On the convergence of the $(1+1)$-ES in noisy spherical environments

Mohamed Jebalia, Anne Auger

On the convergence of the $(1+1)$-ES in noisy spherical environments

Mohamed Jebalia and Anne Auger

TAO Team, INRIA Futurs, France

Abstract. In this paper, we analyze the convergence of a $(1+1)$-ES on unimodal functions perturbed by noise. We investigate two models for the noise: (1) a model where the noise is scaled proportionally to the step-size and (2) a model where the noise is scaled proportionally to the (non-noisy part of the) fitness function. Those models were previously studied in the literature using gaussian noise and asymptotic estimations when the dimension of the search space grows to infinity. Similar results for both of them were obtained. For lower bounded noise (that does not include gaussian noise), we show that those models exhibit different behaviors: premature convergence occurs with probability one for the first model whenever the noise takes strictly negative values, whereas linear convergence always occurs for the second model. Moreover we exhibit for the second model a case where the convergence rate is equal to zero for any choice of the step-size.

1 Introduction

Evolutionary algorithms (EAs) are bio-inspired stochastic search algorithms that iteratively apply operators of variation and selection to a population of candidate solutions. Among EAs, adaptive Evolution Strategies (ESs) are recognized as state of the art algorithms when dealing with continuous optimization problems. Adaptive ESs sequentially adapt the parameters of the search distribution, usually a multivariate normal distribution, based on the history of the search. Several adaptation schemes have been introduced in the past. The one-fifth success rule [11, 9] considers the adaptation of one parameter, referred as the step-size, based on the success probability. The most advanced adaptation scheme, the Covariance Matrix Adaptation (CMA) adapts the full covariance matrix of the multivariate normal distribution [6]. Adaptive ESs do converge with order one, or log-linearly (that will be simply called linear convergence in what follows)[1] on the class of functions defined as

$$\mathrm{g}\left(\mathrm{f_s}(x)\right) = \mathrm{g}\left(\sum_{i=1}^{d} x_i^2\right) = \mathrm{g}\left(\|x\|^2\right) \tag{1}$$

where d is the dimension of the search space, g a strictly monotonically non-decreasing function and $\mathrm{f_s}(x) = \|x\|^2$ the so-called sphere function. Linear convergence has been

[1] The sequence (X_n) converges (log-)linearly if $\exists c \in \mathbb{R}$ such that $\lim_{n\to\infty} \frac{1}{n} \ln \|\mathrm{X}_n\| = c$. In this paper we will also talk about linear convergence whenever $c = 0$.

rigorously proven for adaptive ESs where the step-size is adapted for the class of functions (1) [5, 3].

Noise is present in many real-world optimization problems and can have various origins as measurement limitations or limited accuracy in simulation procedures. The precision of a fitness function evaluation might depend on the computational effort (CE) and noise can be reduced by increasing the CE. For instance, the fitness evaluation can result from an expensive Monte-Carlo (quasi Monte-Carlo) simulation where the number of samples used for the simulation directly controls the precision of the evaluation.

EAs are known to be robust in noisy environment. Theoretical studies of continuous EAs in the presence of noise have been carried out by Rechenberg [11], Arnold and Beyer [2, 1], using asymptotic estimations when the dimension of the search space tends to infinity. In this paper, we investigate the general noisy spherical model

$$\mathrm{f}(x, \eta) = \|x\| + \eta \mathcal{B} \tag{2}$$

where $x \in \mathbb{R}^d$, $\mathcal{B}$ is an independent random variable that models the noise and $\eta \in \mathbb{R}_+^*$[2] is a scaling parameter for the noise level. A natural way to model the noise is to consider that the variance of the noise is proportional to the fitness, *i.e* $\eta = \sigma_\epsilon \|x\|$ in the case of the sphere function (referred as model **pf**). In [1], considering the noisy sphere function, optimal adaptation schemes for the step-size (ie. scale-invariant algorithm) and gaussian noise, the authors approximate the model **pf** by a model where the standard deviation of the noise is proportional to the norm of the parent, or equivalently to the step-size. This model is referred as **pss**. Their justification is the following *"As for finite normalized mutation strength, the distance between parent and offspring is of order of $1/\sqrt{d}$, it can be assumed that the noise strength σ_ϵ at the location of the offspring is well approximated by that at the location of its parent. This assumption significantly simplifies the analysis of the local performance of the algorithm."*. For this model **pss** the noise level can be controlled [12]. For some problems indeed, it is possible to control the accuracy of the fitness evaluation, *i.e.* the parameter η. Consider for instance, a problem where the fitness evaluation comes from a Monte-Carlo estimation, increasing the number of evaluations by N will reduce the noise level by $\sqrt{N}$.

Those two models are considered as asymptotically equivalent for gaussian noise when d converges to ∞ [1]. However, for lower bounded noise, we show a fundamental difference in the two models for finite dimension: if the noise is lower bounded by a negative finite value, premature convergence with probability one occurs for the model **pss** whereas linear convergence occurs for the model **pf**.

