Log-log Convergence for Noisy Optimization

Sandra Cecilia Astete-Morales, Jialin Liu, Olivier Teytaud

Log-log Convergence for Noisy Optimization

S. Astete-Morales, J. Liu, O. Teytaud

TAO (Inria), LRI, UMR 8623 (CNRS - Univ. Paris-Sud), France

Abstract. We consider noisy optimization problems, without the assumption of variance vanishing in the neighborhood of the optimum. We show mathematically that simple rules with exponential number of resamplings lead to a log-log convergence rate. In particular, in this case the log of the distance to the optimum is linear on the log of the number of resamplings. As well as with number of resamplings polynomial in the inverse step-size. We show empirically that this convergence rate is obtained also with polynomial number of resamplings. In this polynomial resampling setting, using classical evolution strategies and an *ad hoc* choice of the number of resamplings, we seemingly get the same rate as those obtained with specific Estimation of Distribution Algorithms designed for noisy setting.
We also experiment non-adaptive polynomial resamplings. Compared to the state of the art, our results provide (i) proofs of log-log convergence for evolution strategies (which were not covered by existing results) in the case of objective functions with quadratic expectations and constant noise, (ii) log-log rates also for objective functions with expectation $\mathbb{E}[f(x)] = ||x - x^*||^p$, where x^* represents the optimum (iii) experiments with different parametrizations than those considered in the proof. These results propose some simple revaluation schemes. This paper extends [1].

1 Introduction

In this introduction, we first present the noisy optimization setting and the local case of it. We then classify existing optimization algorithms for such settings. Afterwards we discuss log-linear and log-log scales for convergence and give an overview of the paper. In all the paper, log represents the natural logarithm and $\mathcal{N}$ is a standard Gaussian random variable (possibly multidimensional, depending on the context), except when it is specified explicitly that $\mathcal{N}$ may be any random variable with bounded density.

Noisy optimization. This term will denote the optimization of an objective function which has internal stochastic effects. When the algorithm requests $fitness(\cdot)$ of a point x, it gets in fact $fitness(x, \theta)$ for a realization of a random variable θ. All calls to $fitness(\cdot)$ are based on independent realizations of the same random variable θ. The goal of a noisy optimization algorithm is to find x such that $\mathbb{E}(fitness(x, \theta))$ is minimized (or nearly minimized).

Local noisy optimization.

Local noisy optimization refers to the optimization of an objective function in which the main problem is noise, and not local minima. Hence, diversity mechanisms as in [2] or [3], in spite of their qualities, are not relevant here. We also restrict our work to noisy settings in which noise does not decrease to 0

around the optimum. This constrain makes our work different from [4]. In [5, 6] we can find noise models related to ours but the results presented here are not covered by their analysis. On the other hand, in [7–9], different noise models (with Bernoulli fitness values) are considered, inclusing a noise with variance which does not decrease to 0 (as in the present paper). They provide general lower bounds, or convergence rates for specific algorithms, whereas we consider convergence rates for classical evolution strategies equipped with resamplings.

Classification of local noisy optimization algorithms. We classify noisy local convergence algorithms in the following 3 families:

- *Algorithms based on sampling, as far as they can, close to the optimum.* In this category, we include evolution strategies[10, 6, 5] and EDA[11] as well as pattern search methods designed for noisy cases[12–14]. Typically, these algorithms are based on noise-free algorithms, and evaluate individuals multiple times in order to cancel (reduce) the effect of noise. Authors studying such algorithms focus on the number of resamplings; it can be chosen by estimating the noise level[15], or using the step-size, or, as in parts of the present work, in a non-adaptive manner.
- *Algorithms which learn (model) the objective function*, sample at locations in which the model is not precise enough, and then assume that the optimum is nearly the optimum of the learnt model. Surrogate models and Gaussian processes[2, 16] belong to this family. However, Gaussian processes are usually supposed to achieve global convergence (i.e. good properties on multimodal functions) rather than local convergence (i.e. good properties on unimodal functions) - in this paper, we focus on local convergence.
- *Algorithms which combine both ideas*, assuming that learning the objective function is a good idea for handling noise issues but considering that points too far from the optimum cannot be that useful for an optimization. This assumption makes sense at least in a scenario in which the objective function cannot be that easy to learn on the whole search domain. CLOP[7, 8] is such an approach.

Log-linear scale and log-log scale: uniform and non-uniform rates. To ensure the convergence of an algorithm and analyze the rate at which it converges are part of the main goals when it comes to the study of optimization algorithms.

