A. Auger and N. Hansen

Theory of Evolution Strategies: a New Perspective

In: A. Auger and B. Doerr, eds. (2010). *Theory of Randomized Search Heuristics: Foundations and Recent Developments*. World Scientific Publishing, pp. 289–325.

Erratum On page 308, before and in Theorem 10.12, it is stated that $Q(x, .)$ needs to have a singularity in zero. Correct is that $Q(x, .)$ needs to have a singularity in x. Note that because the theorem refers to the density associated to $Q(., .)$, it is implicitly assumed that $Q(., .)$ is absolutely continuous with respect to the Lebesgue measure.

Thanks to Alexandre Chotard for pointing out the error to us.

Chapter 10

Theory of Evolution Strategies: A New Perspective

Anne Auger and Nikolaus Hansen
TAO Team, INRIA Saclay — Île-de-France, Orsay, France
anne.auger@inria.fr
nikolaus.hansen@inria.fr

Evolution Strategies (ESs) are stochastic optimization algorithms recognized as powerful algorithms for difficult optimization problems in a black-box scenario. Together with other stochastic search algorithms for continuous domain (like Differential Evolution, Estimation of Distribution Algorithms, Particle Swarm Optimization, Simulated Annealing...) they are so-called *global optimization algorithms*, as opposed to gradient-based algorithms usually referred to as local search algorithms. Many theoretical works on stochastic optimization algorithms focus on investigating convergence to the global optimum with probability one, under very mild assumptions on the objective functions. On the other hand, the theory of Evolution Strategies has been restricted for a long time to the so-called *progress rate theory*, analyzing the one-step progress of ESs on unimodal, possibly noisy functions. This chapter covers global convergence results, revealing slow convergence rates on a wide class of functions, and fast convergence results on more restricted function classes. After reviewing the important components of ESs algorithms, we illustrate how global convergence with probability one can be proven easily. We recall two important classes of convergence, namely sub-linear and linear convergence, corresponding to the convergence class of the pure random search and to the optimal convergence class for rank-based algorithms respectively. We review different lower and upper bounds for adaptive ESs, and explain the link between lower bounds and the progress rate theory. In the last part, we focus on recent results on linear convergence of adaptive ESs for the class of spherical and ellipsoidal functions, we explain how almost sure linear convergence can be proven using different laws of large numbers (LLN).

Contents

10.1. Introduction

In this chapter we focus on numerical optimization where the functions to be optimized are mapping $\mathcal{D}$, a subset of the euclidian space $\mathbb{R}^d$ equipped with the euclidian norm $\|.\|$, into $\mathbb{R}$ and without loss of generality we assume minimization. Moreover, we consider derivative free optimization algorithms, i.e., algorithms that do not use derivatives of the function to optimize. Often, those derivatives do not exist or are too costly to evaluate. For instance, the function f can be given by an expensive numerical simulation, a frequent situation in many real-world optimization problems. Such a context is called black-box optimization. The objective function $f : \mathcal{D} \subset \mathbb{R}^d \to \mathbb{R}$ is modeled as a black-box that is able to compute, for a given $x \in \mathbb{R}^d$, the associated objective function value, $f(x)$. The algorithms considered in this chapter are Evolution Strategies (ESs), known as robust and efficient algorithms for black-box optimization without derivative with numerous successful applications to scientific and industrial problems. ESs are just an instance of adaptive stochastic search algorithms where the search space is explored by sampling probe points according to a (contin-

uous) search distribution that can change (be adapted) during the search process.

We assume for the moment that the optimization goal is to approach, with the least search cost, i.e., the least number of function evaluations, a solution x^* such that $f(x^*) \leq f(x)$ for all $x \in \mathcal{D}$ and refer to Section 10.2.1 for a more suitable definition of x^*. The solution x^* will be called global optimum.

The simplest stochastic search algorithm for black-box optimization is the pure random search (PRS), proposed by Brooks (1958) that samples solutions independently with the same probability distribution defined over $\mathcal{D}$ and takes as approximation, or estimate, of the optimal solution, the best solution visited so far, where best refers to the solution having the smallest objective function value. In PRS, the samples are independent and identically distributed, therefore the exploration is "blind"—no feedback from the objective function values observed is taken into account for guiding the next steps. However, in general, the functions to optimize have an underlying structure that can be exploited by optimization algorithms. This fact was soon recognized after the introduction of the PRS in 1958 and research on the algorithmic point of view focused on finding techniques to adapt the search distribution according to the already observed solutions.

10.1.1. *Adaptive Search Algorithms: A Tour d'Horizon*

Adaptive search algorithms refer here to algorithms where the sampling distribution—as opposed to PRS—is adapted along the search. We denote by X_n the best estimate of the (global) minimum after n iterations and by Y_n a *probe point* or *candidate solution*. In PRS, $Y_0, \ldots, Y_n, \ldots$ are independent and identically distributed (i.i.d.) over $\mathcal{D}$, $X_0 = Y_0$ and

$$\mathrm{X}_{n+1} = \begin{cases} \mathrm{Y}_n \text{ if } f(\mathrm{Y}_n) \leq f(\mathrm{X}_n) \ , \\ \mathrm{X}_n \text{ otherwise.} \end{cases} \tag{10.1}$$

The probably most important concept in an *adaptive* algorithm is to probe preferably in the neighborhood of the current solution since for non pathological functions good solutions are usually close to even better ones. If $W_0, \ldots, W_n, \ldots$ are i.i.d. with a probability measure usually centered in zero, independent of the initial estimate X_0, then $\mathrm{X}_n + W_n$ is the new probe point and the update is given by

$$\mathrm{X}_{n+1} = \begin{cases} \mathrm{X}_n + W_n & \text{if } f(\mathrm{X}_n + W_n) \leq f(\mathrm{X}_n) \ , \\ \mathrm{X}_n & \text{otherwise.} \end{cases} \tag{10.2}$$

This algorithm has different names like *local random search* (Devroye and Krzyżak, 2002) or *Markov monotonous search* (Zhigljavsky and Zilinskas, 2008) and will be referred here—following the terminology for evolutionary algorithms—as (1+1)-Evolution Strategy. The "(1+1)" notation comes from the fact that one point X_n (the first "1" in (1+1)) is used to generate another point $\mathrm{X}_n + W_n$ (the second "1" in (1+1)) and both are compared to achieve the point X_{n+1} (the best among X_n plus $\mathrm{X}_n + W_n$ is kept). In ESs the random vectors $(W_k)_{k\in\mathbb{N}}$ follow typically a multivariate normal distribution with zero mean. The pioneers of Evolution Strategies are I. Rechenberg (Rechenberg, 1973) and H.-P. Schwefel (Schwefel, 1981).

The term *local random search* suggests that a close neighborhood of X_n is explored. The size of this neighborhood is determined by the dispersion (or variance if it exists) of the distribution of W_n which is fixed once for all in the local random search algorithm. However, it seems natural that this dispersion needs to be adapted as well: in the first iterations, exploration of the search space and thus a large dispersion should be preferred and in the last iterations smaller dispersions are needed so as to converge. The question of *how to adapt the dispersion* has been central in the field of stochastic optimization. Consequently, many different methods have been proposed to address this issue and, already in 1971, a survey of different techniques was written by White (1971).

Before to review some important steps in this respect, we set a new framework, restricting the distribution for the new probe point to a spherical multivariate normal distribution. Let $N_0, \ldots, N_n, \ldots$ be i.i.d. multivariate normal random vectors where for each n, each coordinate of N_n follows a standard normal distribution independent of the other coordinates. A new probe point at iteration n is given by $\mathrm{X}_n + \sigma_n N_n$, where σ_n is a strictly positive parameter, called step-size. Since the coordinates of N_n are standard normally distributed, the parameter σ_n corresponds to the standard deviation of each coordinate of N_n. The update for a so-called (1+1)-ES with adaptive step-size reads

$$\mathrm{X}_{n+1} = \begin{cases} \mathrm{X}_n + \sigma_n N_n & \text{if } f(\mathrm{X}_n + \sigma_n N_n) \leq f(\mathrm{X}_n) \ , \\ \mathrm{X}_n & \text{otherwise,} \end{cases} \tag{10.3}$$

for the update of X_n plus an update equation for the step-size σ_n, $(X_0, \sigma_0) \in \mathcal{D} \times \mathbb{R}^+$.

One basic principle behind step-size adaptive algorithms is to try bigger steps if an improvement is observed, and smaller steps if a probe is unsuccessful (Matyas, 1965). Schumer and Steiglitz (1968) propose to maintain a

constant probability of success of around 1/5, where probability of success is defined as the probability that the new probe point has an objective function value smaller than the current solution. This idea can also be found in Devroye (1972) and Rechenberg (1973) and is known in the domain of evolutionary algorithms under the name of 1/5th-success rule[a].