In this paper, we address the question of the convergence of $(1+1)$-ES on the fitness functions (2) when the parameter η is controlled proportionally: (1) to the step-size (Model **pss**), (2) to the non noisy fitness (Model **pf**). In Section 2 we give results concerning the convergence of $(1+1)$-ES without noise and for noise with reevaluation. In Section 2.1 we derive sufficient conditions to prove linear convergence. In Section 2.2 we summarize the important results of the paper. In Section 3.1, we recall some basics about Markov chain theory. In Section 3.2 and Section 3.3 we prove the main results of the paper.

[2] $\mathbb{R}_+^* = \mathbb{R}^+ \backslash \{0\}$

Because of length constraints, we have not been able to include all the proofs. The interested reader can find them in the joint technical report [8].

2 Linear convergence of ESs and Markov Chains

Evolution Strategies (ES) have been originally introduced for minimizing continuous functions f mapping $\mathbb{R}^d$ into $\mathbb{R}$ where d is a positive integer. The simplest ES is a $(1+1)$-ES, where at each iteration (or generation) n, the random variable of $\mathbb{R}^d$ modeling the parent, X_n, is perturbed by the addition of a multivariate isotropic normal distribution $\mathcal{N}(0, I_d)$ scaled by a strictly positive real number σ_n also called step-size[3]. The resulting vector $\tilde{\mathrm{X}}_n$, called offspring, reads:

$$\tilde{\mathrm{X}}_n = \mathrm{X}_n + \sigma_n \mathcal{N}_n(0, I_d) \ . \tag{3}$$

The plus selection "+" means that the offspring is accepted if and only if its fitness value is smaller than the one of the parent. Several heuristics have been introduced for the adaptation of the step-size σ_n, the most popular being the one-fifth success rule [11]. The convergence of ESs is at most linear. Optimal bounds for the convergence rate are reached on the class of functions (1). The (optimal) step-size adaptation scheme giving the optimal bounds is for a so-called scale-invariant algorithm, where σ_n is proportional to the distance to the optimum, *i.e.* $\sigma_n = \sigma \|\mathrm{X}_n\|$ where $\sigma \in \mathbb{R}_+^*$ [4]. For the $(1+1)$-ES this result is stated in the following theorem:

Theorem 1 (Linear convergence and optimal bounds without noise). *For a $(1+1)$-ES, convergence is at most linear: There exists $c(d) < 0$ such that*

$$\liminf_{n\to\infty} \frac{1}{n} \ln\left(\frac{\|\mathrm{X}_n\|}{\|\mathrm{X}_0\|}\right) \geq c(d)$$

The optimal bound $c(d)$ is reached on the class of functions (1) for the scale-invariant algorithm where the step-size σ_n equals $\sigma^*_{\mathrm{opt}}\|\mathrm{X}_n\|$ with σ^*_{opt}, the positive constant minimizing the log-progress (or log-convergence rate) defined as $\varphi_{\ln}(\sigma^*) = E\left(\ln\left(\|\mathrm{e}_1 + \sigma^* \mathcal{N}_n\| \wedge 1\right)\right)$, i.e. $c(d) = E\left(\ln\left(\|\mathrm{e}_1 + \sigma^*_{\mathrm{opt}}\mathcal{N}\| \wedge 1\right)\right)$, where $\mathrm{e}_1 = (1, 0, \ldots, 0)$ and $a \wedge b$ denotes the minimum of a and b.

Proof. We consider one step of a $(1+1)$-ES where the offspring is sampled with a multivariate distribution with standard deviation σ_n: $\sigma_n \mathcal{N}_n$. Then $\mathrm{X}_{n+1} = \mathrm{X}_n + \mathbb{1}_{\{f(\mathrm{X}_n + \sigma_n \mathcal{N}_n) \leq f(\mathrm{X}_n)\}} \sigma_n \mathcal{N}_n$ where $\mathbb{1}_{\{f(\mathrm{X}_n + \sigma_n \mathcal{N}_n) \leq f(\mathrm{X}_n)\}}$ is equal to 1 when the offspring is accepted and 0 otherwise. Since X_{n+1} is either equal to $\mathrm{X}_n + \sigma_n \mathcal{N}_n$ or to X_n, the norm of X_{n+1} is greater than the minimum between $\|\mathrm{X}_n + \sigma_n \mathcal{N}_n\|$ and $\|\mathrm{X}_n\|$, *i.e.*

$$\|\mathrm{X}_{n+1}\| \geq \|\mathrm{X}_n\| \wedge \|\mathrm{X}_n + \sigma_n \mathcal{N}_n\| \ .$$

Taking the log of the previous equation, extracting $\|\mathrm{X}_n\|$ and taking the conditional expectation, we obtain

$$E(\ln \|\mathrm{X}_{n+1}\| | \mathrm{X}_n) \geq \ln \|\mathrm{X}_n\| + E\left(\ln\left(\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \frac{\sigma_n}{\|\mathrm{X}_n\|}\mathcal{N}_n\| \wedge 1\right) | \mathrm{X}_n, \sigma_n\right) \tag{4}$$

[3] The resulting multivariate normal distribution $\sigma_n \mathcal{N}(0, I_d)$ has a diagonal covariance matrix with all diagonal entries equal to σ_n^2.