In the noise-free case, evolution strategies typically converge linearly in log-linear scale, this is, the logarithm of the distance to the optimum typically scales linearly with the number of evaluations (see Section 2.1 for more details on this). The case of noisy fitness values leads to a log-log convergence[9]. We investigate conditions under which such a log-log convergence is possible. In particular, we focus on uniform rates. Uniform means that all points are under a linear curve in the log-log scale. Formally, the rate is the infimum of C such that with probability $1-\delta$, for m sufficiently large, all iterates after m fitness evaluations verify $\log ||x_m|| \leq -C \log m$, where x_m is the m^{th} evaluated individual. This is, all points are supposed to be "good" (i.e. satisfy the inequality); not only the best point of a given iteration. In contrast, a non-uniform rate would be the infimum of C such that $\log ||x_{k_m}|| \leq -C \log k_m$ for some increasing sequence k_m.

The state of the art in this matter exhibits various results. For an objective function with expectation $\mathbb{E}[f(x)] = ||x - x^*||^2$, when the variance is not supposed to decrease in the neighborhood of the optimum, it is known that the best possible slope in this log-log graph is $-\frac{1}{2}$ (see [17]), but without uniform rate. When optimizing $f(x) = ||x||^p + \mathcal{N}$, this slope is provably limited to $-\frac{1}{p}$ under locality assumption (i.e. when sampling far from the optimum does not help, see [9] for a formalization of this assumption), and it is known that some ad hoc EDA can reach $-\frac{1}{2p}$ (see [18]).
For evolution strategies, the slope is not known. Also, the optimal rate for $\mathbb{E}[f(x)] = ||x - x^*||^p$ for $p \neq 2$ is unknown; we show that our evolution strategies with simple revaluation schemes have linear convergence in log-log representation in such a case.
Algorithms considered in this paper. We here focus on simple revaluation rules in evolution strategies, based on choosing the number of resamplings. We start with rules which decide the number of revaluations only depending on the iteration number n. This is, independently of the step-size σ_n, the parents x_n and fitness values. To the best of our knowledge, these simple rules have not been analyzed so far. Nonetheless, they have strong advantages: we get a linear slope in log-log curve simple rules only depending on n whereas rules based on numbers of resamplings defined as a function of σ_n have a strong sensitivity to parameters. Also evolution strategies, contrarily to algorithms with good non-uniform rates, have a nice empirical behavior from the point of view of uniform rates, as shown mathematically by [18].
Overview of the paper. In this paper we show mathematical proofs and experimental results on the convergence of the evolutionary algorithms that will be described in the following sections, which include some resampling rules aiming to cancel the effect of noise. The theoretical analysis presents an exponential number of resamplings together with an assumption of scale invariance. This result is extended to an adaptive rule of resamplings (Section 2.3), in which the number of evaluations depend on the step size only; we also get rid of the scale invariant assumption. Essentially, the algorithms for which we get a proof have the same dynamics as in the noise-free case, they just use enough resamplings for cancelling the noise. This is consistent with the existing literature, in particular [18] which shows a log-log convergence for an Estimation of Distribution Algorithm with exponentially decreasing step-size and exponentially increasing number of resamplings.
In the experimental part, we see that another solution is a polynomially increasing number of resamplings (independently of σ_n; the number of resamplings just smoothly increases with the number of iterations, in a non-adaptive manner), leading to a slower convergence when considering the progress rate per iteration, but the same log-log convergence when considering the progress rate per evaluation. We could get positive experimental results even with the non-proved polynomial number of revaluations (non-adaptive); maybe those results are the most satisfactory (stable) results. We could also get convergence with adaptive

rules (number of resamplings depending on the step-size), however results are seemingly less stable than with non-adaptive methods.

2 Theoretical analysis: exponential non-adaptive rules can lead to log/log convergence.

Section 2.1 is devoted to some preliminaries. Section 2.2 presents results in the scale invariant case, for an exponential number of resamplings and non-adaptive rules. Section 2.3 will focus on adaptive rules, with numbers of resamplings depending on the step-size.

2.1 Preliminary: noise-free case

In the noise-free case, for some evolution strategies, we know the following results, almost surely (see e.g. Theorem 4 in [19], where, however, the negativity of the constant is not proved and only checked by Monte-Carlo simulations): $\log(\sigma_n)/n$ converges to some constant $(-A) < 0$ and $\log(||x_n||)/n$ converges to some constant $(-A') < 0$.

