Reconsidering the Progress Rate Theory for Evolution Strategies in Finite Dimensions

Anne Auger
CoLab Computational Laboratory
ETH Zurich, Switzerland
Anne.Auger@inf.ethz.ch

Nikolaus Hansen
CoLab Computational Laboratory
ETH Zurich, Switzerland
Nikolaus.Hansen@inf.ethz.ch

ABSTRACT

This paper investigates the limits of the predictions based on the classical progress rate theory for Evolution Strategies. We explain on the sphere function why positive progress rates give convergence in mean, negative progress rates divergence in mean and show that almost sure convergence can take place despite divergence in mean. Hence step-sizes associated to negative progress can actually lead to almost sure convergence. Based on these results we provide an alternative progress rate definition related to almost sure convergence. We present Monte Carlo simulations to investigate the discrepancy between both progress rates and therefore both types of convergence. This discrepancy vanishes when dimension increases. The observation is supported by an asymptotic estimation of the new progress rate definition.

Categories and Subject Descriptors: G.1.6 [Numerical Analysis]: Optimization—*Global optimization, Unconstrained optimization*; F.2.1[Analysis of Algorithms and Problem Complexity]:Numerical Algorithms and Problems

General Terms: Algorithms, Theory

Keywords: Evolution Strategy, theory, progress rate, convergence rate

1. INTRODUCTION

Since the introduction of Evolution Strategies (ES) in the mid-sixties, theoretical investigations [3, 6] mainly focused on investigating the so-called progress rate defined as the expected progress from one step of the algorithm to the next. In case of a spherical fitness function[1] the normalized progress rate obeys

$$\varphi^* = d\,\mathbb{E}\left(\frac{\|X_n\| - \|X_{n+1}\|}{\|X_n\|}\,\middle|\, X_n\right) \ , \qquad (1)$$

where d is the dimension of the search space, n the discrete time index and $(X_n)_{n\in\mathbb{N}}$ the sequence of random variables modeling the algorithms evolution in search space that is assumed to take non-zero values except on a set of probability zero. Most estimations given for the progress rate φ^* are asymptotic in the dimension d.[2]

We give a simple example, where the progress definition of Eq. 1 fails to reflect what is actually observed in reality. Consider a sequence of X_n where the search algorithm is stepping forward and backward, and forward and backward,..., always by the same amount respectively; say $\|X_0\| = 1$, $\|X_1\| = 0.5$, $\|X_2\| = 1$, $\|X_3\| = 0.5$, $\|X_4\| = 1$,... Consequently $(\|X_n\| - \|X_{n+1}\|)/\|X_n\|$ equals to $+0.5$ for even n and -1 otherwise, while, in contrary, the same distance is covered in even and odd steps. The resulting negative progress value for even time steps is, in absolute terms, greater than the resulting positive progress value for odd time steps, suggesting divergent behavior of the sequence. Divergence stands in contrast to $\|X_0\| - \|X_n\| \in \{0, 0.5\} \geq 0$ for all time steps n. In this paper we will see that the defect in φ^*, observed in this simple example, carries over to stochastic random sequences $\|X_n\|$ (even though $\|X_n\|$ is deterministic in our example, X_n can be stochastic) and also applies to the evolution strategy. The defect can be resolved by using different means of analysis leading to a new progress definition.

When analyzing the convergence of a sequence of random variables several notions of convergence exist. In this paper we will focus on

1. convergence in mean, associated to the expected progress rate

   A sequence of random variable $(X_n)_{n\in\mathbb{N}}$ converge in mean to X if

   $$\lim_{n\to\infty} \mathbb{E}(|X_n - X|) = 0 \ .$$

2. almost sure convergence, which is more desirable to reflect what is observed in single instances.

   A sequence of random variables $(X_n)_{n\in\mathbb{N}}$ converges almost surely or with probability one to the random variable X if:

   $$\mathbb{P}\{\lim_{n\to\infty} X_n = X\} = 1$$

[1] A spherical (isotropic) fitness function $f : \mathbb{R}^d \to \mathbb{R}$ can be expressed as $f(x) = g(x^T x)$, where $g : \mathbb{R} \to \mathbb{R}$ is strictly increasing. Surfaces of equal fitness of f are hyperspheres.

[2] The conditioning in Eq. 1 should not be omitted in general. However, we will see that for the so-called scale-invariant algorithm the conditional expectation in the right-hand side of Eq. 1 is equal to the (unconditional) expectation (Lemma 1).

GECCO'06, July 8–12, 2006, Seattle, Washington, USA.

which means that the events for which X_n does not converge to X have probability 0. Under these conditions we use the notation

$$\lim_{n\to\infty} X_n = X \text{ a.s.}$$

For the investigation of different types of convergences the reader is also referred to [7].

Positive progress, as defined in Eq. 1, and convergence in mean are equivalent for the following algorithm[3] minimizing $f(x) = \|x\|^2$ [1]:

$$X_{n+1} = \arg\min_{1\leq i\leq\lambda} \left\{ \|X_n + \frac{\sigma^* \|X_n\|}{d} N^i(0, I_d)\|^2 \right\} \quad , \qquad (2)$$

where $\sigma^* \in \mathbb{R}^+$ and $N^i(0, I_d)$ for $i = 1 \ldots \lambda$ are λ independent Gaussian random vectors with mean zero and covariance matrix I_d. This algorithm is related to the $(1, \lambda)$-evolution strategy and the natural question of convergence related to convergence in mean or almost sure convergence arises.

The organization of the paper is the following. In Section 2 we construct a simple example to illustrate the discrepancy between convergence in mean and almost sure convergence. Section 3 explains the link between convergence in mean and maximization of the progress rate and analyzes convergence in mean and almost sure convergence of the algorithm from Eq. 2. From the different convergence rates (associated to different types of convergence), a new progress rate definition related to almost sure convergence is derived. Numerical investigations of the different convergence rates are presented in Section 4. Finally we derive in Section 5 asymptotic approximations for the new progress rate definition. Section 6 discusses the practical relevance of our results and Section 7 gives a summary and conclusion.

2. AN INTRODUCTORY EXAMPLE

We start by constructing a simple example of a sequence of random variables $(X_n)_{n\in\mathbb{N}} \in \mathbb{R}^+$ to illustrate how predictions based on the expectation $\mathbb{E}(X_n)$ can give the wrong intuition of what actually occurs during one simulation.