Because the multivariate normal distribution is isotropic the conditional expectation $E\left(\ln\left(\|\frac{X_n}{\|X_n\|}+\frac{\sigma_n}{\|X_n\|}\mathcal{N}_n\|\wedge 1\right)|X_n,\sigma_n\right)$ is equal to $E\left(\ln\left(\|e_1+\frac{\sigma_n}{\|X_n\|}\mathcal{N}_n\|\wedge 1\right)|X_n,\sigma_n\right)$. Besides the following equation holds

$$E\left(\ln\left(\|e_1+\frac{\sigma_n}{\|X_n\|}\mathcal{N}_n\|\wedge 1\right)|X_n,\sigma_n\right)\geq \min_{\hat{\sigma}} E\left(\ln\left(\|e_1+\hat{\sigma}\mathcal{N}_n\|\wedge 1\right)\right) \tag{5}$$

Let denote σ^* the argmin of the right hand side of the previous equation. Let consider now the scale-invariant algorithm where at each generation n the step-size σ_n is equal to $\sigma\|X_n\|$. In distribution we have the following equality

$$\ln\|X_{n+1}\|=\ln\|X_n\|+\ln\left(\|e_1+\sigma\mathcal{N}_n\|\wedge 1\right)$$

Summing the previous equality for $k=1,\dots n-1$, dividing by n

$$\frac{1}{n}\ln\left(\frac{\|X_n\|}{\|X_0\|}\right)=\frac{1}{n}\sum_{k=0}^{n-1}\ln\frac{\|X_{k+1}\|}{\|X_k\|}=\frac{1}{n}\sum_{k=0}^{n-1}\ln\left(\|e_1+\sigma\mathcal{N}_k\|\wedge 1\right) \tag{6}$$

and applying the Strong Law of Large Numbers for independent random variables we obtain

$$\lim_{n\to\infty}\frac{1}{n}\ln\left(\frac{\|X_n\|}{\|X_0\|}\right)=E\left(\ln\left(\|e_1+\sigma\mathcal{N}_k\|\wedge 1\right)\right)$$

Besides, if $\sigma=\sigma^*$, Eq. 5 is an equality and one obtains that for any $(1+1)$-ES

$$\liminf_{n\to\infty}\frac{1}{n}\ln\left(\frac{\|X_n\|}{\|X_0\|}\right)\geq E\left(\ln\left(\|e_1+\sigma^*\mathcal{N}_k\|\wedge 1\right)\right)$$

□

The convergence rate of the scale-invariant $(1+1)$-ES with $\sigma_n=\sigma^*\|X_n\|$ is equal to $E\left(\ln\left(\|e_1+\sigma^*\mathcal{N}\|\wedge 1\right)\right)$. For any σ^*, the convergence rate is strictly negative and converges to 0 whenever σ grows to ∞. Because they give optimal bounds for the convergence rate, scale-invariant algorithms are important to study. Note that the result that convergence is at most linear is more general: true for any rank-based algorithm [13], or any Hit-and-Run direct search method [7].

The proof of the linear convergence in Theorem 1 relies on the application of the Strong Law of Large Numbers (LLN) for independent random variables (in Eq. 6). The LLN is the main ingredient for proofs of linear convergence of ESs. In most of the cases though, the dependency between consecutive generations precludes the use of LLN for independent variables. The Markovian properties of the ES process can be exploited to use LLN for Markov chains [5, 4].

For the scale-invariant algorithm, we investigate the differences between the noisy sphere, where the noise is scaled proportionally to the fitness (Model **pf**) and where the noise is scaled proportionally to the optimal step-size (Model **pss**). According to the model considered, at each generation, the fitness of the offspring $\tilde{X}_n=X_n+\sigma\|X_n\|\,\mathcal{N}_n$, is equal to

$$\begin{aligned} f_{pss}(\tilde{X}_n)&=\|\tilde{X}_n\|+\sigma_\epsilon\|X_n\|\mathcal{B}_n \qquad \text{for the model } \mathbf{ps}\\ f_{pf}(\tilde{X}_n)&=\|\tilde{X}_n\|+\sigma_\epsilon\|\tilde{X}_n\|\mathcal{B}_n \qquad \text{for the model } \mathbf{pf}\end{aligned} \tag{7}$$

where $\mathcal{B}_n$ is distributed as $\mathcal{B}$ and is independent, σ_ϵ is a strictly positive constant. It is usually assumed that the fitness of one individual is computed once. In the case of the "+" selection investigated here, this implies that the current generation is correlated to the generation where the offspring giving this parent has been accepted. This prevents to use the LLN for independent random variables as it is done in Eq. 6 and one has to use Markov chain theory for the analysis.

However the independent LLN can be applied and Theorem 1 generalized if we introduce reevaluation of the fitness function at each generation. Reevaluation means sample a new random variable $\bar{\mathcal{B}}$ distributed as the noise $\mathcal{B}$ to reevaluate the parent when compared to the offspring. Consequently an offspring $\tilde{\mathrm{X}}_n$ is accepted if its fitness is better than the reevaluated fitness of the parent.