This implies that for any $\rho < A$, $\log(\sigma_n) \leq -\rho n$ for n sufficiently large. So, $\sup_{n\geq 1} \log(\sigma_n) + \rho n$ is finite. With these almost sure results, now consider V the quantile $1 - \delta/4$ of $\exp\left(\sup_{n\geq 1} \log(\sigma_n) + \rho n\right)$. Then, with probability at least $1 - \delta/4$, $\forall n \geq 1, \sigma_n \leq V \exp(-\rho n)$. We can apply the same trick for lower bounding σ_n, and upper and lower bounding $||x_n||$, all of them with probability $1 - \delta/4$, so that all bounds hold true simultaneously with probability at least $1 - \delta$.
Hence, for any $\alpha < A'$, $\alpha' > A'$, $\rho < A$, $\rho' > A$, there exist $C > 0$, $C' > 0$, $V > 0$, $V' > 0$, such that with probability at least $1 - \delta$

$$\forall n \geq 1, \quad C' \exp(-\alpha' n) \leq ||x_n|| \leq C \exp(-\alpha n); \tag{1}$$

$$\forall n \geq 1, \quad V' \exp(-\rho' n) \leq \quad \sigma_n \quad \leq V \exp(-\rho n). \tag{2}$$

We will first show, in Section 2.2, our noisy optimization result (Theorem 1):

(i) in the scale invariant case
(ii) using Eq. 1 (supposed to hold in the noise-free case)

We will then show similar results in Section 2.3:

(i) without scale-invariance
(ii) using Eq. 2 (supposed to hold in the noise-free case)
(iii) with other resamplings schemes

2.2 Scale invariant case, with exponential number of resamplings

We consider Alg. 1, a version of multi-membered Evolution Strategies, the (μ,λ)-ES. μ denotes the number of parents and λ the number of offspring ($\mu \leq \lambda$). In every generation, the selection takes place among the λ offspring, produced from a population of μ parents. Selection is based on the ranking of the individuals

according their $fitness(\cdot)$ taking the μ best individuals among the population. Here x_n denotes the parent at iteration n.

Algorithm 1 An evolution strategy, with exponential number of resamplings. If we consider $K = 1$ and $\zeta = 1$ we obtain the case without resampling. $\mathcal{N}$ is an arbitrary random variable with bounded density (each use is independent of others).

Parameters: $K > 0$,$\zeta \geq 0$, $\lambda \geq \mu > 0$, a dimension $d > 0$.
Input: an initial $x_1 \in \mathbb{R}^d$ and an initial $\sigma_0 > 0$.
$n \leftarrow 1$
while (true) **do**
  Generate λ individuals $i_1, \ldots, i_\lambda$ independently using
  $$i_j = x_n + \sigma_{n,j}\mathcal{N}. \tag{3}$$
  Evaluate each of them $r_n = \lceil K\zeta^n \rceil$ times and average their fitness values.
  Select the μ best individuals $j_1, \ldots, j_\mu$.
  Update: from x, σ_n, $i_1, \ldots, i_\lambda$ and $j_1, \ldots, j_\mu$, compute x_{n+1} and σ_{n+1}.
  $n \leftarrow n + 1$
end while

We now state our first theorem, under log-linear convergence assumption (the assumption in Eq. 5 is just Eq. 1).

Theorem 1. *Consider the fitness function*
$$f(z) = ||z||^p + \mathcal{N} \tag{4}$$
over $\mathbb{R}^d$ *and* $x_1 = (1, 1, \ldots, 1)$.
Consider an evolution strategy with population size λ*, parent population size* μ*, such that without resampling, for any* $\delta > 0$*, for some* $\alpha > 0$*,* $\alpha' > 0$*, with probability* $1 - \delta/2$*, with objective function* $fitness(x) = ||x||$*,*
$$\exists C, C'; \qquad C' \exp(-\alpha' n) \leq ||x_n|| \leq C \exp(-\alpha n). \tag{5}$$
Assume, additionally, that there is scale invariance:
$$\sigma_n = C''||x_n|| \tag{6}$$
for some $C'' > 0$.
Then, for any $\delta > 0$*, there is* $K_0 > 0, \zeta_0 > 0$ *such that for* $K \geq K_0$*,* $\zeta > \zeta_0$*, Eq. 1 also holds with probability at least* $1 - \delta$ *for fitness function as in Eq. 4 and resampling rule as in Alg. 1.*
Remarks: *(i) Informally speaking, our theorem shows that if a scale invariant algorithm converges in the noise-free case, then it also converges in the noisy case with the exponential resampling rule, at least if parameters are large enough (a similar effect of constants was pointed out in [4] in a different setting).*
(ii) We assume that the optimum is in 0 *and the initial* x_1 *at* 1*. Note that these assumptions have no influence when we use algorithms invariant by rotation and translation.*
(iii) We show a log-linear convergence rate as in the noise-free case, but at the cost of more evaluations per iteration. When normalized by the number of

function evaluations, we get $\log ||x_n||$ *linear in the logarithm of the number of function evaluations, as detailed in Corollary 1.*

Proof of the theorem: In all the proof, $\mathcal{N}$ denotes a standard Gaussian random variable (depending on the context, in dimension 1 or d). Consider an arbitrary $\delta > 0$, $n \geq 1$ and $\delta_n = \exp(-\gamma n)$ for some $\gamma > 0$.