Let $X_0 \in \mathbb{R}^+\backslash\{0\}$ and let the random sequence X_n be recursively defined by

$$X_{n+1} = X_n Y_\alpha \quad , \qquad (3)$$

where Y_α is an independent random variable with parameter $\alpha \in \mathbb{R}^+\backslash\{0\}$. As an example we take for Y_α a log-normal distribution multiplied by α:

$$X_{n+1} = X_n \, \alpha \exp\left(N_n(0, 1)\right) \quad , \qquad (4)$$

where $N_n(0, 1)$ are independent normal random variables with mean 0 and standard deviation 1. Since $\mathbb{E}\left(\exp\left(N_n(0, 1)\right)\right) = \exp(1/2)$ we have

$$\mathbb{E}\left(Y_\alpha\right) = \alpha \exp(1/2)$$

and since $\ln(Y_\alpha) = \ln(\alpha) + N_n(0, 1)$ we have

$$\mathbb{E}\left(\ln\left(Y_\alpha\right)\right) = \ln(\alpha) \quad .$$

[3]For convenience and brevity we will simply use the term *algorithm* for a sequence of random variables that models an algorithm running on a function to be minimized.

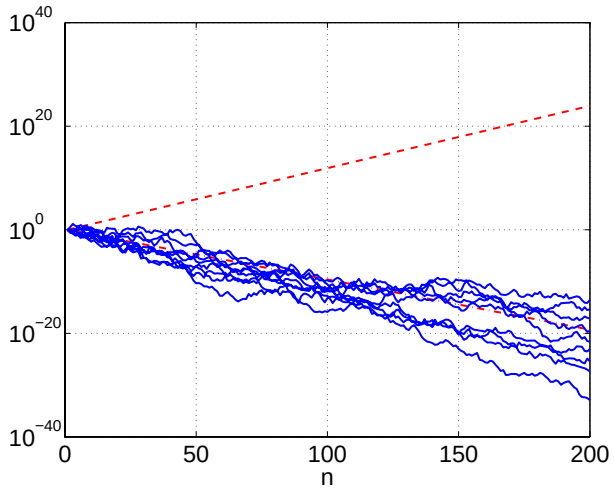

Figure 1: Ten realizations of the random sequence defined in Eq. 4 with $\alpha = 0.8$ and $X_0 = 1$. The dashed line represents $\mathbb{E}(X_n) = (\alpha \exp(1/2))^n$, the dashed-dotted line (within the set of curves) depicts $\exp(\mathbb{E}(\ln X_n)) = \alpha^n$. We can observe convergence of those 10 samples despite the divergence in mean.

Divergence in mean, almost sure convergence. By taking the expectation of Eq. 4 we obtain that

$$\mathbb{E}(X_{n+1}) = \mathbb{E}(X_n) \, \alpha \exp(1/2) \quad . \qquad (5)$$

This equation shows that for $\alpha > 1/\sqrt{e} \approx 0.6065$ the sequence of random variables $(X_n)_{n\in\mathbb{N}}$ diverges in mean. However for $\alpha = 0.8 > 1/\sqrt{e}$ we observe that samples of the sequence of random variables converge to zero as illustrated in **Fig. 1**. We precise in the following proposition when convergence in mean and almost sure convergence take place for the sequence defined with Eq. 4.

PROPOSITION 1. *Let $(X_n)_{n\in\mathbb{N}}$ be the random sequence defined with Eq. 4 with $\mathbb{E}(X_0) < +\infty$.*

1. *The sequence $\ln(X_n)$ is a random walk, i.e.*

$$\ln(X_{n+1}) = \ln(X_n) + \ln(\alpha) + N_n(0, 1) \quad . \qquad (6)$$

2. *Almost surely*

$$\lim_{n\to\infty} \frac{1}{n} \left(\ln(X_n) - \ln(X_0)\right) = \ln(\alpha) \qquad (7)$$

3. *In mean*

$$\mathbb{E}(X_n) = \mathbb{E}(X_0) \left(\alpha \exp(1/2)\right)^n \qquad (8)$$

PROOF: By taking the logarithm of Eq. 4 we obtain Eq. 6. Equation 8 results from iterating Eq. 5. For the almost sure convergence, we sum both sides of Eq. 6 and obtain:

$$\sum_{k=0}^{n-1} \ln(X_{k+1}) = \sum_{k=0}^{n-1} \ln(X_k) + n \ln(\alpha) + \sum_{k=0}^{n-1} N_k(0, 1)$$

which simplifies to

$$\ln(X_n) = \ln(X_0) + n \ln(\alpha) + \sum_{k=0}^{n-1} N_k(0, 1)$$

and dividing by n yields

$$\frac{1}{n}\left(\ln(X_n) - \ln(X_0)\right) = \ln(\alpha) + \frac{1}{n}\sum_{k=0}^{n-1} N_k(0,1) \ .$$

From the Strong Law of Large Numbers we have that almost surely

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} N_k(0,1) = \mathbb{E}(N(0,1)) = 0 \ .$$

Equation 7 follows. □

From Proposition 1 follows that for α such that $\alpha\exp(1/2)$ is greater than 1 and $\ln(\alpha)$ is negative, divergence in mean occurs though almost sure convergence takes place, which is the case illustrated in **Fig. 1**.

COROLLARY 1 . *For $1/\sqrt{e} < \alpha < 1$ there is almost sure convergence of $(X_n)_{n\in\mathbb{N}}$ and divergence in mean.*

For $\alpha < 1/\sqrt{e}$ and $\alpha > 1$ the results almost sure versus in mean agree.

3. NEGATIVE PROGRESS BUT CONVERGENCE TOWARDS ZERO

In this section we prove that the previous phenomenon can also be observed with the progress rate approach for Evolution Strategies. For step-sizes associated to a negative progress (defined in expectation), almost sure convergence of the algorithm can occur.

3.1 Equivalence between progress rate and convergence of the scale-invariant algorithm

We consider a $(1,\lambda)$-ES, a simple non-elitist strategy where negative progress exists. The fitness function that we consider is the so-called d-dimensional sphere function

$$f := (x_1,\ldots,x_d) \in \mathbb{R}^d \to \|x\|^2 = \sum_{i=1}^{d} x_i^2 \ ,$$

to be minimized. At each generation n a parent, X_n, creates λ offspring with the so-called *mutation operator*

$$X_n^i = X_n + \sigma N^i(0, I_d) \quad \text{for } i = 1,\ldots,\lambda \ ,$$

where $\sigma \in \mathbb{R}^+$ is the step-size and $\left(N^i(0,I_d)\right)_{1\leq i\leq\lambda}$ are λ independent instances of a Gaussian random variable with zero mean and identity covariance matrix. The best offspring is selected to become the next parent X_{n+1}. One iteration of the ES running on the sphere function can be summarized by the following equation:

$$X_{n+1} = \arg\min_{1\leq i\leq\lambda}\left\{\|X_n + \sigma N^i(0,I_d)\|^2\right\} \ . \qquad (9)$$

Due to the invariance of the ES against order preserving transformations of the fitness values, the previous equation is equivalent to

$$X_{n+1} = \arg\min_{1\leq i\leq\lambda}\left\{\|X_n + \sigma N^i(0,I_d)\|\right\} \ . \qquad (10)$$