Theorem 2. *The scale-invariant* $(1+1)$*-ES with* $\sigma_n = \sigma\|\mathrm{X}_n\|$ *(*σ *strictly positive) and reevaluation converges linearly for the noisy sphere function* $\mathrm{f}_{\mathrm{pss}}$ *(resp.* f_{pf}*):*

$$\frac{1}{n}\ln\|\mathrm{X}_n\| \xrightarrow[n\to\infty]{} E\left[\ln(\|\mathrm{e}_1 + \delta\sigma\mathcal{N}(0, I_d)\|)\right],$$

where δ *is the Bernoulli random variable equals* 1 *when*

$$\|\mathrm{e}_1 + \sigma\mathcal{N}(0, I_d)\| + \sigma_\epsilon\mathcal{B} \leq 1 + \sigma_\epsilon\bar{\mathcal{B}} \text{ if model } \textbf{pss}$$

$$\left(\textit{resp. } \|\mathrm{e}_1 + \sigma\mathcal{N}(0, I_d)\|\,(1 + \sigma_\epsilon\mathcal{B}) \leq 1 + \sigma_\epsilon\bar{\mathcal{B}} \text{ if model } \textbf{pf}\right)$$

where $\mathcal{B}$ *and* $\bar{\mathcal{B}}$ *are two independent samplings of the noise distribution.*

In the previous theorem, for every σ the Bernoulli random variable is non trivial (non identically equal to zero). Therefore, the convergence rate is non zero for almost every σ. In the sequel, no reevaluation is considered.

We introduce now some notations and definitions needed for the study of the scale-invariant algorithm on $\mathrm{f}_{\mathrm{pss}}$ and f_{pf} without reevaluations. Let b_n denote the noise associated to the parent X_n, *i.e.* $\mathrm{f}_\times(\mathrm{X}_n) = \|\mathrm{X}_n\| + b_n$ where $\mathrm{f}_\times$ stands for $\mathrm{f}_{\mathrm{pss}}$ or f_{pf}. Let δ_n denote the random variable equals 1 whenever the offspring is accepted and zero otherwise. For Model **pss**, $\delta_n = 1$ if and only if

$$\|\tilde{\mathrm{X}}_n\| + \sigma_\epsilon\|\mathrm{X}_n\|\mathcal{B}_n \leq \|\mathrm{X}_n\| + b_n \tag{8}$$

and for Model **pf**, $\delta_n = 1$ if and only if

$$\|\tilde{\mathrm{X}}_n\|\,(1 + \sigma_\epsilon\mathcal{B}_n) \leq \|\mathrm{X}_n\| + b_n \tag{9}$$

The update of the random vector X_n has the same form for both models :

$$\mathrm{X}_{n+1} = \delta_n\tilde{\mathrm{X}}_n + (1-\delta_n)\mathrm{X}_n \tag{10}$$

$$= \mathrm{X}_n + \delta_n\sigma\|\mathrm{X}_n\|\mathcal{N}_n \tag{11}$$

And the difference between the two models is hidden in δ_n, that do not share the same distribution. The update of the random variable b_n is

$$b_{n+1} = \delta_n\left(\sigma_\epsilon\|\tilde{\mathrm{X}}_n\|\mathcal{B}_n\right) + (1-\delta_n)b_n \text{ for Model } \textbf{pss} \tag{12}$$

$$b_{n+1} = \delta_n\left(\sigma_\epsilon\|\mathrm{X}_n\|\mathcal{B}_n\right) + (1-\delta_n)b_n \text{ for Model } \textbf{pf}\ . \tag{13}$$

2.1 Sufficient condition for linear convergence

It is classical when studying stochastic search algorithms that linear convergence of X_n can be studied investigating the stability of $\|\mathrm{X}_{n+1}\|/\|\mathrm{X}_n\|$. This idea was introduced in the context of ESs by Bienvenüe and François [5] and exploited in [3]. Here, the stability of $\|\mathrm{X}_{n+1}\|/\|\mathrm{X}_n\|$ will follow from the one of $b_n/\|\mathrm{X}_n\|$. This is formalized in the following propositions:

Proposition 1 (Model pss). *Let Z_n be the Markov chain defined as:*

$$Z_{n+1} = \delta_n(Z_n)\left(\frac{\sigma_\epsilon \mathcal{B}_n}{\|\mathrm{e}_1 + \sigma\mathcal{N}_n\|}\right) + (1-\delta_n(Z_n))Z_n \tag{14}$$

where $\mathcal{B}_n$ is distributed as $\mathcal{B}$, $\mathcal{N}_n$ is a gaussian vector with mean zero and covariance matrix identity, $\mathrm{e}_1 = (1,0,\ldots,0)$, and $\delta_n(Z_n)$ equals 1 if $\|\mathrm{e}_1+\sigma\mathcal{N}_n\| + \sigma_\epsilon\mathcal{B}_n - 1 \leq Z_n$ and 0 otherwise. Then Z_n and $b_n/\|\mathrm{X}_n\|$, where b_n and X_n are defined in Eq. 12 and Eq. 10, follow the same distribution. If Z_n is positive-Harris recurrent, linear convergence for $\|\mathrm{X}_n\|$ occurs (in probability):

$$\frac{1}{n}\ln(\|\mathrm{X}_n\|) \xrightarrow[n\to\infty]{} \int E\left[\ln(\|\mathrm{e}_1 + \delta(z)\sigma\mathcal{N}\|)\right] d\mu(z) \tag{15}$$

where μ is the invariant probability measure of Z_n and $\delta(z)$ equals 1 if $\|\mathrm{e}_1+\sigma\mathcal{N}\| + \sigma_\epsilon\mathcal{B} - 1 \leq z$ and 0 otherwise.