Define p_n the probability that two generated points, e.g. i_1 and i_2, are such that $\mid ||i_1||^p - ||i_2||^p \mid \leq \delta_n$.

Step 1: Using Eq. 3 and Eq. 6, we show that

$$p_n \leq B' \exp(-\gamma' n) \tag{7}$$

for some $B' > 0, \gamma' > 0$ depending on γ, d, p, C', C'', α'.

Proof of step 1: with $\mathcal{N}_1$ and $\mathcal{N}_2$ two d-dimensional independent standard Gaussian random variables,

$$p_n \leq P(\mid ||1 + C''\mathcal{N}_1||^p - ||1 + C''\mathcal{N}_2||^p \mid \leq \delta_n / ||x_n||^p). \tag{8}$$

Define $densityMax$ the supremum of the density of $\mid ||1+C''\mathcal{N}_1||^p - ||1+C''\mathcal{N}_2||^p \mid$ we get

$$p_n \leq densityMax C'^{-p} \exp((p\alpha' - \gamma)n),$$

hence the expected result with $\gamma' = \gamma - p\alpha'$ and $B' = densityMax(C')^{-p}$. Notice that $densityMax$ is upper bounded.

In particular, γ' is arbitrarily large, provided that γ is sufficiently large.

Step 2: Consider now $p_n^{(1)}$ the probability that there exists i_1 and i_2 such that $\mid ||i_1||^p - ||i_2||^p \mid \leq \delta_n$. Then, $p_n^{(1)} \leq \lambda^2 p_n \leq B' \lambda^2 \exp(-\gamma' n)$.

Step 3: Consider now $p_n^{(2)}$ the probability that $|\mathcal{N}/\sqrt{K\zeta^n}| \geq \delta_n/2$. First, we write $p_n^{(2)} = P(\mathcal{N} \geq \frac{\delta_n}{2}\sqrt{K\zeta^n})$. So by Chebychev inequality, $p_n^{(2)} \leq B'' \exp(-\gamma'' n)$ for $\gamma'' = \log(\zeta) - 2\gamma$ arbitrarily large, provided that ζ is large enough, and $B'' = 4/K$.

Step 4: Consider now $p_n^{(3)}$ the probability that $|\mathcal{N}/\sqrt{K\zeta^n}| \geq \delta_n/2$ at least once for the λ evaluated individuals of iteration n. Then, $p_n^{(3)} \leq \lambda p_n^{(2)}$.

Step 5: In this step we consider the probability that two individuals are misranked due to noise. Let us now consider $p_n^{(4)}$ the probability that at least two points i_a and i_b at iteration n verify

$$||i_a||^p \leq ||i_b||^p \tag{9}$$

$$\text{and} \qquad noisyEvaluation(i_a) \geq noisyEvaluation(i_b) \tag{10}$$

where $noisyEvaluation(i)$ is the average of the multiple evaluations of individual i. Eqs. 9 and 10 occur simultaneously if either two points have very similar fitness (difference less than δ_n) or the noise is big (larger than $\delta_n/2$). Therefore, $p_n^{(4)} \leq p_n^{(1)} + p_n^{(3)} \leq \lambda^2 p_n + \lambda p_n^{(2)}$ so $p_n^{(4)} \leq (B' + B'')\lambda^2 \exp(-\min(\gamma', \gamma'')n)$.

Step 6: Step 5 was about the probability that at least two points at iteration n are misranked due to noise. We now consider $\sum_{n\geq 1} p_n^{(4)}$, which is an upper bound on the probability that in at least one iteration there is a misranking of two individuals.

If γ' and γ'' are large enough, $\sum_{n\geq 1} p_n^{(4)} < \delta$.

This implies that with probability at least $1 - \delta$, provided that K and ζ have

been chosen large enough for γ and γ' to be large enough, we get the same rankings of points as in the noise free case - this proves the expected result. □
The following corollary shows that this is a log-log convergence.