We now introduce the notation $N^*(0,I_d)$ to denote the selected random variable, *i.e.*

$$X_{n+1} = X_n + \sigma N^*(0,I_d) \ . \qquad (11)$$

The classical progress rate approach. The progress rate approach [3, 6] consists in looking for σ maximizing the progress rate defined in Eq. 1. This progress can be rewritten as

$$\varphi^* = d\left(1 - \mathbb{E}\left(\frac{\|X_{n+1}\|}{\|X_n\|}\middle| X_n\right)\right)$$

or, in case of the $(1,\lambda)$-ES on the sphere model, we can write

$$\varphi^* = d\left(1 - \mathbb{E}\left(\|\frac{X_n}{\|X_n\|} + \frac{\sigma}{\|X_n\|}N^*(0,I_d)\|\middle| X_n\right)\right) \ .$$

Usually the so-called normalized step-size $\sigma^* = \frac{\sigma d}{\|X_n\|}$ is introduced [3] and the approach consists in trying to find σ^* maximizing

$$\varphi^* = d\left(1 - \mathbb{E}\left(\|\frac{X_n}{\|X_n\|} + \frac{\sigma^*}{d}N^*(0,I_d)\|\middle| X_n\right)\right) \ . \qquad (12)$$

Actually the distribution of the random variable $\|\frac{X_n}{\|X_n\|} + \frac{\sigma^*}{d}N^*(0,I_d)\|$ does not depend on X_n and more specifically we have:

LEMMA 1. *The distribution of the random variable $\|\frac{X_n}{\|X_n\|} + \frac{\sigma^*}{d}N^*(0,I_d)\|$ does not depend on X_n and is the same as the distribution of*

$$\|e_1 + \frac{\sigma^*}{d}N^*(0,I_d)\| = \min_{1\leq i\leq\lambda}\left\{\|e_1 + \frac{\sigma^*}{d}N^i(0,I_d)\|\right\} \qquad (13)$$

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

This result has been for instance used in [4]. With Lemma 1 Eq. 12 simplifies to

$$\boxed{\varphi^* = d\left(1 - \mathbb{E}\left(\|e_1 + \frac{\sigma^*}{d}N^*(0,I_d)\|\right)\right) \ , \qquad (14)}$$

specifying the classical progress rate of the $(1,\lambda)$-ES on any unimodal isotropic function.

Convergence of a scale-invariant algorithm. We recall in this section how the progress rate given in Eq. 14 relates to the convergence of the scale-invariant algorithm $(X_n)_{n\in\mathbb{N}}$ [1] defined as:

$$\boxed{X_{n+1} = \arg\min_{1\leq i\leq\lambda}\left\{\|X_n + \frac{\sigma^*\|X_n\|}{d}N^i(0,I_d)\|\right\} \qquad (15)}$$

derived from Eq. 10 where the step-size σ is chosen at each step proportional to the norm of the parent, $\sigma = \frac{\sigma^*\|X_n\|}{d}$.

LEMMA 2. *For the algorithm defined in Eq. 15, the following holds*

$$\|X_{n+1}\| = \|X_n\|\|\frac{X_n}{\|X_n\|} + \frac{\sigma^*}{d}N^*(0,I_d)\| \qquad (16)$$

where the random variable $\frac{\|X_{n+1}\|}{\|X_n\|} = \|\frac{X_n}{\|X_n\|} + \frac{\sigma^*}{d}N^*(0,I_d)\|$ is independent of $\|X_n\|$ and is distributed as $\|e_1 + \frac{\sigma^*}{d}N^*(0,I_d)\|$ defined in Eq. 13.

PROOF: Taking the norm in both sides of Eq. 15 we have that

$$\|X_{n+1}\| = \min_{1\leq i\leq\lambda}\left\{\|X_n + \frac{\sigma^*\|X_n\|}{d}N^i(0,I_d)\|\right\}$$

where we can factorize $\|X_n\|$:

$$\|X_{n+1}\| = \|X_n\| \min_{1\leq i\leq\lambda} \left\{ \| \frac{X_n}{\|X_n\|} + \frac{\sigma^*}{d} N^i(0, I_d) \| \right\} \ .$$

Using the $*$ notation for the selected $(N^i(0, I_d))_{1\leq i\leq\lambda}$ we have:

$$\|X_{n+1}\| = \|X_n\| \| \frac{X_n}{\|X_n\|} + \frac{\sigma^*}{d} N^*(0, I_d) \| \ .$$

The independence is implied from Lemma 1 stating that the distribution of $\| \frac{X_n}{\|X_n\|} + \frac{\sigma^*}{d} N^*(0, I_d) \|$ is independent of X_n. □

From Lemma 2 and Lemma 1 we deduce the convergence in mean for $(X_n)_{n\in\mathbb{N}}$ defined in Eq. 15.

PROPOSITION 2. *For the algorithm defined in Eq. 15, with* $\mathbb{E}(\|X_0\|) < +\infty$ *the following holds:*

$$\mathbb{E}(\|X_{n+1}\| \,|X_n) = \|X_n\| \, \mathbb{E}\left(\|e_1 + \frac{\sigma^*}{d} N^*(0, I_d)\| \right) \tag{17}$$

and

$$\mathbb{E}(\|X_n\|) = \mathbb{E}(\|X_0\|) \left(\mathbb{E}\left(\|e_1 + \frac{\sigma^*}{d} N^*(0, I_d)\| \right) \right)^n \tag{18}$$

$$= \mathbb{E}(\|X_0\|) \left(1 - \frac{\varphi^*}{d} \right)^n \tag{19}$$

PROOF: We take the conditional expectation of Eq. 16. According to Lemma 1 the distribution of $\| \frac{X_n}{\|X_n\|} + \sigma^* N^*(0, I_d) \|$ is the same as the distribution of $\|e_1 + \sigma^* N^*(0, I_d)\|$ and is independent of $\|X_n\|$, therefore Eq. 17 follows. Taking again the expectation gives

$$\mathbb{E}(\|X_{n+1}\|) = \mathbb{E}(\|X_n\|) \mathbb{E}\left(\|e_1 + \frac{\sigma^*}{d} N^*(0, I_d)\| \right) .$$

Iterating yields Eq. 18, substituting with Eq. 14 yields Eq. 19. □

Proposition 2 states that convergence in mean of the algorithm defined in Eq. 15 occurs for positive progress with a convergence rate equal to

$$\mathbb{E}\left(\|e_1 + \frac{\sigma^*}{d} N^*(0, I_d)\| \right) = 1 - \frac{\varphi^*}{d} \tag{20}$$

and divergence occurs for negative progress. Moreover we see with Eq. 20 that the highest convergence rate is associated to the σ^* that maximizes the progress rate φ^*. According to Eq. 17 for positive progress or equivalently whenever convergence in mean occurs, $\|X_n\|$ is a positive supermartingale, *i.e.* $\mathbb{E}(\|X_{n+1}\| |X_n) \leq \|X_n\|$. The positivity and the supermartingale property implies almost sure convergence [9]. In the next section we find for some values of σ^* almost sure convergence whereas $\mathbb{E}\|X_n\|$ diverges.