Proposition 2 (Model pf). *Let Z_n be the Markov chain defined as:*

$$\begin{aligned} Z_0 &= \sigma_\epsilon\mathcal{B}_0 \\ Z_{n+1} &= \delta_n(Z_n)\sigma_\epsilon\mathcal{B}_n + (1-\delta_n(Z_n))Z_n \end{aligned} \tag{16}$$

where $\mathcal{B}_n$ is distributed as $\mathcal{B}$, $\mathrm{e}_1 = (1,0,\ldots,0)$, and $\delta_n(Z_n)$ equals 1 if $\|\mathrm{e}_1 + \sigma\mathcal{N}_n\|(1+\sigma_\epsilon\mathcal{B}_n) - 1 \leq Z_n$ and 0 otherwise (where $\mathcal{N}_n$ is a gaussian vector with mean zero and covariance matrix identity). Then Z_n and $b_n/\|\mathrm{X}_n\|$, where b_n and X_n are defined in Eq. 13 and Eq. 10, follow the same distribution. If Z_n is positive-Harris recurrent, linear convergence for $\|\mathrm{X}_n\|$ occurs in probability:

$$\frac{1}{n}\ln(\|\mathrm{X}_n\|) \xrightarrow[n\to\infty]{} \int E\left[\ln(\|\mathrm{e}_1 + \delta(z)\sigma\mathcal{N}\|)\right] d\mu(z) \tag{17}$$

where μ is the invariant probability measure of Z_n and $\delta(z)$ equals 1 if $\|\mathrm{e}_1 + \sigma\mathcal{N}\|(1+\sigma_\epsilon\mathcal{B}) - 1 \leq z$ and 0 otherwise.

2.2 Main results

The study of the stability of (Z_n) depends on the support of the noise $\mathcal{B}$. We make the following assumptions:

Assumption 1 *1. $\mathcal{B}$ is absolutely continuous with respect to the Lebesgue measure and $p_\mathcal{B}$ denotes the associated density.*

2. *The support of the noise satisfies $supp(p_\mathcal{B}) = [m_\mathcal{B}, M_\mathcal{B}[$ where $m_\mathcal{B} \in]-\infty, 0]$ and $M_\mathcal{B} \in]0, +\infty]$.*
3. *The random variable $\mathcal{B}$ is integrable and $E(|\mathcal{B}|) < \infty$.*

Model pss For this model, we show in the remainder of this paper that convergence of the $(1+1)$-ES depends on the support of the noise $\mathcal{B}$. The following theorem summarizes our results:

Theorem 3. *Let* X_n *be the random sequence defined in Section 2 and following the model* **pss**. *Under Assumption 1, the following holds:*

- *If the lower bound of the noise is strictly negative, i.e.* $m_{\mathcal{B}} < 0$, X_n *converges to* $x_0 \neq 0$ *with probability one. In other words, premature convergence to a non-optimal point occurs with probability one.*
- *If the noise is positive i.e.* $m_{\mathcal{B}} = 0$, Z_n, *defined in Proposition 1, is positive Harris recurrent and* $\|\mathrm{X}_n\|$ *converges linearly (in probability):*

$$\frac{1}{n}\ln(\|\mathrm{X}_n\|) \xrightarrow[n\to\infty]{} \int E\left[\ln(\|\mathrm{e}_1 + \delta(z)\sigma\mathcal{N}\|)\right] d\mu(z)$$

where μ *is the invariant probability measure of* Z_n *and* $\delta(z)$ *equals 1 if* $\|\mathrm{e}_1 + \sigma\mathcal{N}\| + \sigma_\epsilon\mathcal{B} - 1 \leq z$ *and* 0 *otherwise. The invariant measure* μ *is absolutely continuous and the convergence rate is non-zero for almost every* σ.

Proof. See Theorems 7 and 8 for each case. □

Model pf The following Theorem summarizes the results obtained when the noise is proportional to the location:

Theorem 4. *Let* X_n *be the random sequence defined in Section 2 and following the model* **pf**. *Under Assumption 1,* Z_n *defined in Proposition 2 is positive Harris recurrent and* $\|\mathrm{X}_n\|$ *converges linearly (in probability):*

$$\frac{1}{n}\ln(\|\mathrm{X}_n\|) \xrightarrow[n\to\infty]{} \int E\left[\ln(\|\mathrm{e}_1 + \delta(z)\sigma\mathcal{N}\|)\right] d\mu(z)$$

where μ *is the invariant probability measure of* Z_n *and* $\delta(z)$ *equals 1 if* $\|\mathrm{e}_1 + \sigma\mathcal{N}\|(1 + \sigma_\epsilon\mathcal{B}) - 1 \leq z$ *and* 0 *otherwise.*
If $\sigma_\epsilon m_{\mathcal{B}} = -1$, *the invariant measure* μ *is singular and is equal to* $\delta_{\{-1\}}$. *This implies that for all* σ, *the convergence rate is equal to* 0. *If* $\sigma_\epsilon m_{\mathcal{B}} \neq -1$, *the invariant measure is absolutely continuous and the convergence rate is non-zero for almost every* σ.