Corollary 1: log-log **convergence with exponential resampling.** *With e_n the number of evaluations at the end of iteration n, we have $e_n = K\zeta\frac{\zeta^n-1}{\zeta-1}$. We then get, from Eq. 1,*

$$\log(||x_n||)/\log(e_n) \to -\frac{\alpha}{\log\zeta} \tag{11}$$

with probability at least $1-\delta$. Eq. 11 is the convergence in log/log *scale.*
We have shown this property for an exponentially increasing number of resamplings, which is indeed similar to R-EDA[18], which converges with a small number of iterations but with exponentially many resamplings per iteration. In the experimental section 3, we will check what happens in the polynomial case.

2.3 Extension: adaptive resamplings and removing the scale invariance assumption

We have assumed above a scale invariance. This is obviously not a nice feature of our proof, because scale invariance does not correspond to anything real; in a real setting we do not know the distance to the optimum. We show below an extension of the result above using the assumption of a log-linear convergence of σ_n as in Eq. 2 instead of the scale invariance used before.
In the corollary below, we also get rid of the non-adaptive rule with exponential number of resamplings, replaced by a number of resamplings depending on the step-size σ_n only, as in Eq. 2. In one corollary, we switch to both (i) adaptive resampling rule and (ii) no scale invariance; each change can indeed be proved independently of the other.

Algorithm 2 An evolution strategy, with number of resamplings polynomial in the step-size. The case without resampling means $Y = 1$ and $\eta = 0$. $\mathcal{N}$ is an arbitrary random variable with bounded density (each use is independent of others).

Parameters: $Y > 0$,$\eta \geq 0$, $\lambda \geq \mu > 0$, a dimension $d > 0$.
Input: an initial $x_1 \in \mathbb{R}^d$ and an initial $\sigma_0 > 0$.
$n \leftarrow 1$
while (true) **do**
  Generate λ individuals $i_1, \ldots, i_\lambda$ independently using

$$i_j = x_n + \sigma_{n,j}\mathcal{N}. \tag{12}$$

  Evaluate each of them $r_n = \lceil Y{\sigma_n}^{-\eta}\rceil$ times and average their fitness values.
  Select the μ best individuals $j_1, \ldots, j_\mu$.
  Update: from x, σ_n, $i_1, \ldots, i_\lambda$ and $j_1, \ldots, j_\mu$, compute x_{n+1} and σ_{n+1}.
  $n \leftarrow n + 1$
end while

Corollary 2: adaptive resampling and no scale-invariance. *The proof of Theorem [1] also holds without scale invariance, under the following assumptions:*

- *For any $\delta > 0$, there are constants $\rho > 0, V > 0, \rho' > 0, V' > 0$ such that with probability at least $1-\delta$, Eq. 2 holds.*

– *The number of revaluations is*

$$Y\left(\frac{1}{\sigma_n}\right)^{\eta} \tag{13}$$

with Y and η sufficiently large.

– *Individuals are still randomly drawn using $x_n + \sigma_n \mathcal{N}$ for some random variable $\mathcal{N}$ with bounded density.*

Remark: *This setting is useful in cases like self-adaptive algorithms, in which we do not use directly a Gaussian random variable, but a Gaussian random variable multiplied e.g. by $\exp(\frac{1}{\sqrt{d}})Gaussian$, with Gaussian a standard Gaussian random variable. For example, SA-ES algorithms as in [19] are included in this proof because they converge log-linearly as explained in Section 2.1.*

Proof of corollary 2: Two steps of the proof are different, namely step 1 and step 2. We here adapt the proofs of these two steps.

Adapting step 1: Eq. 8 becomes Eq. 14:

$$p_n \leq P(\mid ||1 + C''_n \mathcal{N}_1||^p - ||1 + C''_n \mathcal{N}_2||^p \mid \leq \delta_n / ||x_n||^p). \tag{14}$$

where $C''_n = \sigma_n / ||x_n|| \geq t' \exp(-tn)$ for some $t > 0, t' > 0$ depending on ρ, ρ', V, V' only. Eq. 14 leads to

$$p_n \leq (C''_n)^{-d} densityMax C'^{-p} \exp((p\alpha' - \gamma)n),$$

hence the expected result with $\gamma' = \gamma - p\alpha' - dt$. $densityMax$ is upper bounded due to the third condition of corollary 2.

Adapting step 2: It is sufficient to show that the number of resamplings is larger (for each iteration) than in the Theorem 1, so that step 2 still holds.
Eq. 13 implies that the number of revaluations at step n is at least $Y\left(\frac{1}{V}\right)^{\eta} \exp(\rho\eta n)$. This is more than $K\zeta^n$, at least if Y and η are large enough. This leads to the same conclusion as in the Theorem 1, except that we have probability $1 - 2\delta$ instead of $1 - \delta$ (which is not a big issue as we can do the same with $\delta/2$). □
The following corollary is here for showing that our result leads to the log-log convergence.