3.2 Convergence despite a negative progress

For the scale-invariant algorithm (Eq. 15), Lemma 2 shows that $\|X_{n+1}\|$ is the product of $\|X_n\|$ and the independent random variable $\| \frac{X_n}{\|X_n\|} + \frac{\sigma^*}{d} N^*(0, I_d) \|$. Therefore Eq. 15 is an instantiation of Eq. 3. As pointed out in Section 2 the analysis of the expectation of Eq. 3 does not necessarily reflect what is observed for samples of the algorithm. **Figure 2** illustrates that for $\sigma^* = 3.133$, associated to negative progress in dimension 3 (or equivalently to divergence

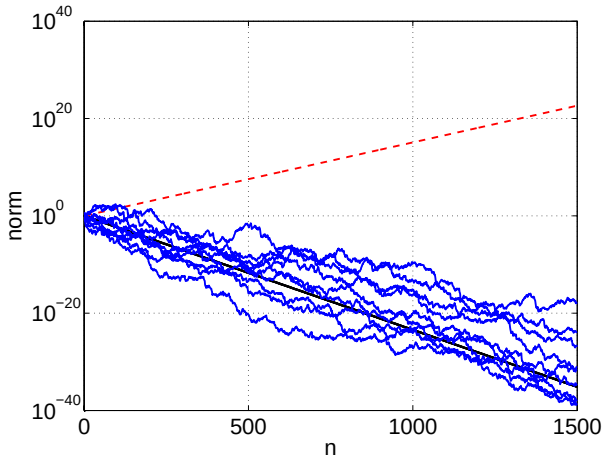

Figure 2: Ten realizations of the algorithm from Eq. 15 for $\sigma^* = 3.133$, dimension $d = 3$, $\lambda = 5$, and $\|X_0\| = 1$. The increasing dashed line depicts $\|X_0\| \times (1 - \varphi^*/d)^n$, the decreasing straight line depicts $\|X_0\| \times \exp(-\varphi^*_{\ln}/d)^n$. Both terms are modeling $\|X_n\|$.

in mean of the random sequence given in Eq. 15), almost sure convergence occurs. This is related to the same effect than the one observed in Section 2: almost sure convergence can happen without convergence in mean. The almost sure convergence of the algorithm Eq. 15 was analyzed in [4].

PROPOSITION 3 (ALMOST SURE CONVERGENCE). *For the algorithm defined in Eq. 15, the following holds*

$$\lim_{n\to\infty} \frac{1}{n} (\ln \|X_n\| - \ln \|X_0\|) \tag{21}$$

$$= \mathbb{E}\left(\ln \|e_1 + \frac{\sigma^*}{d} N^*(0, I_d)\| \right) \ a.s.$$

or equivalently

$$\lim_{n\to\infty} \left(\frac{\|X_n\|}{\|X_0\|} \right)^{1/n} = \exp\left(\mathbb{E}\left(\ln \|e_1 + \frac{\sigma^*}{d} N^*(0, I_d)\| \right) \right) \ a.s.$$

The limites are independent of X_0.

PROOF: The proof is similar to the proof of Proposition 1 using the fact that $\|X_n\|$ and $\| \frac{X_n}{\|X_n\|} + \frac{\sigma^*}{d} N^*(0, I_d)\|$ are independent (from Lemma 2). □

The proposition implies that almost sure convergence of X_n toward zero will occur for all σ^* such that

$$\mathbb{E}\left(\ln \|e_1 + \frac{\sigma^*}{d} N^*(0, I_d)\| \right) < 0 \ . \tag{22}$$

Additionally Eq. 22 gives the convergence rate for $\mathbb{E}(\ln(\|X_n\|))$:

PROPOSITION 4. *For the algorithm defined in Eq. 15 with* $\mathbb{E}(\ln(\|X_0\|)) < +\infty$, *the expectation of* $\ln(\|X_n\|)$ *obeys*

$$\frac{1}{n} (\mathbb{E}(\ln(\|X_n\|) - \ln(\|X_0\|))) \tag{23}$$

$$= \mathbb{E}\left(\ln \|e_1 + \frac{\sigma^*}{d} N^*(0, I_d)\| \right) \ .$$

PROOF: We take the logarithm of Eq. 16, use the result of Lemma 1, and take the expectation. □

The convergence rate of Eq. 23 is reflected by the slope of the decreasing straight line in Fig. 2.

When convergence in mean occurs, we did conclude from the positivity and the martingale property implied by Eq. 17 that almost sure convergence takes places. This can be deduced as well from Jensen's inequality implying that

$$\ln \mathbb{E}\left(\|e_1 + \frac{\sigma^*}{d} N^*(0, I_d)\| \right) > \mathbb{E}\left(\ln \|e_1 + \frac{\sigma^*}{d} N^*(0, I_d)\| \right) .$$

Motivated by Eq. 21 we suggest an alternative progress rate definition relating more to what is actually observed in single runs:

$$\varphi^*_{\ln} = d\, \mathbb{E}\left(\ln \frac{\|X_n\|}{\|X_{n+1}\|} \middle| X_n \right) , \tag{24}$$

where we did multiply by d to have the same normalization as φ^*. We refer to $\varphi^*_{\ln}$ as log-progress in the following.

4. FINITE DIMENSION PROGRESSES

We will now quantify on the sphere model the differences of progress definitions and formulas and evaluate their predictive power in terms of actually observed convergence of the evolution strategy. For this we simulate the classical progress rate φ^* using the relation

$$\varphi^* = d\left(1 - \mathbb{E}\left(\|e_1 + \frac{\sigma^*}{d} N^*(0, I_d)\| \right)\right) , \tag{25}$$

and the log-progress rate representing the convergence rate for almost sure convergence using the relation

$$\varphi^*_{\ln} = -d\, \mathbb{E}\left(\ln \|e_1 + \frac{\sigma^*}{d} N^*(0, I_d)\| \right) , \tag{26}$$

a consequence of Lemma 2 substituting in Eq. 24. We use a Monte Carlo simulation of the expectations with empirical averages of $10^6/\sqrt{d}$ samples from the random variables $\|e_1 + \frac{\sigma^*}{d} N^*(0, I_d)\|$ and $\ln \|e_1 + \frac{\sigma^*}{d} N^*(0, I_d)\|$ respectively.

