[\times]m_\xi, M_\xi[$ and is dominated by $\frac{2}{\sqrt{2\pi}} S x e^{-\frac{x^2}{2}}$ for $(x, \sigma, y, z) \in \mathbb{R}^+\times]0, S]\times[0, +\infty[\times]m_\xi, M_\xi[\times]m_\xi, M_\xi[$ which gives the continuity of γ^+ with respect to σ by the Lebesgue dominated convergence theorem. For γ^-, and after the change of variables $v = \sigma x$, the integrand will be dominated by $\frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}} \frac{\ln(|1+v|)}{v}$ for $(v, \sigma, y, z) \in]-2, 0]\times]0, +\infty[\times]m_\xi, M_\xi[\times]m_\xi, M_\xi[$ and the continuity of γ^- with respect to σ follows from the dominated convergence theorem. $\square$