Adaptivity of random search algorithms is not limited to a single parameter, like the step-size. The multivariate normal distribution used to sample the new probe point has, besides its mean value, $(d^2 + d)/2$ variation parameters—variances and covariances. These parameters reflect a quadratic model. More recently, a surprisingly effective method has been introduced to adapt all variances and covariances in the *covariance matrix adaptation* (CMA) evolution strategy (Hansen and Ostermeier, 2001). Three main ideas are exploited: (1) the search path (evolution path) over a backward time horizon of about d iterations is pursued and its length and direction are analyzed, (2) new probe points are favored in directions where previously probe points were successful, (3) invariance properties are maintained and the method is in particular invariant under the choice of the coordinate system.

The methods outlined above sample only one new probe point. An important ingredient of *evolutionary algorithms* however is to sample a *population* of probe points. ESs loop over the following steps: (1) sample new solutions from a multivariate normal distribution, (2) evaluate the objective function value of those solutions and (3) adapt the parameters of the multivariate normal distribution (mean vector, step-size and/or covariance matrix) using the observed data. The last step is a crucial step for an ES to converge faster than random search as we will see later.

Following the terminology used for evolutionary algorithms, we might, in the sequel, call *parent* the current search point X_n and *offspring* the new probe points generated from X_n.

10.1.2. *What to Analyze Theoretically?*

Given an optimization algorithm, the first question usually investigated is the one of convergence that can be formulated as: will the algorithm, when time grows to infinity, get arbitrarily close to an optimum of the optimization problem?

[a] A simple implementation can be found in (Kern *et al.*, 2004): the step-size is multiplied by $\alpha > 1$ in case of success and divided by $\alpha^{1/4}$ otherwise. This algorithm will be analyzed in Section 10.4.3.2.

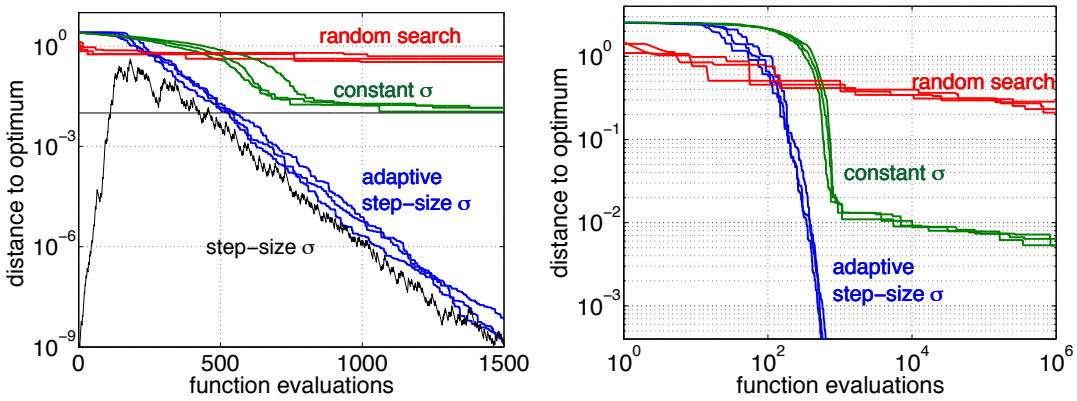

Fig. 10.1. Time evolution of the distance to the global minimum $\|\mathrm{X}_n - x^*\|$ of three different (1+1)-algorithms on the function $f : x \mapsto g(\|x - x^*\|)$, where g is strictly monotonically increasing. For each algorithm, three trials on the left and three on the right are shown. On the left, the jagged black line starting at 10^{-9} shows additionally the step-size of one trial with adaptive step-size. The constant horizontal line shows the step-size of the constant step-size algorithm, 10^{-2}. The initial step-size of the adaptive step-size algorithm was intentionally chosen much smaller than 10^{-2} in order to observe the effect of the adaptation. On the right, the trials are shown in a log-log plot and the adaptive step-size algorithm has an initial step-size equal to 10^{-2} to have a fair comparison with the algorithm with constant step-size. From the right plot, we estimate that the algorithm with constant step-size reaches a distance to the optimum of 10^{-3} in at least 10^{12} function evaluations and is therefore more than 1 billion times slower than the adaptive step-size algorithm.

Convergence is illustrated in Figure 10.1. The objective function to be minimized is the sphere function defined as $f(x) = \|x\|$. The global minimum is here unique and equals 0. The optimization goal is thus to approach 0 as fast as possible. Shown are 18 realizations (or runs) from three different algorithms. The y-axis depicts the objective function value of the best solution reached so far (note the log-scale) plotted against the number of objective function evaluations (x-axis) regarded as execution time[b]. The three algorithms converge to the optimum: they approach zero arbitrarily close when the time grows to infinity.

Global convergence can either refer to convergence to a local optimum independently of the starting point or convergence to a global optimum. The former is usually meant in the "deterministic" optimization community, where it is often formulated as convergence to zero of the sequence of

[b]We assume that the search cost (and thus the overall execution time) is mainly due to the objective function evaluations and not to the internal operations of the algorithm. This assumption is reasonable for many problems that are handled with evolutionary algorithms.

gradients associated to the candidate solutions. The latter is usually meant in the "stochastic" optimization community.

Convergence results alone are of little interest from a practical viewpoint because they do not tell us how long it takes to approach a solution. The three algorithms presented in Figure 10.1 all converge with probability one to the optimum. However, we clearly observe that they do not converge at the same speed. The (1+1) with constant step-size is estimated to be 1 billion times slower than the (1+1) with adaptive step-size to reach a distance to the optimum of 10^{-3}.

Therefore, it is important to investigate, together with convergence, the speed of convergence or convergence rate of an algorithm, i.e., to study how fast the algorithm approaches the global minimum. The question of speed of convergence can be tackled in different ways.

(1) We can study the speed at which the distance to x^*, $\|\mathrm{X}_n - x^*\|$, decreases to zero.
(2) We can study the *hitting time* τ_ϵ of an ϵ-ball $B_\epsilon(x^*)$ around the optimum, $\tau_\epsilon = \inf\{n : \mathrm{X}_n \in B_\epsilon(x^*)\}$ or the hitting time of a sublevel set $\{x|f(x) \leq f(x^*) + \epsilon\}$. Assume $E[\tau_\epsilon]$ is finite, studying the rate of convergence amounts to studying how $E[\tau_\epsilon]$ depends on ϵ when ϵ goes to zero.

Convergence speeds are classified into classes (quadratic convergence, linear convergence, sublinear convergence, ...) that we will define more precisely in Section 10.2.3. For instance, the fastest algorithm depicted in Figure 10.1, left, is in the linear convergence class (the logarithm of the distance to the optimum decreases linearly with the number of function evaluations). Within a class, constants, usually referred to as convergence rates, determine more precisely the speed of convergence.

After investigating for a given class of functions to which convergence class an algorithm belongs to, one is usually interested in estimating the constants (convergence rates). Determining the convergence rate will allow to compare the speed of convergence of two algorithms from the same convergence class. For instance, for the fastest algorithm depicted in Figure 10.1, left, for which the logarithm of the distance to the optimum decreases linearly, the convergence rate corresponds to the coefficient for the linear decrease, i.e., the averaged slope of the three curves represented in Figure 10.1. However, determining the constants analytically is in general much more difficult than finding the convergence class (Nekrutkin and

Tikhomirov, 1993) and requires further assumptions. Related to this last question, one is interested in the scaling of the constants with respect to the dimension of the search space.

An analysis in discrete search spaces is different in several respects. In a discrete space one can consider the hitting time of the optimum, $\tau_{x^*} = \inf\{n : \mathrm{X}_n = x^*\}$. Then, only the scaling of $E[\tau_{x^*}]$ with respect to the size of the search space is investigated. For instance, if the domain is the set of bit-strings of length d, $\mathcal{D} = \{0,1\}^d$, one investigates the scaling of $E[\tau_{x^*}]$ with respect to d.

10.1.3. *Notations and Structure of the Chapter*

We denote $\mathbb{N}$ the set of non-negative integers $\{0, 1, \ldots\}$, denote $\mathbb{R}^+$ the set $(0, +\infty)$, denote $B_\epsilon(x)$ the ball in $\mathbb{R}^d$ of center x and radius ϵ for the euclidian norm, i.e., $B_\epsilon(x) = \{y \in \mathbb{R}^d, \|y - x\| \leq \epsilon\}$ and denote $\mathrm{V}ol(.)$ the volume in $\mathbb{R}^d$ for the Lebesgue measure, i.e., $\mathrm{V}ol(A) = \int_A dx$. The volume of the ball $B_\epsilon(x)$ for all x is, for a fixed dimension d, proportional to ϵ^d, more precisely

$$\mathrm{V}ol(B_\epsilon(x)) = \frac{\pi^{d/2}\epsilon^d}{\Gamma(\frac{d}{2}+1)}, \tag{10.4}$$

where $\Gamma(.)$ is the gamma function. The Borel σ-algebra on $\mathbb{R}^d$ is denoted $\mathcal{B}(\mathbb{R}^d)$. The binary operator $\wedge$ denotes either the logical conjunction or the minimum between two real numbers. Technically, when a is a real number and b a random variable, $b : \Omega \to \mathbb{R}$ then $a \wedge b$ is a random variable such that for each $\omega \in \Omega$, $(a \wedge b)(\omega) = \min\{a, b(\omega)\}$. A vector distributed according to a multivariate normal distribution with zero mean vector and identity as covariance matrix is said to be distributed as $N(0, I_d)$. The set of strictly increasing transformations on $\mathbb{R}$ is denoted $\mathcal{M}$, namely $\mathcal{M} = \{g : \mathbb{R} \to \mathbb{R}, \forall x, y \text{ such that } x < y, g(x) < g(y)\}$.