Figure 3 shows the results for dimensions $d = 3, 10, 30$ in comparison with the progress rate approximation formula $\varphi^* \approx c_{1,\lambda}\sigma^* - 0.5\,\sigma^{*2}$ and its refinements Eq. 3.242 and Eq. 3.241 from [3] (respectively lower and upper light line). For larger dimensions all graphs become quite similar within the shown range (for larger σ^* they still disagree significantly). Already for $d = 10$ Eq. 3.241 from [3] and $\varphi^*_{\ln}$ are in a good agreement, while for $d = 3$ the deviations are remarkable.

To confirm the relevance of $\varphi^*_{\ln}$ we simulate the algorithm of Eq. 15 for $\sigma^* = 3.133$ and 3.516, where $d = 3$. These σ^*-values are shown as $*$ on the $\varphi = 0$ axis in Fig. 3, upper left. Only for the $\varphi^*_{\ln}$ measure the value is positive for the smaller σ^* and negative for the larger one. **Figure 4** shows ten realizations for the two σ^*-values respectively (crimped lines) for $d = 3$. The final norm values for the smaller σ^* are smaller than 10^{-20}, relating to positive progress, while the final values for the larger σ^* are larger than 10^{20}, relating to negative progress. This outcome is predicted only by the progress graph of $\varphi^*_{\ln}$ in Fig. 3. The results for $d = 10$ are similar (not shown).

Besides the ten realizations for the smaller σ^* the 10^{-5}, 10^{-3}, 0.5, $1 - 10^{-3}$, and $1 - 10^{-5}$-quantiles of 2.5×10^5 realizations are shown in Fig. 4 (lesser crimped lines) for $d = 3$. These graphs are estimates of the cumulative distribution function of X_n for the probabilities given by the quantiles.

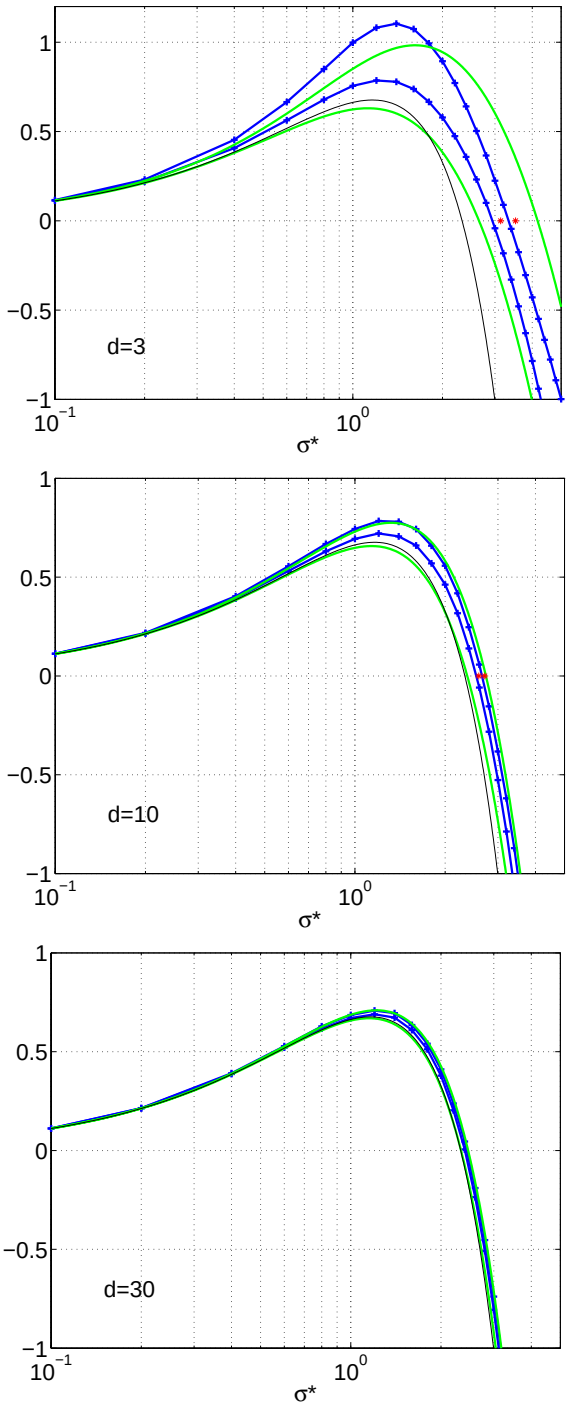

Figure 3: Monte Carlo simulation of φ^* and $\varphi^*_{\ln}$ (respectively lower and upper dark graph with crosses), the classical progress rate approximation formula $c_{1,\lambda}\,\sigma^* - 0.5\,\sigma^{*2}$ (thin dark line) and two refinements of the formula (related to φ^*) conferring to Eq. 3.242 and Eq. 3.241 from [3] (lower and upper light line, respectively), for dimensions $d = 3, 10, 30$. While for large dimensions a good agreement between the different progress lines is observed, the graphs deviate considerably for small dimensions.

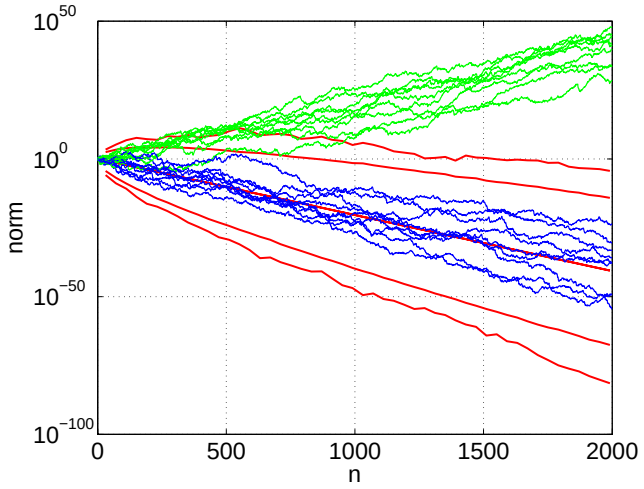

Figure 4: Twenty realizations of the algorithm from Eq. 15 with $\lambda = 5$, ten for $\sigma^* = 3.133$ and ten for $\sigma^* = 3.516$, where dimension $d = 3$, and $\|X_0\| = 1$. The lesser crimped lines depict for some time steps the 10^{-5}, 10^{-3}, 0.5, $1-10^{-3}$, and $1-10^{-5}$-quantiles of 2.5×10^5 realizations. For example, the lowest line is an estimate for $x(n)$ where $P(X_n < x(n)) = 10^{-5}$.