Proof. See Section 3.4. □

3 Stability

3.1 Basics about Markov chains and definitions

In Propositions 1 and 2, we have seen that linear convergence can be implied from the stability of the chain Z_n. Before investigating the stability we recall some definitions and results about φ-irreducible Markov Chains that will be widely used in the sequel. We refer to the Meyn and Tweedie book for a complete presentation of this theory [10].

In the remainder of the paper, $\mathfrak{B}(X)$ will denote the Borel σ-algebra on a subset X of $\mathbb{R}^d$.
For a Markov chain $Z_n \subset \mathbb{R}$, the transition kernel $P(.,.)$ is defined for all $z \in \mathbb{R}$, for all $A \in \mathfrak{B}(\mathbb{R})$ as

$$P(z, A) = P(Z_1 \in A | Z_0 = z).$$

The first notion of stability is given by the φ-irreducibility. A chain (Z_n) is irreducible with respect to a measure φ if:

$$\forall (z, A) \in \mathbb{R} \times \mathfrak{B}(\mathbb{R}) \text{ such that } \varphi(A) > 0, \exists n_0 \geq 0 \text{ such that } P^{n_0}(z, A) > 0 \quad (18)$$

or equivalently $\forall z \in \mathbb{R}, A \in \mathfrak{B}(\mathbb{R})$ such that $\varphi(A) > 0,\ P_z(\tau_A < \infty) > 0$ where, τ_A is the hitting time of Z_n on A, *i.e.*

$$\tau_A = \min\{n \geq 0 \text{ such that } Z_n \in A\}.$$

If the last term of Eq. 18 is equal to one, the chain is recurrent. A φ-irreducible chain Z_n is Harris recurrent if:

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

where η_A is the occupation time of A, i.e. $\eta_A = \sum_{n=1}^{\infty} \mathbb{1}_{\{Z_n \in A\}}$.

A chain (Z_n) which is Harris-recurrent admits an invariant measure, *i.e.* a measure π on $\mathfrak{B}(\mathbb{R})$ satisfying:

$$\pi(A) = \int_{\mathbb{R}} \pi(\mathrm{d}z) P(z, A), A \in \mathfrak{B}(\mathbb{R})$$

If in addition this measure is a probability measure, the chain is called positive. To prove Harris-reccurrence and positivity of Z_n we will make use of practical drift conditions. First we need the notion of small sets. A set C is called small set if $\exists \delta > 0, n > 0$, and some non trivial probability measure ν, such that:

$$P^n(z, A) \geq \delta \nu(A), z \in C \quad (19)$$

A φ-irreducible chain (Z_n) is called aperiodic if, for some (and then for every) small set with $\varphi(C) > 0$, 1 is the g.c.d of all values n for which Eq. 19 holds.
The drift operator is defined as

$$\Delta V(z) = E\left[V(Z_1) | Z_0 = z\right] - V(z) = \int P(z, dy) V(y) - V(z).$$

Drift conditions can be used to show Harris recurrence and positivity. We will use the following theorem wich gives drift conditions allowing to prove geometric ergodicity which is a stability criterium stronger than Harris-recurrence and positivity:

Theorem 5 (Condition for geometric ergodicity [10]). *Suppose that Z_n is φ-irreducible and aperiodic. The following drift condition is a sufficient and necessary condition for geometric ergodicity.*
There exists a function $V \geq 1$, finite at least for one z, and a small set C, such that for some $\lambda_C > 0$, $b_C < \infty$ the following drit condition holds

$$\Delta V \leq -\lambda_C V + b_C \mathbb{1}_C \quad (20)$$

Geometric ergodicity implies positivity and Harris recurrence. Therefore Eq. 20 is a sufficient condition for positivity and Harris-recurrence. Positive, Harris-reccurent chains satisfy the Strong-Law of Large Numbers (LLN):

Theorem 6 (LLN for Harris Positive chains). *Suppose that Z_n is a positive Harris chain with invariant probability measure π, then the LLN holds for any f satisfying $\pi(f) = \int f \mathrm{d}\pi < \infty$, i.e.*

$$\frac{1}{n} \lim_{n\to\infty} \sum_{k=1}^{n} f(Z_k) = \pi(f) \tag{21}$$

The Lebesgue measure in $\mathbb{R}^d$ will be denoted $\mu^{\mathrm{Leb,d}}$. When d is equal to one it will be simply denoted μ^{Leb}. For a measurable set E of $\mathbb{R}^d$, we will denote $\mu_{\mathrm{E}}^{\mathrm{Leb,d}}$ the measure defined for all measurable set A by $\mu_{\mathrm{E}}^{\mathrm{Leb,d}}(\mathrm{A}) = \mu^{\mathrm{Leb,d}}(\mathrm{A} \cap \mathrm{E})$. In particular $\mu_{\mathbb{R}^+}^{\mathrm{Leb}}$ is the measure induced by the Lebesgue measure on $\mathbb{R}^+$.