For a real-valued function $x \mapsto h(x)$, we introduce its positive part $h^+(x) := \max\{0, h(x)\}$ and negative part $h^- = (-h)^+$. In other words $h = h^+ - h^-$ and $|h| = h^+ + h^-$. In the sequel, we denote by e_1 a unitary vector in $\mathbb{R}^d$ and w.l.o.g. $e_1 = (1, 0, \ldots, 0)$.

The organization of the chapter is the following. In Section 10.2 we start by giving a rigorous definition of the optimization goal; define different modes of convergence for sequences of random vectors; define linear convergence and sub-linear convergence; illustrate simple proofs for convergence (without convergence rate) for the (1+1)-ES and explain the invariance to

strictly increasing transformations of ESs together with its consequences. In Section 10.3 we analyze in detail the convergence class of the pure random search and review lower and upper bounds of non-adaptive ESs. In Section 10.4, we present tight lower bounds for step-size adaptive ESs and link those results with the progress rate theory. We then explain how linear convergence of adaptive ESs can be proven.

10.2. Preliminary Definitions and Results

In this section we come back on the mathematical formulation of a search problem and define different modes of convergence needed throughout the chapter, as well as important convergence classes for optimization. We illustrate convergence proofs for the simple (1+1)-ES with fixed sample distribution. We also explain the invariance of ESs to order preserving transformations of the objective function and its consequences.

10.2.1. *Mathematical Formulation of the Search Problem*

The goal in numerical optimization is to approximate the global minimum of a real valued function f defined on a subset $\mathcal{D}$ of $\mathbb{R}^d$. Without further assumptions on f, this problem may have no solution. First of all, f may have no minimum in $\mathcal{D}$, take for instance $f(x) = x$ for $x \in \mathcal{D} = (0, 1)$ or $\mathcal{D} = \mathbb{R}$, or the minimum may not be unique as for $f(x) = \sin(x)$ in $\mathcal{D} = \mathbb{R}$. However, even if f admits a unique global minimum on $\mathcal{D}$, this minimum can be impossible to approach in practice: take for instance $f(x) = x^2$ for all $x \in \mathbb{R}\backslash\{1\}$ and $f(1) = -1$. Then the global minimum of f is located in 1, however, it would be impossible to approach it in a black-box scenario. To circumvent this, we take the approach from measure theory considering classes of functions (instead of functions) where two functions belong to the same class if and only if they are equal except on a set of measure zero. Let ν be a measure on $\mathcal{D}$, and $f, g : \mathcal{D} \to \mathbb{R}$, then g belongs to the class of f, denoted $[f]$, if g and f are equal almost everywhere, that is $\nu\{x, f(x) \neq g(x)\} = 0$. The generalization of the minimum of a function to classes of functions is the so-called essential infimum defined for a function f as

$$m_\nu(f) = \operatorname{ess\,inf} f = \sup\{b \in [-\infty, \infty], \nu(\{x \in \mathcal{D} : f(x) < b\}) = 0\}, \tag{10.5}$$

and which is constant for all g in $[f]$. When the context is clear, we write m_ν instead of $m_\nu(f)$ and we will not differentiate between the minimum of the function class $[f]$ and of any $g \in [f]$.

If X is a random variable with probability measure ν, then $\Pr[f(X) < m_\nu] = 0$ and $\Pr[f(X) < m_\nu + \epsilon] > 0$ for all $\epsilon > 0$. The value m_ν depends on ν but will be the same for equivalent measures—measures having the same null sets. We denote in the following with m the essential infimum with respect to the Lebesgue measure on $\mathcal{D}$ or, equivalently, with respect to any probability measure with a strictly positive density everywhere on $\mathcal{D}$.

From now on, we assume that there exists a unique point x^* in the domain closure $\overline{\mathcal{D}}$, such that $\nu\{x \in B_\epsilon(x^*) : f(x) < m + \delta\} > 0$ for all $\epsilon, \delta > 0$. The definition of x^* is independent of the choice of a function in $[f]$. Then, for any $g \in [f]$, the well-defined optimization goal is to approach the unique *global optimum* $x^* \in \overline{\mathcal{D}}$.

With the above definitions we obtain $x^* = 0$ for $f(x) = x$ and $\mathcal{D} = (0, 1)$, and $x^* = 0$ for $f(x) = x^2$ for $x \in \mathbb{R}\backslash\{1\}$ and $f(1) = -1$. However, not all functions admit a global optimum according to our definition, take for instance $f(x) = x$ and $\mathcal{D} = \mathbb{R}$. For solving such a function in practice, one expects an algorithm to generate X_n with $\lim_{n\to\infty} X_n = -\infty$.

10.2.2. *Different Modes of Convergence*

If $(x_n)_{n\in\mathbb{N}}$ is a deterministic sequence of $\mathbb{R}^d$, all possible definitions of convergence are equivalent to the following: x_n converges to x^* if for all $\epsilon > 0$, there exists n_1 such that for all $n \geq n_1 \in \mathbb{N}$, $\|x_n - x^*\| \leq \epsilon$. Notations used are $\lim_{n\to\infty} \|x_n - x^*\| = 0$ or $\lim_{n\to\infty} x_n = x^*$.

However, for a sequence $(\mathrm{X}_n)_{n\in\mathbb{N}}$ of random vectors there exist different modes of convergence inducing different definitions of convergence which are not equivalent.

Definition 10.1 (Almost sure convergence). *The sequence* X_n *converges to a random variable* X *almost surely (a.s.) or with probability one if*

$$\Pr[\lim_{n\to\infty} X_n = X] = 1 \ .$$

Except on an event of probability zero, all single instances of the sequence $(\mathrm{X}_n)_n$ converge to X. Both notations $\lim_{n\to\infty} \|X_n - X\| = 0$ a.s. or $\lim_{n\to\infty} X_n = X$ a.s. will be used, where a.s. stands for almost surely.

A weaker type of convergence is convergence in probability.

Definition 10.2 (Convergence in probability). *The sequence X_n converges in probability towards X if for all $\epsilon > 0$*

$$\lim_{n\to\infty} \Pr\left[\|X_n - X\| \geq \varepsilon\right] = 0.$$

Almost sure convergence implies convergence in probability.

Definition 10.3 (Convergence in p-mean or in L_p).
The sequence X_n converges towards X in p-mean or in L_p for $p \geq 1$ if $E[\|\mathrm{X}_n\|^p] < \infty$ and

$$\lim_{n\to\infty} E[\|\mathrm{X}_n - X\|^p] = 0 \ .$$

If $p = 1$ we say simply that X_n converges in mean or in expectation towards X.

10.2.3. *Convergence Order of Deterministic Sequences*

Let a deterministic sequence $(z_n)_{n\in\mathbb{N}}$ converge to $z \in \mathbb{R}^d$ with $z_n \neq z$ for all n, and let

$$\lim_{n\to\infty} \frac{\|z_{n+1} - z\|}{\|z_n - z\|^q} = \mu, \text{ with } q \geq 1 \text{ and } \mu \in (0, 1) \ . \tag{10.6}$$

Depending on q and μ we define different *orders of convergence*.

Super-linear convergence, if $\mu = 0$ or $q > 1$. If $\mu > 0$, we speak about convergence with order $q > 1$ and about quadratic convergence if $q = 2$.

Linear convergence, if $q = 1$ and $\mu \in (0, 1)$, where we have consequently

$$\lim_{n\to\infty} \frac{\|z_{n+1} - z\|}{\|z_n - z\|} = \mu \in (0, 1) \ . \tag{10.7}$$

Sub-linear convergence, if $q = 1$ and $\mu = 1$. Furthermore, the sequence $(z_n)_n$ converges sub-linear *with degree* $p > 0$, if

$$\frac{\|z_{n+1} - z\|}{\|z_n - z\|} = 1 - c_n\|z_n - z\|^{1/p} \text{ and } c_n \to c > 0 \ . \tag{10.8}$$

The distance to z is reduced in each iteration by a factor that approaches one[c]. For $p = \infty$ and $c < 1$, linear convergence is recovered.