5. ASYMPTOTIC ESTIMATION OF PROGRESSES

The simulations shown in Section 4 suggest that for high dimension or sufficiently small σ^* both φ^* and φ^*_{ln} converge to the same values. We formalize this result in Theorem 1 stating that asymptotic estimations of φ^* and φ^*_{ln} (zero order in $1/d$ and second order in σ^*) are equal. The proof relies on the following proposition stating that asymptotically in the dimension, $N_1^* + \frac{\sigma^*}{2d}\|N^*\|^2$ (where N_1^* is the first coordinate of the selected vector) converges to $N_{1:\lambda} + \frac{\sigma^*}{2}$ where $N_{1:\lambda}$ is the minimum of λ normal distributions. This result is useful to derive rigorously asymptotic estimations of φ^* and, to the best of our knowledge, here proven for the first time.

PROPOSITION 5. *Let N_1^* be the first coordinate of the selected vector N^* (implicitly) defined in Eq. 13. The following holds:*

$$\lim_{d\to\infty} N_1^* + \frac{\sigma^*}{2d}\|N^*\|^2 = N_{1:\lambda} + \frac{\sigma^*}{2} \text{ a.s.} \tag{27}$$

where $N_{1:\lambda} = \min\{N_1, \ldots, N_\lambda\}$ is the first order statistics among λ normal random variables.

$$\lim_{d\to\infty} \mathbb{E}\left(N_1^* + \frac{\sigma^*}{2d}\|N^*\|^2\right) = -c_{1,\lambda} + \frac{\sigma^*}{2} \tag{28}$$

where $c_{1,\lambda} = -\mathbb{E}(N_{1:\lambda})$ is called progress coefficient [6, 3].

PROOF: Let $(N^i)_{1\le i\le\lambda}$ be λ independent normal random vectors with mean zero and covariance matrix identity. For $1 \le j \le d$, let $(N_j^i)_{1\le i\le\lambda}$ denote the coordinate j of $(N^i)_{1\le i\le\lambda}$. The equation

$$\|e_1 + \frac{\sigma^*}{d}N^*\|^2 = \min_{1\le i\le\lambda}\{\|e_1 + \frac{\sigma^*}{d}N^i\|^2\} \tag{29}$$

can be rewritten as

$$1 + \frac{\sigma^*}{d}\left(\frac{\sigma^*}{d}\|N^*\|^2 + 2N_1^*\right) \tag{30}$$
$$= \min_{1\le i\le\lambda}\left\{1 + \frac{\sigma^*}{d}\left(\frac{\sigma^*}{d}\|N^i\|^2 + 2N_1^i\right)\right\} ,$$

which implies that

$$\frac{\sigma^*}{d}\|N^*\|^2 + 2N_1^* = \min_{1\le i\le\lambda}\{\frac{\sigma^*}{d}\|N^i\|^2 + 2N_1^i\} . \tag{31}$$

Moreover for all $1 \le i \le \lambda$, $\|N^i\|^2 = \sum_{j=1}^d (N_j^i)^2$ is the sum of d independent random variables with finite expectation. With the Strong Law of Large Numbers

$$\lim_{d\to\infty}\frac{1}{d}\|N^i\|^2 = \mathbb{E}((N_1^i)^2) = 1 \text{ a.s.}$$

and therefore $\frac{\sigma^*}{d}\|N^i\|^2 + 2N_1^i$ converges almost surely to $\sigma^* + 2N_1^i$ and $\min_{1\le i\le\lambda}\{\frac{\sigma^*}{d}\|N^i\|^2 + 2N_1^i\}$ converges almost surely to

$$\min_{1\le i\le\lambda}\{\sigma^* + 2N_1^i\} = \sigma^* + 2N_{1:\lambda}$$

where $N_{1:\lambda}$ is the first order statistics among λ normal distributions. With Eq. 31, we obtain that

$$\lim_{d\to\infty}\frac{\sigma^*}{d}\|N^*\|^2 + 2N_1^* = \sigma^* + 2N_{1:\lambda} \text{ a.s.} \tag{32}$$

In addition, $\frac{\sigma^*}{d}\|N^*\|^2 + 2N_1^*$ is uniformly integrable since

$$\mathbb{E}(\|N^*\|^2) \le \mathbb{E}(\max_{1\le i\le\lambda}\chi_d^i) \le \lambda d , \tag{33}$$

where χ_d^i are chi-squared distributions and

$$\mathbb{E}(|N_1^*|) \le \mathbb{E}(\max_{1\le i\le\lambda}|N_i|) < +\infty . \tag{34}$$

Therefore

$$\lim_{d\to\infty}\mathbb{E}\left(\frac{\sigma^*}{d}\|N^*\|^2 + 2N_1^*\right) = \sigma^* + 2\mathbb{E}(N_{1:\lambda})$$
$$= \sigma^* - 2c_{1,\lambda} .$$

□

THEOREM 1 . *For a $(1,\lambda)$-ES on the sphere model, a zero order in $1/d$ and second order in σ^* approximation of φ^* and φ^*_{ln} is given by:*

$$c_{1,\lambda}\sigma^* - \frac{1}{2}(\sigma^*)^2 + \mathcal{O}\left(\frac{(\sigma^*)^3}{d}\right) \tag{35}$$

where $c_{1,\lambda} = -\mathbb{E}(N_{1:\lambda}) = \mathbb{E}(N_{\lambda:\lambda})$ is the expectation of the largest order statistics among normal random variables.

PROOF: Let N_1^* and N_{d-1}^* be respectively the first and last $d-1$ components of the selected vector $N^*(0, I_d)$. We decompose the vector $\|e_1 + \frac{\sigma^*}{d}N^*(0, I_d)\|$ in the following way.

$$
\begin{aligned}
&\|e_1 + \frac{\sigma^*}{d} N^*(0, I_d)\| \\
&= \left((1 + \frac{\sigma^*}{d} N_1^*)^2 + \left(\frac{\sigma^*}{d}\right)^2 \|N_{d-1}^*\|^2 \right)^{1/2} \\
&= \left(1 + 2\frac{\sigma^*}{d} N_1^* + \left(\frac{\sigma^*}{d}\right)^2 (N_1^*)^2 + \left(\frac{\sigma^*}{d}\right)^2 \|N_{d-1}^*\|^2 \right)^{1/2} \\
&= \left(1 + 2\frac{\sigma^*}{d} N_1^* + \left(\frac{\sigma^*}{d}\right)^2 \|N^*\|^2 \right)^{1/2}
\end{aligned}
$$

We use $(1+h)^{1/2} = 1 + \frac{1}{2}h + \mathcal{O}(h^2)$ to obtain:

$$
\begin{aligned}
&\|e_1 + \frac{\sigma^*}{d} N^*(0, I_d)\| \qquad (36) \\
&= 1 + \frac{\sigma^*}{d} N_1^* + \frac{1}{2}\left(\frac{\sigma^*}{d}\right)^2 \|N^*\|^2 + \mathcal{O}(\frac{(\sigma^*)^3}{d^2})
\end{aligned}
$$