Corollary 3: log-log **convergence for adaptive resampling.** *With e_n the number of evaluations at the end of iteration n, we have $e_n = Y\left(\frac{1}{V}\right)^{\eta} \exp(\rho\eta) \frac{\exp(\rho\eta n) - 1}{\exp(\rho\eta) - 1}$. We then get, from Eq. 1,*

$$\log(||x_n||) / \log(e_n) \to -\frac{\alpha}{\rho\eta} \tag{15}$$

with probability at least $1 - \delta$. Eq. 15 is the convergence in log/log *scale.*

3 Polynomial number of resamplings: experiments

We here consider a polynomial number of resamplings, as in Alg. 3.

Algorithm 3 An evolution strategy, with polynomial number of resamplings. The case without resampling means $K = 1$ and $\zeta = 0$.

Parameters: $K > 0$,$\zeta \geq 0$, $\lambda \geq \mu > 0$, a dimension $d > 0$.
Input: an initial $x_1 \in \mathbb{R}^d$ and an initial $\sigma_0 > 0$.
$n \leftarrow 1$
while (true) **do**
  Generate λ individuals $i_1, \ldots, i_\lambda$ independently using

$$\sigma_{n,j} = \sigma_n \times \exp(\frac{1}{\sqrt{d}}\mathcal{N})$$
$$i_j = x_n + \sigma_{n,j}\mathcal{N}. \quad (16)$$

  Evaluate each of them $r_n = \lceil Kn^\zeta \rceil$ times and average their fitness values.
  Select the μ best individuals $j_1, \ldots, j_\mu$.
  Update: from x, σ_n, $i_1, \ldots, i_\lambda$ and $j_1, \ldots, j_\mu$, compute x_{n+1} and σ_{n+1}.
  $n \leftarrow n + 1$
end while

Experiments are performed in a "real" setting, without scale invariance. Importantly, our mathematical results hold only log-log convergence under the assumption that constants are large enough. We present results with fitness function $f(x) = ||x||^p + \mathcal{N}$ with $p = 2$ in Fig. 1.

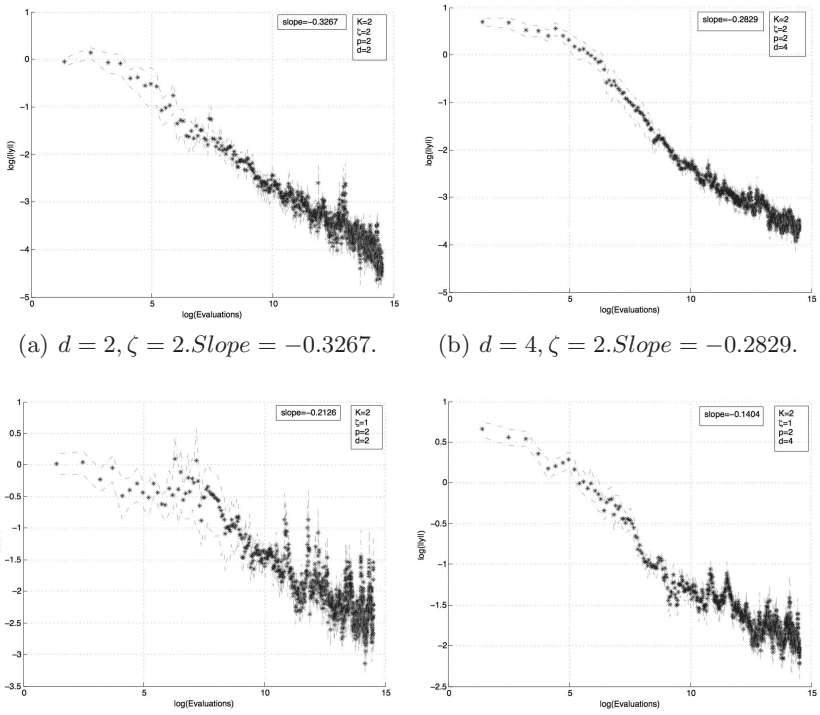
In experiments with the following parameters (as recommended in [10, 20]): $p = 1$ or $p = 4$, dimension $2, 3, 4, 5$, $\zeta = 1, 2, 3$, $\mu = \min(d, \lceil \lambda/4 \rceil)$, $\lambda = \lceil d\sqrt{d} \rceil$, slopes are usually better than $-1/(2p)$ for $\zeta = 2$ or $\zeta = 3$ and worse for $\zeta = 1$. Non-presented experiments show that $\zeta = 0$ performs very poorly. Seemingly results for ζ large are farther from the asymptotic regime. We conjecture that the asymptotic regime is $-1/(2p)$ but that it is reached later when ζ is large. R-EDA[18] reaches $-1/(2p)$; we seemingly get slightly better but this might be due to a non-asymptotic effect. Fig. 1 provides results with high numbers of evaluations.