[c]This definition does not allow to characterize the convergence of all sequences converging sub-linearly, for instance $1/(n\log(n))$ converges sub-linearly to zero but lies in between convergence of degree 1 and $1 + \epsilon$ for all $\epsilon > 0$.

Sub-linear convergence with degree p implies that

$$\|z_n - z\| \sim \left(\frac{p}{c}\right)^p \frac{1}{n^p}. \tag{10.9}$$

See Stewart (1995).

10.2.4. *Log- and Sub-linear Convergence of Random Variables*

Since the limit of $\frac{\|z_{n+1}-z\|}{\|z_n-z\|}$ is a random variable, defining convergence like with Equations (10.6) or (10.7) is in general not appropriate. We could use the deterministic sequence $E[\|z_n - z\|]$ instead of $\|z_n - z\|$. However, we rather use a weaker definition of linear convergence implied by Equation (10.7): we say that a sequence of random variables converges log-linearly, if there exits $c < 0$ such that

$$\lim_n \frac{1}{n} \ln \frac{\|z_n - z\|}{\|z_0 - z\|} = c \ . \tag{10.10}$$

The definition of linear convergence in Equation (10.10) is implied by Equation (10.7) since Equations (10.7) is equivalent to $\lim_k \ln(\|z_{k+1} - z\|/\|z_k - z\|) = \ln(\mu)$ and by the Cesàro mean result we obtain $\lim_n \frac{1}{n}\sum_{k=0}^{n-1} \ln(\|z_{k+1} - z\|/\|z_k - z\|) = \ln(\mu)$ which after simplification gives Equation (10.10) where $c = \ln(\mu)$. Equation (10.10) implies also that the logarithm of $\|z_n - z\|$ converges to $-\infty$ like cn, suggesting the name *log-linear convergence*. We will say that log-linear convergence or linear convergence holds almost surely if there exists $c < 0$ such that Equation (10.10) holds almost surely. We will say that log-linear convergence or linear convergence holds in mean or expectation if there exists $c < 0$ such that

$$\lim_n \frac{1}{n} E\left[\ln \frac{\|z_n - z\|}{\|z_0 - z\|}\right] = c \ . \tag{10.11}$$

Log-linear convergence is illustrated in Figure 10.2.

Following Equation (10.9), we will say that a sequence of random variables converges sub-linearly with degree p if $\|z_n - z\|$ converges to zero like $\frac{1}{n^p}$. Sub-linear convergence is illustrated in Figure 10.3.

As explained in Section 10.1.2, speed of convergence and thus linear and sublinear convergence can be alternatively defined with respect to the hitting time of an ϵ-ball around the optimum, $E[\tau_{B_\epsilon}]$. Linear convergence corresponds to $E[\tau_{B_\epsilon}]$ increasing like $K(-\ln(\epsilon)) = K\ln(1/\epsilon)$ where $K > 0$. The constants $-1/K$ and c play the same role and will typically decrease like

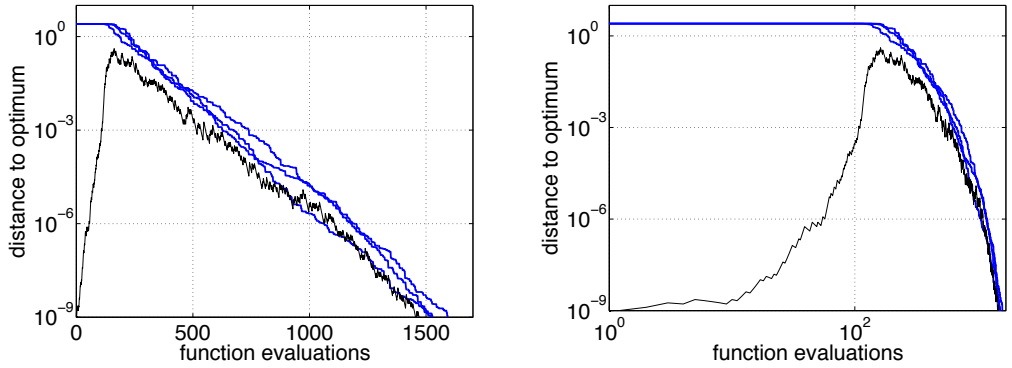

Fig. 10.2. Illustration of the (log)-linear convergence: Time evolution of the distance to the optimum of the (1+1)-ES with one-fifth success rule, three trials respectively, on the function $f : x \in \mathbb{R}^{10} \mapsto g(\|x\|)$, where g is strictly monotonically increasing. The jagged line starting at 10^{-9} shows additionally the step-size of one of the trials with adaptive step-size. The initial step-size has been chosen very small (10^{-9}) compared to the initial search points to show the capability to increase the step-size (during the first 100 evaluations). **Left:** log scale on y-axis: after about 100 evaluations, the logarithm of the distance to the optimum decreases linearly, the slope of the curves corresponds to the convergence rate c defined in Equation (10.10). **Right:** log scale on y-axis and x-axis.

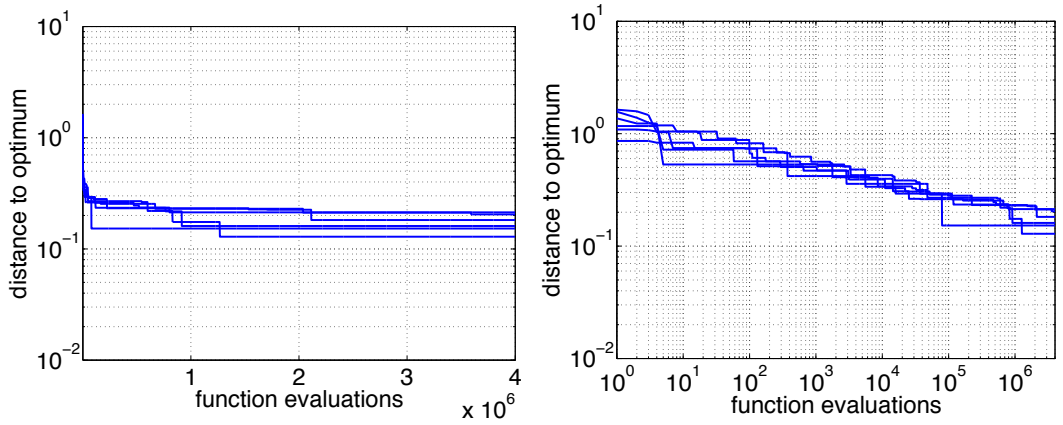

Fig. 10.3. Illustration of sub-linear convergence with degree p: Time evolution of the distance to the optimum from six runs of pure random search in a log-linear plot (left) and a log-log plot (right). Each probe point is sampled uniformly at random in $[-0.2, 0.8]^{10}$ on the function $f : x \mapsto g(\|x\|)$, where g is strictly monotonically increasing. Apart from the scale of the x-axis both plots are identical. In the log-log plot the graphs shape up as linear with only stochastic deviations. According to Theorem 10.8, $p = 1/d$ and here $d = 10$.

$1/d$, i.e., K will grow like d for evolution strategies. Sub-linear convergence with degree p corresponds to $E[\tau_{B_\epsilon}]$ growing like $(1/\epsilon)^{1/p}$.

10.2.5. *Simple Proofs for Convergence*

Convergence alone can be quite easy to establish as we illustrate in the first part of this section. We first recall the following corollary from the Borel-Cantelli Lemma useful for proving almost sure convergence.

Lemma 10.4 (Sufficient condition for convergence). *Let Y_n be a sequence of random variables and Y be a random variable. If for all $\epsilon > 0$, $\sum_n \Pr[|Y_n - Y| > \epsilon] < \infty$, then Y_n converges almost surely to Y.*

Proof. The proof can be found in probability textbooks, see for instance (Karr, 1993, p. 138). □

The previous lemma is implicitly or explicitly used to investigate convergence of evolution strategies in (Baba, 1981; Rudolph, 2001; Greenwood and Zhu, 2001).

In the sequel, we will prove convergence for the local random search given in Equation (10.2) where $(W_k)_{k\in\mathbb{N}}$ are i.i.d. with multivariate normal distribution $\sigma N(0, I_d)$ for $\sigma > 0$ (thus $\mathcal{D} = \mathbb{R}^d$) as common law. The algorithm is thus a (1+1)-ES with fixed step-size equal to σ. The selection is termed elitist as a new candidate solution is selected among the parent and the offspring such that the best solution ever found cannot be lost. The convergence is proven for functions with bounded sublevel sets.