3.2 Model pss: non convergence for lower bounded noise

In this section we study the model **pss** when the lower bound of the noise is strictly negative, *i.e.* $m_{\mathcal{B}} < 0$. We show that premature convergence occurs with probability one. The sketch of the proof is the following. Let denote $\alpha_{\mathcal{B}} = -1 + \sigma_\epsilon m_{\mathcal{B}}$. We remark that $\alpha_{\mathcal{B}} < -1 < 0$. We are going to show that for all $Z_0 = z$, the probability that Z_n enters $]-\infty, \alpha_{\mathcal{B}}]$ in finite time is equal to one. Besides, whenever Z_n is in $]-\infty, \alpha_{\mathcal{B}}]$, Z_n is constant. Moreover, there is equivalence between Z_n being in $]-\infty, \alpha_{\mathcal{B}}]$ and X_n stuck in a point that is non-zero.

Theorem 7. *If $m_{\mathcal{B}} < 0$ and $E(|\mathcal{B}|) < \infty$, almost surely the algorithm does not converge to zero, i.e. the algorithm is stuck in $x_0 \neq 0$ with probability one.*

3.3 Model pss: linear convergence for positive noise

In this section, we assume that the noise is positive *i.e.* $m_{\mathcal{B}} = 0$.

Theorem 8. *If $m_{\mathcal{B}} = 0$ and $E(|\mathcal{B}|) < \infty$, Z_n is positive, Harris-reccurent. Therefore, linear convergence of X_n holds:*

$$\frac{1}{n} \ln(\|\mathrm{X}_n\|) \xrightarrow[n\to\infty]{} \int E\left[\ln(\|\mathrm{e}_1 + \delta(z)\sigma\mathcal{N}\|)\right] d\mu(z)$$

where μ is the stationary measure of Z_n and $\delta(z)$ equals 1 if $\|\mathrm{e}_1 + \sigma\mathcal{N}\| + \sigma_\epsilon\mathcal{B} - 1 \leq z$ and 0 otherwise.

Proof. As the noise is positive, $(Z_n) \subset \mathbb{R}^+$. Next, Propositions 3, 4, 5 show that the chain (Z_n) is $\mu_{\mathbb{R}^+}^{\mathrm{Leb}}$-irreducible, Harris recurrent and positive then it satisfies the Strong Law of Large Numbers. □

Proposition 3. *Suppose that $0 \in supp(p_{\mathcal{B}}) \subset \mathbb{R}^+$. Then the chain $(Z_n)_{n\geq 0} \subset \mathbb{R}^+$ is $\mu_{\mathbb{R}^+}^{\mathrm{Leb}}$-irreducible.*

Proposition 4. *The chain $(Z_n)_{n\geq 0}$ is aperiodic and non-empty compact of $\mathbb{R}^+$ are some small sets.*

Proposition 5. *For $V(z) = z^\alpha + 1,\ 0 < \alpha < 1$, the chain verifies drift conditions for geometric ergodicity.*

3.4 Model pf: linear convergence for lower bounded noise

In this case, in Proposition 6, we show that $(Z_n) \subset \sigma_\epsilon \mathrm{supp}(p_\mathcal{B})$. Next, Propositions 7, 8 and 9 show that Z_n is positive, Harris-reccurent. Therefore(Z_n) satisfies the Strong Law of Large Numbers and linear convergence of X_n holds as stated in Theorem 4.

Proposition 6. $(Z_n) \subset \sigma_\epsilon$*supp*$(p_\mathcal{B}) = [\sigma_\epsilon m_\mathcal{B}, \sigma_\epsilon M_\mathcal{B}[$.

Proposition 7. *1. If $-1 < \sigma_\epsilon m_\mathcal{B} \leq 0$, the chain $(Z_n)_{n\geq 0} \subset \sigma_\epsilon$supp$(p_\mathcal{B})$ is $\mu^{\mathrm{Leb}}_{[\sigma_\epsilon m_\mathcal{B}, \sigma_\epsilon M_\mathcal{B}[}$-irreducible.*
2. If $-\infty < \sigma_\epsilon m_\mathcal{B} < -1$, we have the following:
  - *The chain $(Z_n)_{n\geq 0} \subset \sigma_\epsilon$supp$(p_\mathcal{B})$ is not $\mu^{\mathrm{Leb}}_{[\sigma_\epsilon m_\mathcal{B}, \sigma_\epsilon M_\mathcal{B}[}$-irreducible.*
  - *The chain $(Z_n)_{n\geq 0} \subset \sigma_\epsilon$supp$(p_\mathcal{B})$ is $\mu^{\mathrm{Leb}}_{[\sigma_\epsilon m_\mathcal{B}, -1]}$-irreducible.*