4 Experiments with adaptivity: $Y\sigma_n^{-\eta}$ revaluations

We here show experiments with Alg. 2. The algorithm should converge linearly in log-log scale as shown by Corollary 3, at least for large enough values of Y and η. Notice that we consider values of μ, λ for which log-linear convergence is proved in the noise-free setting (see Section 2.1).In all this section, $\mu = \min(d, \lceil \lambda/4 \rceil)$, $\lambda = \lceil d\sqrt{d} \rceil$.

Slopes as estimated on the case $\eta = 2$ (usually the most favorable, and an important case naturally arising in sufficiently differentiable problems) are given in Table 1 (left) for dimension $d = 100$. In this case we are far from the asymptotic regime.

We get results close to $-\frac{1}{2}$ in all cases This slope of $-\frac{1}{2}$ is reachable by algorithms which learn a model of the fitness function, as e.g. [7]. In this case of high dimension we are far from the slope $1/(-2p)$, which might be the case for the asymptotic results. This is suggested by experiments in dimension 10



(a) $d = 2, \zeta = 2. Slope = -0.3267$. (b) $d = 4, \zeta = 2. Slope = -0.2829$.

(c) $d = 2, \zeta = 1. Slope = -0.2126$. (d) $d = 4, \zeta = 1. Slope = -0.1404$.

Fig. 1: Experiments in dimension 2, 3, 4 with $\zeta = 1, 2$ (number of evaluations shown by x-axis) for $r_n = K\lceil n^\zeta \rceil$ (i.e. polynomial, non-adaptive) with $\mu = 2$, $\lambda = 4$, $p = 2$ and $K = 2$. The slope is evaluated on the second half of the iterations. We get slopes close to $-1/(2p)$. All results are averaged over 20 runs.

summarized in Table 1 (right). We also point out that the known complexity bounds is $-\frac{1}{p}$ (from [9]), and maybe the slope can reach $-\frac{1}{p}$ in some cases.
Results with $Y\left(\frac{1}{\sigma}\right)^\eta$ are moderately stable (impact of Y, in particular). This supports our preference for stable rules, such as non-adaptively choosing n^2 revaluations per individual at iteration n.

5 Conclusion

We have shown mathematically log-log convergence results and studied experimentally the slope in this convergence. These results were shown for evolution strategies, which are known for having good uniform rates, rather than good non-uniform rates. We summarize these two parts below and give some research directions.

Log-log convergence. We have shown that the log-log convergence (i.e. linear convergence with x-axis the log of the number of evaluations and y-axis the log of the distance to the optimum) occurs in various cases:

Table 1: **Left: Dimension 100.** Estimated slope for the adaptive rule with $r_n = \lceil \left(\frac{1}{\sigma_n}\right)^2 \rceil$ resamplings at iteration n. Slopes are estimated on the second half of the curve. **Right: Dimension 10.** Estimated slope for the adaptive rule with $r_n = \lceil Y \left(\frac{1}{\sigma_n}\right)^2 \rceil$ resamplings at iteration n ($Y = 1$ as in previous curves, and $Y = 20$ for checking the impact of convergence; the negative slope (apparent convergence) for $Y = 20$ is stable, as well as the divergence or stagnation for $Y = 1$ for $p = 4$). Slopes are estimated on the second half of the curve.

$d = 100$	
p	slope for $Y = 1$
1	-0.52
2	-0.51
4	-0.45

$d = 10$		
p	slope for $Y = 1$	slope for $Y = 20$
1	-0.51	-0.50
2	-0.18	-0.17
4	>0	-0.08

- Non-adaptive rules, with number of resamplings exponential in the iteration counter. Here we have a mathematical proof, which includes the assumption of scale invariance; as shown by Corollary 2, this can be extended to non scale-invariant algorithms;
- Adaptive rules, with number of resamplings polynomial in $1/\sigma_n$ with σ_n the step-size. Here we have a mathematical proof; however, there is a strong sensitivity to constants Y and η which participate in the number of resamplings per individual, $Y \left(\frac{1}{\sigma_n}\right)^\eta$;
- Non-adaptive rule, with polynomial number of resamplings. This case is a quite convenient scheme experimentally but we have no proof.