Assumption 1 (bounded sublevel sets). *For the function f, the sublevel set $\{x|f(x) \leq \alpha\}$ is bounded for every $\alpha \in \mathbb{R}$.*

We are now going to prove that $f(\mathrm{X}_n)$ converges almost surely to m by using Lemma 10.4. For $\delta > 0$, we need to prove that

$$\sum_n \Pr[|f(\mathrm{X}_n) - m| > \delta] < \infty \ . \tag{10.12}$$

We denote A_δ the set $\{x \in \mathbb{R}^d, f(x) \leq m+\delta\}$, its complementary A_δ^c satisfies thus $A_\delta^c = \{x \in \mathbb{R}^d, f(x) > m + \delta\}$. Rewriting the condition (10.12), we need to show that

$$\sum_n \Pr[\mathrm{X}_n \in A_\delta^c] < \infty \ . \tag{10.13}$$

Equation (10.13) will follow from (a) a constant lower bound for the probability to hit A_δ at any step and (b) elitist selection that prevents to escape A_δ once it has been hit. First, we sketch the proof idea for (a).

Lemma 10.5. *For any $R > 0$, there exists $\gamma > 0$ such that for all $x \in B_R(0)$, $\Pr[x + W_0 \in A_\delta] \geq \gamma$.*

Proof. We sketch the idea of the proof and leave the technical details to the reader. The definition of m (essential infimum of f) guarantees that $\mathrm{V}ol(A_\delta)$ is strictly positive. The probability that $x + W_0$ hits A_δ is given by the integral over A_δ of the density of a multivariate normal distribution centered in x. This integral will be always larger or equal to the integral when x is placed in $B_R(0)$ and at the largest distance from A_δ. The constant γ will be this latter integral. □

The previous lemma together with elitist selection implies the following lemma:

Lemma 10.6. *Let $R > 0$ such that $\{x, f(x) \leq f(X_1)\} \subset B_R(0)$ and γ the constant from Lemma 10.5, then $\Pr[\mathrm{X}_n \in A_\delta^c] \leq (1-\gamma)^n$.*

Proof. Note first that R does exist because of Assumption 1 and that the elitist selection ensures that for all n, $\mathrm{X}_n \in B_R(0)$. Therefore, $\Pr[\mathrm{X}_n \in A_\delta^c] = \Pr[\mathrm{X}_n \in A_\delta^c \cap B_R(0)]$. Because of the elitist selection, once X_n enters A_δ, the next iterates stay in A_δ such that if X_n does not belong to A_δ it means that X_{n-1} did not belong to A_δ and thus $\Pr[\mathrm{X}_n \in A_\delta^c \cap B_R(0)] = \Pr[\mathrm{X}_n \in A_\delta^c \cap B_R(0), X_{n-1} \in A_\delta^c \cap B_R(0)]$. By the Bayes formula for conditional probabilities, we can write $\Pr[\mathrm{X}_n \in A_\delta^c \cap B_R(0), X_{n-1} \in A_\delta^c \cap B_R(0)]$ as $\Pr[\mathrm{X}_n \in A_\delta^c \cap B_R(0) | X_{n-1} \in A_\delta^c \cap B_R(0)]$ times $\Pr[X_{n-1} \in A_\delta^c \cap B_R(0)]$. However, by Lemma 10.5, $\Pr[\mathrm{X}_n \in A_\delta^c \cap B_R(0) | X_{n-1} \in A_\delta^c \cap B_R(0)] \leq 1 - \gamma$ such that we have now that

$$\Pr[\mathrm{X}_n \in A_\delta^c \cap B_R(0)] \leq (1-\gamma) \Pr[X_{n-1} \in A_\delta^c \cap B_R(0)]$$

and then by induction we obtain that $\Pr[\mathrm{X}_n \in A_\delta^c] = \Pr[\mathrm{X}_n \in A_\delta^c \cap B_R(0)] \leq (1-\gamma)^n$. □

A direct consequence of Lemma 10.6 is the convergence of $f(\mathrm{X}_n)$ in probability since we have that $\Pr[\mathrm{X}_n \in A_\delta^c]$ converges to zero when n grows to infinity. However, using now Lemma 10.6, we can also obtain the stronger convergence, that is almost sure convergence. Indeed, we deduce from

Lemma 10.6 that

$$\sum_{n=1}^{\infty} \Pr[\mathrm{X}_n \in A_\delta^c] \leq \sum_{n=1}^{\infty} (1-\gamma)^n = \frac{1}{\gamma}$$

and thus applying Lemma 10.4, the almost sure convergence of $f(\mathrm{X}_n)$ to m holds. In conclusion, we have proven the following theorem:

Theorem 10.7 (Almost sure convergence of (1+1)-ES). *For f satisfying Assumption 1, the (1+1)-ES with constant step-size $\sigma > 0$ converges almost surely to m the essential infimum of f in the sense that*

$$f(\mathrm{X}_n) \to m \quad a.s.$$

The techniques illustrated in this section can be also used for disproving convergence. Almost sure convergence implies convergence in probability such that a necessary condition for convergence with probability one is

$$\Pr[\mathrm{X}_n \in A_\delta^c] \to 0 \ .$$

Rudolph is using this fact to disprove convergence with probability one of the (1+1)-ES with one-fifth success rule on a multi-modal function (Rudolph, 2001).

Over a compact set and without further assumptions on f, convergence towards x^* holds if $\sigma_n \sqrt{\ln(n)} \to \infty$ (Devroye and Krzyżak, 2002), i.e., if σ_n decreases not too fast, global convergence is preserved. We will see later that for step-size adaptive ESs, the step-size σ_n converges much faster to 0, more precisely $\lim_{n\to\infty} \sigma_n \alpha^n < \infty$ for some $\alpha > 1$.

10.2.6. *Invariance to Order Preserving Transformations*

One important aspect of evolution strategies is their invariance with respect to monotonically increasing transformations of the objective function. Remind that $\mathcal{M}$ denotes the set of strictly increasing transformations on $\mathbb{R}$. For an ES, optimizing f or $g \circ f$ for $g \in \mathcal{M}$ is exactly the same, in that the exact same sequence $(\mathrm{X}_n)_{n\in\mathbb{N}}$ is constructed when optimizing f or $g \circ f$, provided the independent random numbers needed to sample the probe points are the same. This is due to the fact that all the updates are *only* based on the ranking of new solutions (which is preserved if $g \circ f$ is considered instead of f) and not on their absolute objective function value. This property is not true for all stochastic search algorithms, in particular not for the simulated annealing algorithm where the selection of a new probe depends

on the fitness difference between the current solution and the probe point or for genetic algorithms with fitness proportionate selection.

Invariance to strictly increasing transformations is illustrated in Figure 10.4 where the function $f : x \mapsto \|x\|^2$ is plotted (in dimension 1 for the sake of illustration) on the left together with two functions $x \mapsto g(f(x))$ for $g \in \mathcal{M}$ (middle and right). Though the left function seems to be easier to optimize (the function being convex and quadratic), the behavior of ESs on the three functions will be identical, in the sense that the sequences $(\mathrm{X}_n)_n$ will be indistinguishable.

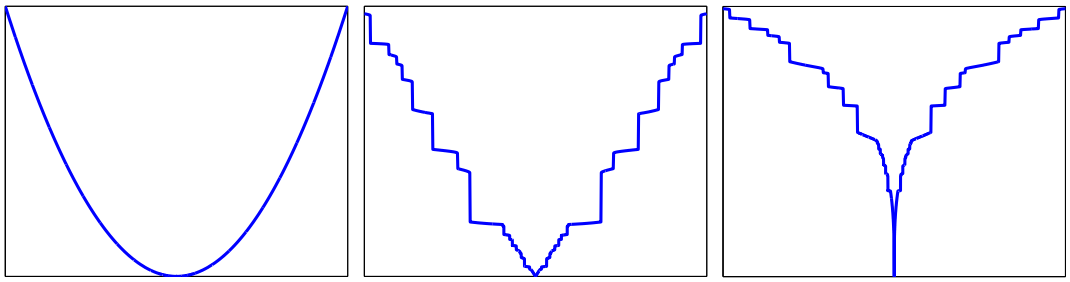

Fig. 10.4. **Left:** sphere function $f : x \mapsto \|x\|^2$. **Middle and right:** $x \mapsto g(f(x))$ for two different $g \in \mathcal{M}$ and $d = 1$. ESs are invariant to $\mathcal{M}$ in the sense that $(\mathrm{X}_n)_n$ generated when optimizing $g \circ f$ for any $g \in \mathcal{M}$ will be the same.

10.3. Rate of Convergence of Non-adaptive Algorithms

In this section, *non-adaptive* means that dispersion parameters of the sampling distribution, like its width and shape, remain constant, however, the sampling distribution can be shifted during the search. The local random search algorithm described in Equation (10.2) is non-adaptive in this sense. The (1+1)-ES described in Equation (10.3) is adaptive, in case σ_n changes over time. Rates of convergence are investigated for the pure random search first and for local random search algorithms afterwards.