3. If $\sigma_\epsilon m_\mathcal{B} = -1$, the chain $(Z_n)_{n\geq 0} \subset \sigma_\epsilon$supp$(p_\mathcal{B})$ is $\delta_{\{-1\}}$-irreducible.[4]

Proposition 8. *The chain $(Z_n)_{n\geq 0}$ is aperiodic and non-empty compact of $\sigma_\epsilon$supp$(p_\mathcal{B})$ are some small sets.*

Proposition 9. *For $V(z) = |z| + 1$, if $E\left[|\mathcal{B}|\right] < \infty$, the chain verifies drift conditions for geometric ergodicity.*

4 Discussion and conclusion

In this paper we have analyzed the convergence of the scale-invariant $(1 + 1)$-ES for the noisy sphere function. Two models for the noise have been analyzed: the model **pss**, where the noise is scaled proportionally to the step-size and therefore to the norm of the parent; the model **pf**, where the noise is scaled proportionally to location of the individual or to the non-noisy part of the fitness.

In the case where reevaluation takes place every generation, we show that the results obtained for the non-noisy case can be extended in a straightforward way. In particular, the linear convergence is implied from the application of the Strong Law of Large Numbers for independent random variables.

The analyses of the cases without reevaluation imply the use of Markov Chains. The Strong Law of Large Numbers is deduced from the stability of underlying Markov Chains.

For both models, we assume that the support of the noise is an interval $[m_\mathcal{B}, M_\mathcal{B}[$ where $m_\mathcal{B} \in]-\infty, 0]$ and $M_\mathcal{B} \in]0, +\infty]$. In particular the analysis does not include yet the case of gaussian noise.

[4] The measure $\delta_{\{-1\}}$ is the measure such that for$A \in \mathfrak{B}(\mathbb{R})$, $\delta_{\{-1\}}(A) = 1$ if $-1 \in A$ and 0 otherwise.

We show that both models exhibit different behaviors. For the model **pss** with a noise that can take negative values ($m_{\mathcal{B}} < 0$), premature convergence (to a non-optimal point) occurs with probability one. However when $m_{\mathcal{B}} = 0$ linear convergence occurs. For the model **pf**, linear convergence occurs in any case. In the scale-invariant algorithm analyzed, the step-size σ_n is proportional to the norm of the parent, $\sigma_n = \sigma\|\mathrm{X}_n\|$. The convergence rate associated to the different models is a continuous function of the parameter σ. In general, the convergence rate is non-zero for almost every σ. We exhibit a specific case where this is not true: the model **pf** and the lower bound of the scaled noise equal to -1. In that case, we proved that for all σ, the convergence rate is zero.

Acknowledgments

We would like to thank Nikolaus Hansen for fruitful discussions, valuable comments about this work and for indirectly encouraging us to use several footnotes within the paper. This work was partially funded by the ANR/RNTL project Optimisation Multidisciplinaire (OMD) and by the ACI CHROMALGEMA.

References

1. 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.
2. D. V. Arnold and H.-G. Beyer. A general noise model and its effects on evolution strategy performance. *IEEE Transactions on Evolutionary Computation*, 10(4):380–391, 2006.
3. A. Auger. Convergence results for (1,λ)-SA-ES using the theory of φ-irreducible markov chains. *Theoretical Computer Science*, 334:35–69, 2005.
4. A. Auger and N. Hansen. Reconsidering the progress rate theory for evolution strategies in finite dimensions. In A. Press, editor, *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2006)*, pages 445–452, 2006.
5. A. Bienvenüe and O. François. Global convergence for evolution strategies in spherical problems: some simple proofs and difficulties. *Theor. Comput. Sci.*, 306(1-3):269–289, 2003.
6. N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.
7. J. Jägersküpper. Lower bounds for hit-and-run direct search. In *Stochastic Algorithms: Foundations and Applications - SAGA 2007, LNCS 4665*. Springer Berlin, Heidelberg.
8. M. Jebalia and A. Auger. On the convergence of the $(1+1)$-ES in noisy spherical environments. Technical report, INRIA, 2007.
9. S. Kern, S. Müller, N. Hansen, D. Büche, J. Ocenasek, and P. Koumoutsakos. Learning Probability Distributions in Continuous Evolutionary Algorithms - A Comparative Review. *Natural Computing*, 3:77–112, 2004.
10. S. Meyn and R. Tweedie. *Markov Chains and Stochastic Stability*. Springer-Verlag, New York, 1993.
11. I. Rechenberg. *Evolutionstrategie: Optimierung Technisher Systeme nach Prinzipien des Biologischen Evolution*. Fromman-Hozlboog Verlag, Stuttgart, 1973.
12. O. Teytaud and A. Auger. On the adaptation of the noise level for stochastic optimization. In *IEEE Congress on Evolutionary Computation - CEC 2007*. IEEE, 2007.
13. O. Teytaud and S. Gelly. General lower bounds for evolutionary algorithms. In 10^{th} *International Conference on Parallel Problem Solving from Nature (PPSN 2006)*, 2006.