We take the expectation of both sides and use Eq. 14:

$$
\varphi^* = -\sigma^* \left(\mathbb{E}\left(N_1^* + \frac{\sigma^*}{2d} \|N^*\|^2 \right) \right) + \mathcal{O}\left(\frac{(\sigma^*)^3}{d} \right)
$$

Using Proposition 5 we obtain

$$
\varphi^* = c_{1,\lambda}\sigma^* - \frac{(\sigma^*)^2}{2} + \mathcal{O}\left(\frac{(\sigma^*)^3}{d} \right) \quad .
$$

To obtain an estimation for $\varphi^*_{\ln}$ we take the log in Eq. 36 using the fact that $\ln(1+h) = h + \mathcal{O}(h^2)$:

$$
\ln \|e_1 + \frac{\sigma^*}{d} N^*(0, I_d)\| = \frac{\sigma^*}{d} N_1^* + \frac{1}{2}\left(\frac{\sigma^*}{d}\right)^2 \|N^*\|^2 + \mathcal{O}(\frac{(\sigma^*)^3}{d^2})
$$

We now take the expectation of both sides and multiply by d:

$$
\begin{aligned}
&d\,\mathbb{E}\left(\ln \|e_1 + \frac{\sigma^*}{d} N^*(0, I_d)\| \right) = \\
&\qquad \sigma^* \mathbb{E}(N_1^*) + (\sigma^*)^2 \frac{1}{2d} \mathbb{E}\left(\|N^*\|^2 \right) + \mathcal{O}\left(\frac{(\sigma^*)^3}{d} \right)
\end{aligned}
$$

Using Proposition 5 we obtain

$$
\varphi^*_{\ln} = c_{1,\lambda}\sigma^* - \frac{(\sigma^*)^2}{2} + \mathcal{O}\left(\frac{(\sigma^*)^3}{d} \right) \quad .
$$

□

The main reason why both models asymptotically agree is the correspondence of the Taylor expansions $\sqrt{1+h} - 1$ and $\ln \sqrt{1+h}$ in $h = 0$.

6. PRACTICAL RELEVANCE

An important and challenging task in Evolution Strategies is to find an efficient method to adapt the step-size. Several techniques have been introduced for this purpose. The first work in this direction is the one-fifth-success rule [6] where the step-size is adapted based on the rate of successful mutations. Then self-adaptation was proposed, where the step-size is mutated and selected according to the individuals fitness [8]. Cumulative path length control was introduced to overcome certain shortcomings of self-adaptation and is implemented in the CMA-ES [5].

One may question the relevance of the scale-invariant algorithm where the step-size is chosen proportional to the distance to the optimum, $\frac{\sigma^* \|X_n\|}{d}$, because in practice the location of the optimum is not known and is not used in algorithms that adapt the step-size. First we prove in Theorem 2 that the convergence rate of the scale-invariant $(1,\lambda)$-ES associated to an optimal choice of σ^* (denoted σ^*_{opt}) is the optimal convergence rate for adaptive $(1,\lambda)$-ES. Second we argue that the adaptive techniques mentioned above do achieve, on the sphere function, an adaptation where the step-size is proportional to the distance to the optimum.

We consider a $(1,\lambda)$-ES minimizing $f : \mathbb{R}^d \to \mathbb{R}^+$ where an adaptive step-size method is implemented:

$$
Y_{n+1} = \arg\min_{1 \le i \le \lambda} \left\{ f\left(Y_n + \sigma_n N^i(0, I_d) \right) \right\} \quad , \qquad (37)
$$

where the random variable σ_n denotes the step-size at generation n.

THEOREM 2 . *Let $f : \mathbb{R}^d \to \mathbb{R}^+$ be a fitness function with an unique global optimum x_{opt} and without loss of generality $x_{opt} = 0$. Let $Y_n \in \mathbb{R}^d$ be defined as in Eq. 37, where $(\sigma_n)_{n\in\mathbb{N}}$ is a sequence of strictly positive random variables and the random vector Y_0 is zero only with zero probability and $\mathbb{E}(\ln(\|Y_0\|)) < +\infty$. Then the following holds*

$$
\mathbb{E}\left(\ln \|Y_n\| - \ln \|Y_{n+1}\| \; | Y_n, \sigma_n \right) \le \max_{\sigma^*} \frac{\varphi^*_{\ln}(\sigma^*)}{d} = \frac{\varphi^*_{\ln}(\sigma^*_{opt})}{d}
$$

where $\varphi^*_{\ln}$ is defined in Eq. 24 and

$$
\varphi^*_{\ln}(\sigma^*) = -d\,\mathbb{E}\left(\ln \|e_1 + \frac{\sigma^*}{d} N^*(0, I_d)\| \right) \quad ,
$$

holds.

This bound is reached on the sphere function for

$$
\sigma_n = \frac{\sigma^*_{opt}}{d\|X_n\|} \quad .
$$

PROOF: Because $\|Y_{n+1}\| \ge \min_{1\le i\le\lambda} \|Y_n + \sigma_n N^i(0, I_d)\|$ we obtain:

$$
\begin{aligned}
&\mathbb{E}\left(\ln \|Y_n\| - \ln \|Y_{n+1}\| \; | Y_n, \sigma_n \right) \\
&\le \mathbb{E}\left(\ln \|Y_n\| - \ln \left(\min_{1\le i\le\lambda} \|Y_n + \sigma_n N^i(0, I_d)\| \right) \; | Y_n, \sigma_n \right) \\
&= \mathbb{E}\left(-\ln \left(\min_{1\le i\le\lambda} \| \frac{Y_n}{\|Y_n\|} + \frac{\sigma_n}{\|Y_n\|} N^i(0, I_d)\| \right) | Y_n, \sigma_n \right) \\
&= \mathbb{E}\left(-\ln \left(\min_{1\le i\le\lambda} \|e_1 + \frac{\sigma_n}{\|Y_n\|} N^i(0, I_d)\| \right) | Y_n, \sigma_n \right) \\
&\le \max_{\sigma^*} \mathbb{E}\left(-\ln \left(\min_{1\le i\le\lambda} \|e_1 + \sigma^* N^i(0, I_d)\| \right) \right) \\
&= \frac{\varphi^*_{\ln}(\sigma^*_{opt})}{d}
\end{aligned}
$$

□

The previous theorem states that the optimal adaptation scheme (on the sphere) chooses $\sigma \propto \|X_n\|$ as in the scale-invariant algorithm and $\sigma^* = \sigma^*_{opt}$. One can argue that in practice it is not possible to choose the step-size proportional to the norm. Simple experiments on the sphere show the contrary: with all adaptive techniques mentioned above, $\ln \|Y_n\|$ and $\ln(\sigma_n)$ decrease linearly[4] at the same

[4]Because $\ln \|Y_n\|$ and $\ln(\sigma_n)$ are random variables, the linear decrease is superposed by stochastic deviations as can be observed for instance in Fig. 2.

rate. Moreover this property is proven for self-adaptive schemes, where $\|Y_n\|/\sigma_n$ admits a stationary measure and converges "fast" to this measure, and sufficient conditions for ensuring this are derived [2]. As a corollary, the linear convergence of $\ln\|Y_n\|$ and $\ln(\sigma_n)$ is deduced [2]:

$$\lim_{n\to\infty}\frac{1}{n}\ln\|Y_n\| = \lim_{n\to\infty}\frac{1}{n}\ln\sigma_n = \mathbf{c}$$

where $\mathbf{c}$ is expressed in terms of the stationary distribution of $\|Y_n\|/\sigma_n$.