10.3.1. *Pure Random Search*

We investigate the pure random search from Equation (10.1). We denote μ the probability measure of the random vectors $(\mathrm{Y}_n)_{n\in\mathbb{N}}$. The probability measure μ characterizes the probability to hit (Borel) sets of $\mathcal{D}$: for a set A included in $\mathcal{D}$, $\Pr[\mathrm{Y}_n \in A] = \mu(A)$. We start by looking at a simple case:

Theorem 10.8 (Case of uniform sampling on the unit cube). *Let f be the spherical function $f(x) = g(\|x - x^*\|)$ where $g \in \mathcal{M}$ and with $x^* \in \mathcal{D} =]0,1[^d$. Let μ be the uniform distribution on $\mathcal{D}$, then for ϵ such that $B_\epsilon(x^*) \subset \mathcal{D}$, the first hitting time of $B_\epsilon(x^*)$ by X_n, i.e., $\tau_{B_\epsilon} = \inf\{n \geq 1 \,|X_n \in B_\epsilon(x^*)\}$ satisfies*

$$E[\tau_{B_\epsilon}] = \frac{\Gamma(\frac{d}{2}+1)}{\pi^{d/2}} \frac{1}{\epsilon^d}. \tag{10.14}$$

Proof. Since the objective function is the sphere function, the hitting time of $B_\epsilon(x^*)$ by X_n corresponds to the hitting time of $B_\epsilon(x^*)$ by Y_n, such that $\tau_{B_\epsilon} = \inf\{n \geq 1 \,|Y_n \in B_\epsilon(x^*)\}$. Let us denote the success probability as $p = \Pr[\mathrm{Y}_n \in B_\epsilon(x^*)]$. Since the random variables $(\mathrm{Y}_n)_{n\in\mathbb{N}}$ are independent, the hitting time τ_{B_ϵ} is the moment where the first success is obtained in the experiment that consists in repeating independently trials with two outcomes "$\mathrm{Y}_n \in B_\epsilon(x^*)$"—corresponding to success—and "$\mathrm{Y}_n \notin B_\epsilon(x^*)$"—corresponding to failure—with probability p and $1-p$. Since the random variables $(\mathrm{Y}_n)_{n\in\mathbb{N}}$ are independent, τ_{B_ϵ} follows a geometric distribution with parameter p and $E[\tau_{B_\epsilon}] = 1/p$.

Since μ is the uniform distribution on $\mathcal{D}$ and $B_\epsilon(x^*) \subset \mathcal{D}$, the probability $p = \Pr[\mathrm{Y}_n \in B_\epsilon(x^*)] = \frac{\mathrm{V}ol(B_\epsilon(x^*))}{\mathrm{V}ol(\mathcal{D})}$. Since $\mathrm{V}ol(\mathcal{D}) = 1$, with Equation (10.4) we obtain the result. □

The arguments used in Theorem 10.8 transfer to more general objective functions and sampling distributions if we consider the hitting time of the set $A_\delta = \{x, f(x) \leq m + \delta\}$, i.e.,

$$\tau_{A_\delta} = \inf\{n,\, \mathrm{X}_n \in A_\delta\} = \inf\{n,\, f(\mathrm{X}_n) \leq m + \delta\} \ . \tag{10.15}$$

We now assume that the search domain $\mathcal{D}$ is bounded and the sampling distribution μ admits a density $p_\mu(x)$ with respect to the Lebesgue measure that satisfies the following assumption.

Assumption 2 (Bounded sample distribution). *There exist $c_2, c_1 > 0$ such that $c_1 \leq p_\mu(x) \leq c_2$ for all $x \in \mathcal{D}$.*

In order to derive the dependence in ϵ and d of the expected hitting time, we also need to make an assumption on the objective function:

Assumption 3 (Bounded volume of A_δ). *There exists $\delta_0 > 0$ such that for all $\delta > 0$ and $\delta \leq \delta_0$, there exists two constants K_1 and K_2 such that*

$$K_1 \epsilon^d \leq \mathrm{V}ol(A_\delta) \leq K_2 \epsilon^d \ .$$

We can now state the following theorem for PRS:

Theorem 10.9. *Let f satisfy A 3 and the sampling distribution of PRS satisfy A 2, then $E[\tau_{A_\delta}] = \Theta(1/\epsilon^d)$. Specifically, the expected hitting time of A_δ satisfies*

$$\frac{1}{K_2 c_2}\frac{1}{\epsilon^d} \leq E[\tau_{A_\delta}] \leq \frac{1}{K_1 c_1}\frac{1}{\epsilon^d}\,. \tag{10.16}$$

Proof. The hitting time τ_{A_δ} defined in Equation (10.15) can be expressed as the hitting time of the variable Y_n, i.e., $\tau_{A_\delta} = \inf\{n,\, \mathrm{Y}_n \in A_\delta\}$. Note that this would not have been the case if we would have considered the hitting time of $B_\epsilon(x^*)$. The same arguments used in Theorem 10.8 hold now here: at each iteration, Y_n hits A_δ with probability $p = \Pr[\mathrm{Y}_n \in A_\delta]$. Because the Y_n are independent, the expectation of τ_{A_δ} equals $1/p$. It remains now to estimate p. Since $p = \int_{A_\delta} p_\mu(x)dx$, using A 2, we have that $c_1 \int_{A_\delta} dx \leq p \leq c_2 \int_{A_\delta} dx$, in other words $c_1 \mathrm{V}ol(A_\delta) \leq p \leq c_2 \mathrm{V}ol(A_\delta)$. Using the bounds of $\mathrm{V}ol(A_\delta)$ from A 3 we obtain the result. □

Both theorems state sub-linear convergence with degree $1/d$ of the pure random search algorithm which is illustrated for the spherical function in Figure 10.3.

10.3.2. *Lower and Upper Bounds for Local Random Search*

The local random search defined via Equation (10.2) includes a (1+1)-ES with constant step-size and is a particular case of the so-called Markov monotonous search algorithms.

10.3.2.1. *A Detour Through Markov Monotonous Search (Zhigljavsky and Zilinskas, 2008)*

Definition 10.10 (Markov montonous search). *A Markov chain sequence $(\mathrm{X}_n)_{n\in\mathbb{N}}$ that moreover satisfies $f(\mathrm{X}_{n+1}) \leq f(\mathrm{X}_n)$ with probability one is called a Markov monotonous search sequence.*

Informally, a sequence $(\mathrm{X}_n)_n$ is a Markov chain if the value of the n-th variable depends on the past variables only through the immediate predecessor. A transition kernel can be associated to a Markov chain:

Definition 10.11 (Transition kernel). *For a Markov chain $(\mathrm{X}_n)_n$ the transition kernels are the collection of $P_n(.,.) : \mathbb{R}^d \times \mathcal{B}(\mathbb{R}^d) \to [0,1]$ where for each n, for each $A \in \mathcal{B}(\mathbb{R}^d)$, $P_n(.,A)$ is a measurable mapping and for all*

$x \in \mathbb{R}^d$, $P_n(x, .)$ is a probability measure that characterizes the distribution of the sequence: $P_n(x, A) = \Pr[\mathrm{X}_{n+1} \in A | \mathrm{X}_n = x]$ for all x, A.

When P_n is independent of n, the Markov chain is homogeneous and thus the local random search is an *homogeneous* Markov monotonous search algorithm. However, if the random variable W_n is, for example, scaled using a cooling schedule, say $1/(n+1)$, i.e., a new probe point is $\mathrm{X}_n + \frac{1}{n+1} W_n$ the algorithm is not homogeneous.

Let Q be the transition kernel $Q(x, A) = \Pr[x + W_0 \in A]$ representing the probability that $x + W_0$ belongs to A. The transition kernel for the local random search can be written as

$$P(x, A) = \delta_x(A) Q(x, B_f^c(x)) + Q(x, A \cap B_f(x)) \ , \tag{10.17}$$

where $B_f(x) = \{y \in \mathbb{R}^d, f(y) \leq f(x)\}$ and $B_f^c(x)$ denote its complement and $\delta_x(.)$ is the probability measure concentrated at x, i.e., $\delta_x(A) = 1$ if $x \in A$ and 0 otherwise.