Slope in log-log convergence. Experimentally, the best slope in the log-log representation is often close to $-\frac{1}{2p}$ for fitness function $f(x) = ||x||^p + \mathcal{N}$. It is known that under modeling assumptions (i.e. the function is regular enough for being optimized by learning), it is possible to do better than that (the slope becomes $-1/2$ for parametric cases, see [7] and references therein), but $-\frac{1}{2p}$ is the best known exponent under locality assumption. Basically, locality assumption ensures that most points are reasonably good, whereas some specialized noisy optimization algorithms sample a few very good points and essentially sample individuals far from the optimum (see e.g. [7]).

Further work. The main further work is the mathematical analysis of the polynomial number of resamplings in the non-adaptive case. Also, a combination of adaptive and non-adaptive rules might be interesting; adaptive rules are intuitively satisfactory, but non-adaptive polynomial rules provide simple efficient solutions, with empirically easy (no tuning) results. If our life depended on a scheme, we would for the moment choose a simple polynomial rule with a number of revaluations quadratic in the number of evaluations, in spite of (maybe) moderate elegance due to lack of adaptivity.

References

1. Morales, S.A., Liu, J., Teytaud, O.: Noisy optimization convergence rates. In: GECCO (Companion). (2013) 223–224
2. Jones, D., Schonlau, M., Welch, W.: (Efficient global optimization of expensive black-box functions)
3. Auger, A., Jebalia, M., Teytaud, O.: Xse: quasi-random mutations for evolution strategies. In: Proceedings of Evolutionary Algorithms, 12 pages. (2005)
4. Jebalia, M., Auger, A., Hansen, N.: Log linear convergence and divergence of the scale-invariant (1+1)-ES in noisy environments. Algorithmica (2010)
5. Arnold, D.V., Beyer, H.G.: A general noise model and its effects on evolution strategy performance. IEEE Transactions on Evolutionary Computation **10** (2006) 380–391
6. Finck, S., Beyer, H.G., Melkozerov, A.: Noisy optimization: a theoretical strategy comparison of es, egs, spsa & if on the noisy sphere. In: GECCO. (2011) 813–820
7. Coulom, R.: Clop: Confident local optimization for noisy black-box parameter tuning. In: Advances in Computer Games. Springer Berlin Heidelberg (2012) 146–157
8. Coulom, R., Rolet, P., Sokolovska, N., Teytaud, O.: Handling expensive optimization with large noise. In: Foundations of Genetic Algorithms. (2011)
9. Teytaud, O., Decock, J.: Noisy Optimization Complexity. In: FOGA - Foundations of Genetic Algorithms XII - 2013, Adelaide, Australie (2013)
10. Beyer, H.G.: The Theory of Evolution Strategies. Natural Computing Series. Springer, Heideberg (2001)
11. Yang, X., Birkfellner, W., Niederer, P.: Optimized 2d/3d medical image registration using the estimation of multivariate normal algorithm (EMNA). In: Biomedical engineering. (2005)
12. Anderson, E.J., Ferris, M.C.: A direct search algorithm for optimization with noisy function evaluations. SIAM Journal on Optimization **11** (2001) 837–857
13. Lucidi, S., Sciandrone, M.: A derivative-free algorithm for bound constrained optimization. Comp. Opt. and Appl. **21** (2002) 119–142
14. Kim, S., Zhang, D.: Convergence properties of direct search methods for stochastic optimization. In: Proceedings of the Winter Simulation Conference. WSC '10, Winter Simulation Conference (2010) 1003–1011
15. Hansen, N., Niederberger, S., Guzzella, L., Koumoutsakos, P.: A method for handling uncertainty in evolutionary optimization with an application to feedback control of combustion. IEEE Transactions on Evolutionary Computation **13** (2009) 180–197
16. Villemonteix, J., Vazquez, E., Walter, E.: An informational approach to the global optimization of expensive-to-evaluate functions. Journal of Global Optimization (2008) 26 pages
17. Fabian, V.: Stochastic Approximation of Minima with Improved Asymptotic Speed. Annals of Mathematical statistics **38** (1967) 191–200
18. Rolet, P., Teytaud, O.: Bandit-based estimation of distribution algorithms for noisy optimization: Rigorous runtime analysis. In: Proceedings of Lion4 (accepted); presented in TRSH 2009 in Birmingham. (2009) 97–110
19. Auger, A.: (Convergence results for (1,λ)-SA-ES using the theory of φ-irreducible Markov chains)
20. Fournier, H., Teytaud, O.: Lower bounds for comparison based evolution strategies using vc-dimension and sign patterns. Algorithmica (2010)