7. SUMMARY AND CONCLUSION

In this paper we have investigated the limits of the predictions based on the classical progress rate. We show how the definition of progress φ^* in Eq. 1 is related to the convergence of the scale-invariant $(1,\lambda)$-evolution strategy on a spherical fitness function as defined in Eq. 15: using

$$\|X_{n+1}\| = \|X_n\|\,\|e_1 + \frac{\sigma^*}{d}N^*(0,I_d)\| \ ,$$

we get

$$\left(\frac{\mathbb{E}(\|X_n\|)}{\mathbb{E}(\|X_0\|)}\right)^{1/n} = \mathbb{E}\left(\|e_1 + \frac{\sigma^*}{d}N^*(0,I_d)\|\right) \ , \qquad (38)$$

where the coefficient

$$\mathbb{E}\left(\|e_1 + \frac{\sigma^*}{d}N^*(0,I_d)\|\right) = 1 - \frac{\varphi^*}{d} \qquad (39)$$

determines the convergence or divergence in mean. From Eq. 39 we see that positive progress corresponds to convergence in mean and negative progress to divergence in mean.

Convergence in mean and almost sure convergence are not equivalent. In reality single samples of the algorithm are observed and thus almost sure results reflect what we observe. Almost sure convergence is given in Proposition 3 by

$$\lim_{n\to\infty}\left(\frac{\|X_n\|}{\|X_0\|}\right)^{1/n} = \exp\mathbb{E}\left(\ln\|e_1 + \frac{\sigma^*}{d}N^*(0,I_d)\|\right) \text{ a.s. } , \qquad (40)$$

where the sign of the coefficient

$$\mathbb{E}(\ln\|e_1 + \frac{\sigma^*}{d}N^*(0,I_d)\|) = -\frac{\varphi^*_{\ln}}{d} \qquad (41)$$

determines whether almost sure convergence or divergence holds. Equations 40 and 41 suggest the definition of a new progress rate, the log-progress

$$\varphi^*_{\ln} = d\,\mathbb{E}\left(\ln\frac{\|X_n\|}{\|X_{n+1}\|}\,\middle|\,X_n\right) \ , \qquad (42)$$

and the replacement of $(1-\varphi^*/d)^n$, as a (deterministic) model for the sequence $\|X_n\|/\|X_0\|$, by the almost sure convergence rate

$$\exp\left(-\frac{\varphi^*_{\ln}}{d}\right)^n \ . \qquad (43)$$

According to Eq. 41 and Eq. 39 there exists step-sizes σ^* where divergence in mean occurs, *i.e.* $1-\varphi^*/d > 1$, while almost sure convergence takes place, *i.e.* $\exp(-\varphi^*_{\ln}/d) < 1$. One simulation of this phenomenon was given in Fig. 2. *That means there exists step-sizes (close to the right border of the evolution window) for which the progress rate φ^* is negative but almost sure convergence is observed (and $\varphi^*_{\ln}$ is positive).* We show that the zero order in $1/d$ and second order in σ^* estimations of φ^* and $\varphi^*_{\ln}$ coincide (Section 5).

We come back to our simple example from Section 1, where the algorithm oscillates between two points and $\|X_n\|$ equals 1 for even n and 0.5 otherwise. Using the new progress definition Eq. 42 the quantity $\ln(\|X_n\|/\|X_{n+1}\|)$ equals to $\ln(2)$ for even n and $-\ln(2)$ for odd n, reflecting well that the algorithm is not stepping ahead in time.

Many of our results are based on the scale-invariant algorithm from Eq. 15, where the step-size is chosen proportional to the distance to the optimum. Given $\sigma^* = \sigma^*_{\text{opt}}$ the scale-invariant algorithm running on the sphere function achieves the maximum possible log-progress and the fastest convergence to zero for the $(1,\lambda)$-ES (Theorem 2). While in the scale-invariant algorithm $\sigma_n \propto \|X_n\|$ is chosen, theoretical and empirical results reveal that this seemingly unrealistic setting turns out to be what is approximately achieved by realistic and well-known adaptive techniques: $\|X_n\|/\sigma_n$ converges to a stationary measures.

We conclude with three final remarks.

- The convergence results that we provide do not only state convergence but are associated with *convergence rates*. Therefore they give insights into the finite time behavior and they are a realistic measure of efficiency.
- The log-progress rate corresponds to almost sure convergence and achieves correct convergence predictions even in finite dimensions for any unimodal spherical fitness function. It can be used as an empirical performance measure using Monte-Carlo integration. Any analysis aiming at finite dimensional progress rates should use as a starting point $\varphi^*_{\ln}$ rather than φ^* to reflect more what is actually observed in single runs.
- The theory of progress rates and the log-progress $\varphi^*_{\ln}$ is not confined to the sphere function alone but can be naturally extended to any function $f(x) = g(x^T H x)$ with Hessian matrix H and a strictly increasing function $g : \mathbb{R}^+ \to \mathbb{R}$. The norm in the progress definition is then replaced by the respective Mahalanobis metric.

8. REFERENCES

[1] A. Auger, C. Le Bris, and M. Schoenauer. Dimension-independent Convergence Rate for Non-isotropic $(1,\lambda)-$ES. In E. Cantu-Paz et al., editor, *Proceedings of the Genetic and Evolutionary Conference 2003*, pages 512–524, 2003.

[2] 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.

[3] H.-G. Beyer. *The Theory of Evolution Strategies*. Springer, Heidelberg, 2001.

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

[5] N. Hansen and A. Ostermeier. Completely Derandomized Self-Adaptation in Evolution Strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

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

[7] G. Rudolph. *Convergence Properties of Evolutionary Algorithms*. Kovac, Hamburg, 1997.

[8] H. P. Schwefel. *Numerical Optimization of Computer Models*. John Wiley & Sons, New-York, 1995 – 2^{nd} edition edition, 1981.

[9] D. Williams. *Probability with Martingales*. Cambridge University Press, Cambridge, 2000.