10.3.2.2. *Upper Bounds for Convergence Rates of Certain Homogeneous Markov Monotonous Search*

The question of how to choose the distribution $Q(x, .)$ to converge "fast" has been addressed by Nekrutkin and Tikhomirov (1993) and Tikhomirov (2006). The main result is that for an appropriate choice of $Q(x, .)$ one can upper bound the expected hitting time of M_ϵ defined as

$$M_\epsilon = \{y \in B_\epsilon(x^*) : f(y) < f(z) \text{ for every } z \in B_\epsilon^c(x^*)\}$$

by $O(\ln^2(1/\epsilon))$ under mild assumptions on the objective function. The result holds for a fixed dimension and thus the constant hidden in the O notation depends on d. We now prove that the density associated to the distribution $Q(x, .)$ needs to have a singularity ~~in 0~~. in x

Theorem 10.12. *The density associated to the distribution $Q(x, .)$ needs to have a singularity ~~in zero~~.* *in x*

Proof. If the density of the sampling distribution is upper bounded, non-adaptive algorithms cannot be faster than random search, because there exists a PRS with uniform density larger than the upper bound over M_ϵ (for small enough ϵ). This PRS has in each iteration a larger probability for hitting M_ϵ and consequently a smaller expected hitting time. Therefore, if the density is upper bounded, the upper bound of the expected hitting

time cannot be essentially faster than $1/\epsilon^d$, i.e., cannot belong to $o(1/\epsilon^d)$. □

The results by Tikhomirov (2006) have been established in the context of homogeneous Markov symmetric search where Q admits a density $q(x, y)$ that can be written as $q(x, y) = g(\|x - y\|)$ where g is a non-increasing, non-negative left-continuous function defined on $\mathbb{R}^+$ and normalized such that $q(0, .)$ is a density. More precisely for a fixed precision ϵ, Tikhomirov (2006) proves that there exists a function g_ϵ that depends on the precision ϵ such that for the associated Markov Monotonous search algorithm, the hitting time of M_ϵ satisfies on average $E_x[\tau_\epsilon] \leq O(\ln^2(1/\epsilon))$. A slightly worse upper bound with a function g independent of the precision ϵ can be obtained (Zhigljavsky, 1991).

10.4. Rate of Convergence of Adaptive ESs

We focus now on the rate of convergence of *adaptive* ESs where the dispersion of the sampling distribution is adapted during the course of the search process. Various results show that adaptive ESs cannot converge faster than linear, with a convergence rate decreasing like $1/d$ when d goes to infinity. This has been in particular proven by Nekrutkin and Tikhomirov (1993) in the context of Markov Monotonous search (i.e., for a (1+1)-ES) without showing the dependency in $1/d$ for the convergence rate though, by Jägersküpper (2008) for (1+λ)-ES with isotropic sampling distributions and by Teytaud and Gelly (2006) for general rank-based algorithms.

10.4.1. *Tight Lower Bounds for Adaptive ESs*

In this section, we establish *tight constants for the lower bounds* associated to the (1+1)-ES with adaptive step-size defined in Equation (10.3) and explain how the results generalize to the (1,λ)-ES with adaptive step-size. The parent and step-size of an adaptive (1+1) or (1,λ)-ES are denoted (X_n, σ_n). Both algorithms sample new points by adding to X_n random vectors following a *spherical multivariate normal distribution* scaled by σ_n. Whereas a single new probe point is created for the (1+1)-ES, λ *independent* probe points are created in the case of the (1,λ)-ES where the best among the λ points becomes subsequently the new parent.

We start by defining a specific artificial step-size adaptation rule where the step-size is proportional to the distance to the optimum of the function

to optimize. This algorithm is essentially relevant for spherical functions as we will illustrate, however it can be defined for an arbitrary function f:

Definition 10.13 (Scale-invariant step-size). *Let x^* be the optimum of a function f to optimize and (X_n, σ_n) be the parent and step-size at iteration n of a (1,λ) or (1+1)-ES. If $\sigma_n = \sigma\|X_n - x^*\|$ for $\sigma > 0$, the step-size is called scale-invariant.*

We now define the expected log-progress towards a solution x^*.

Definition 10.14 (Expected log-progress). *Let (X_n, σ_n) be the parent and step-size of an adaptive ES, we define the expected conditional log-progress towards a solution x^* at iteration n as*

$$\varphi_{\ln}(\mathrm{X}_n, \sigma_n) := E\left[\ln \frac{\|X_n - x^*\|}{\|X_{n+1} - x^*\|}\Big| X_n, \sigma_n\right] \ . \tag{10.18}$$

The previous definition implicitly assumes that (X_n, σ_n) is a Markov Chain but can be adapted to more general settings by replacing the conditioning with respect to (X_n, σ_n) by conditioning with respect to the past in Equation (10.18). In the next lemma, we define the function $F_{(1+1)}$. We will then give its interpretation in terms of expected log-progress.

Lemma 10.15 (Jebalia *et al.* (2008)). *Let $\mathcal{N}$ be a random vector of distribution $N(0, I_d)$. The map $F_{(1+1)} : [0, +\infty] \to [0, +\infty]$ defined by $F_{(1+1)}(\sigma) := E\left[\ln^-(\|e_1 + \sigma\mathcal{N}\|)\right]$, $F_{(1+1)}(+\infty) := 0$ and that can be written as*

$$F_{(1+1)}(\sigma) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} \ln^- \|e_1 + \sigma x\| e^{-\frac{\|x\|^2}{2}} dx \,, \tag{10.19}$$

otherwise, is continuous on $[0, +\infty]$ (endowed with the usual compact topology), finite valued and strictly positive on $]0, \infty[$.

We will now prove that for $\sigma \in [0, +\infty[$, $F_{(1+1)}(\sigma)$ is equal to (i) the expected log-progress achieved on the spherical functions $g(\|x\|), g \in \mathcal{M}$ by the (1+1)-ES starting from $e_1 = (1, 0, \ldots)$ and with step-size σ; (ii) the expected log-progress of the (1+1)-ES with scale-invariant step-size. We formalize and prove those results in the next lemma.

Lemma 10.16. *Let $\sigma \in [0, +\infty[$, on the class of spherical functions $f(x) = g(\|x\|), g \in \mathcal{M}$, $F_{(1+1)}(\sigma)$ coincides with (i) the expected log-progress of a (1+1)-ES starting from e_1 and with step-size σ, i.e.,*

$$E\left[\ln \frac{\|X_n\|}{\|X_n + \sigma_n\mathcal{N}\| \wedge 1}\Big| X_n = e_1, \sigma_n = \sigma\right] = F_{(1+1)}(\sigma) \ ,$$

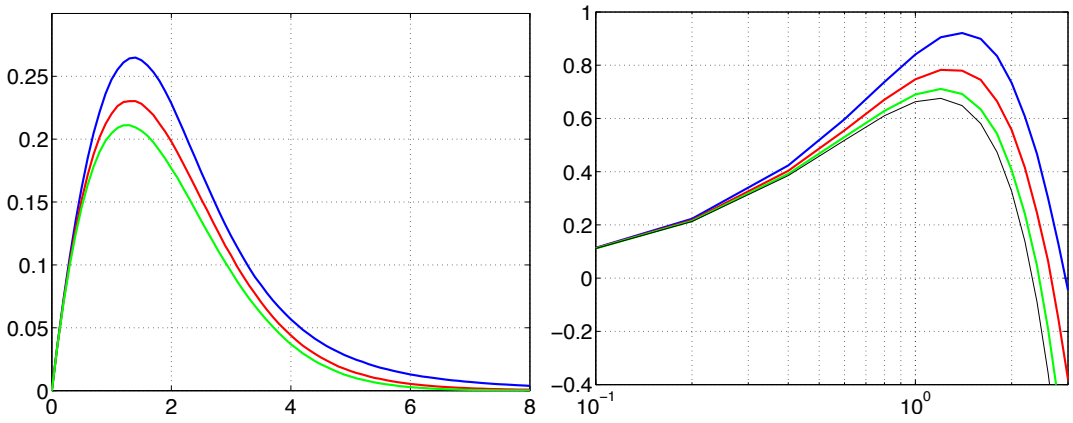

Fig. 10.5. **Left:** Plot of $\sigma \mapsto dF_{(1+1)}(\sigma/d)$ where $F_{(1+1)}$ is defined in Equation (10.19) for $d = 5, 10, 30$ (top to bottom). **Right:** Plot of $\sigma \mapsto dF_{(1,5)}(\sigma/d)$ where $F_{(1,5)}$ is defined in Equation (10.25) for $d = 5, 10, 30$, top to bottom. The lowest line is the limit of $\sigma \mapsto dF_{(1,5)}(\sigma/d)$ for d to infinity given in Equation (10.26).

(ii) the expected log-progress of the (1+1)-ES with scale-invariant step-size ($\sigma_n = \sigma\|\mathrm{X}_n\|$) at any iteration n, i.e., for all $n \in \mathbb{N}$

$$\varphi_{\ln}(\mathrm{X}_n, \sigma_n) = F_{(1+1)}(\sigma) \quad .$$

Proof. Starting from $X_n = e_1$, a new search point sampled with a step-size σ denoted $e_1 + \sigma\mathcal{N}$ is accepted if its objective function $g(\|e_1 + \sigma\mathcal{N}\|)$ is smaller than the objective function of e_1, which is equal to $g(1)$. Thus, $\|\mathrm{X}_{n+1}\|$ is the minimum between $g(\|e_1 + \sigma\mathcal{N}\|)$ and $g(1)$. The expected log-progress will therefore be equal to $\ln\|e_1\| - E\left[\ln\left(\|e_1 + \sigma\mathcal{N}\| \wedge 1\right)\right]$ which simplifies to $E[\ln^- \|e_1 + \sigma\mathcal{N}\|]$ with $\ln^-(x) := \max(0, -\ln(x))$. The second point (ii) will be proven together with Theorem 10.17 (see Equation (10.43)). □

Plots of the function $F_{(1+1)}$ for different dimensions are given in Figure 10.5. For a given dimension d, *minus the maximum of the function $F_{(1+1)}$ is a lower bound for the convergence rate of step-size adaptive (1+1)-ESs* as stated in the following theorem.

Theorem 10.17 (Convergence at most linear). *Let $(\mathrm{X}_n, \sigma_n)_n$ be generated by a (1+1)-ES with adaptive step-size, on any function f. Let y^* be any vector in $\mathbb{R}^d$ and $E[|\ln\|\mathrm{X}_n - y^*\|\|] < +\infty$ for all n. Then*

$$E\left[\ln\|\mathrm{X}_{n+1} - y^*\|\right] \geq E\left[\ln\|\mathrm{X}_n - y^*\|\right] - \tau\,, \tag{10.20}$$

where τ is a strictly positive constant defined as $\tau = \sup F_{(1+1)}([0, +\infty])$ where $F_{(1+1)}$ is the real valued function defined in Equation (10.19). In

particular, the convergence is at most linear with the best convergence rate being $-\tau$ in the sense that for all $n_0 \in \mathbb{N}$

$$\inf_{n\in\mathbb{N}, n>n_0} \frac{1}{n-n_0} E \ln \frac{\|\mathrm{X}_n - y^*\|}{\|X_{n_0} - y^*\|} \geq -\tau \ . \tag{10.21}$$

The proof of Theorem 10.17 is presented on page 322.

Remark 10.18. Optimality can also be formulated with respect to the expectation of $\|\mathrm{X}_n - y^*\|$ instead of $\ln \|\mathrm{X}_n - y^*\|$ in the following manner: Assume $E[\|\mathrm{X}_n - y^*\|] < \infty$ for all n, then

$$E\|\mathrm{X}_{n+1} - y^*\| \geq E[\|\mathrm{X}_n - y^*\|] \ \tau^{'} \ , \tag{10.22}$$

where $\tau^{'} = \min \tilde{F}_{(1+1)}([0, +\infty])$ with $\tilde{F}_{(1+1)}(\sigma) = E[\|e_1 + \sigma\mathcal{N}\| \wedge 1]$.

The two formulations are not equivalent. The constants $-\tau$ and $\ln(\tau^{'})$ play the same role but are not equal due to Jensen's inequality that implies for all σ

$$-F_{(1+1)}(\sigma) = E[\ln(\|e_1 + \sigma\mathcal{N}\| \wedge 1)] < \ln E[\|e_1 + \sigma\mathcal{N}\| \wedge 1] = \ln(\tilde{F}_{(1+1)}(\sigma)) \ .$$

The formulation with the logarithm inside the expectation is compatible with almost sure convergence (Auger and Hansen, 2006).

We will now prove that the lower bound given in Equation (10.21) is reached on spherical functions by the (1+1)-ES with scale-invariant step-size and an appropriate choice of σ. However before to state this result we formulate the linear convergence in expectation of the (1+1)-ES with scale-invariant step-size.

Proposition 10.19. *On spherical functions, $f(x) = g(\|x - x^*\|), g \in \mathcal{M}$, the (1+1)-ES with scale-invariant step-size ($\sigma_n = \sigma\|\mathrm{X}_n - x^*\|$) converges linearly in expectation, moreover for all $n_0 \in \mathbb{N}$ and for all $n > n_0$*

$$\frac{1}{n-n_0} E\left[\ln \frac{\|\mathrm{X}_n - x^*\|}{\|X_{n_0} - x^*\|}\right] = -F_{(1+1)}(\sigma) \ . \tag{10.23}$$

The proof of Proposition 10.19 is presented on page 323. As a consequence, lower bounds are reached for the (1+1)-ES with scale-invariant step-size and σ chosen to maximize $F_{(1+1)}$.

Corollary 10.20. (Lower bound reached for ES with scale-invariant step-size) *The lower bound in Equation (10.20) is reached on spherical functions $f(x) = g(\|x - x^*\|)$ with $g \in \mathcal{M}$ for the scale-invariant*

step-size rule where at each $n \in \mathbb{N}$, $\sigma_n = \sigma_{(1+1)}\|\mathrm{X}_n - x^*\|$ *with* $\sigma_{(1+1)} > 0$ *such that* $F_{(1+1)}(\sigma_{(1+1)}) = \tau$. *Moreover, for all* $n_0 \in \mathbb{N}$ *and for all* $n > n_0$

$$\frac{1}{n - n_0} E\left[\ln \frac{\|\mathrm{X}_n - x^*\|}{\|X_{n_0} - x^*\|}\right] = -\tau \ . \tag{10.24}$$

The proof is presented together with the proof of Proposition 10.19 on page 323.

Equations (10.23) and (10.24) imply linear convergence in expectation as defined in Equation (10.11). However it does *not only* hold *asymptotically* but also for any *finite* n and n_0. We will prove in addition later that almost sure linear convergence also holds.

The constant $-\tau$ corresponds thus to the convergence rate on spherical functions of the (1+1)-ES with scale-invariant step-size where at each iteration n, $\sigma_n = \sigma_{(1+1)}\|X_n - x^*\|$. Because the convergence rate is reached, the lower bound $-\tau$ is tight. Those results were presented in (Jebalia *et al.*, 2008). They hold for the (1+1)-ES, but the same analysis can be applied to a (1,λ)-ES, resulting in optimality of the scale-invariant step-size (1,λ)-ES where $\sigma_{(1,\lambda)}$ realizes the maximum of $F_{(1,\lambda)}$ defined as the expected log-progress of a (1,λ)-ES

$$F_{(1,\lambda)}(\sigma) = -E\left[\ln \min_{1 \le i \le \lambda} \|e_1 + \sigma \mathcal{N}_i\|\right] , \tag{10.25}$$

where $\mathcal{N}_i$ are λ independent random vectors distributed as $N(0, I_d)$ (Auger and Hansen, 2006).

10.4.2. *Link with Progress Rate Theory*

Developments of ESs are closely related to the so-called progress-rate theory (Rechenberg, 1973) that constitutes the main core of the book called "The Theory of Evolution Strategies" (Beyer, 2001). In this section we explain the progress rate approach and its connexions with the convergence of scale-invariant step-size ESs.

The progress rate is defined as a one-step expected progress towards the optimal solution. Assuming w.l.o.g. that the optimum is $x^* = 0$, we can define the normalized expected progress φ^* as

$$\varphi^* = d\, E\left[\frac{\|\mathrm{X}_n\| - \|\mathrm{X}_{n+1}\|}{\|\mathrm{X}_n\|}\right] = d\left(1 - E\left[\frac{\|\mathrm{X}_{n+1}\|}{\|\mathrm{X}_n\|}\right]\right) \ .$$

The normalized log-progress can also be considered

$$\varphi^*_{\ln} = d\left(E\left[\ln \frac{\|\mathrm{X}_n\|}{\|\mathrm{X}_{n+1}\|}\right]\right) \ .$$

It coincides with the expectation of Equation (10.18) times d. For an ES with a spherical search distribution and on the sphere function, we define additionally

$$\sigma^* = d\,\sigma_n / \|\mathrm{X}_n\| \ .$$

As we will see in Section 10.4.3.2, the sequence $(\sigma_n/\|\mathrm{X}_n\|)_n$ is an homogeneous Markov chain. To take out the dependency of σ^* in n, it is in addition assumed that σ^* *is constant*, and thus that the step-size is scale-invariant with $\sigma_n = \sigma^* \|\mathrm{X}_n\|/d$. Consequently, the normalized progress φ^* and φ^*_{ln} are functions of σ^* that are independent of n (the proof of this fact is similar to the proof of Lemma 10.16) and of further initial values. Moreover φ^*_{ln} *equals the convergence rate of ESs with scale-invariant step-size multiplied by* d, for example for the (1,λ), for all σ^*

$$dF_{(1,\lambda)}(\sigma^*/d) = \varphi^*_{\mathrm{ln}}(\sigma^*) \ ,$$

or see Proposition 10.19 for the case of the (1+1)-ES. The function φ^*_{ln} as a function of σ^* was plotted in Figure 10.5. Progress rate derivations are in general asymptotic for d to infinity so as to provide comprehensive quantitative estimates for convergence rates. In general, in the limit for $d \to \infty$, φ^* and φ^*_{ln} coincide (Auger and Hansen, 2006). As an example of a simple asymptotic formula, we give the asymptotic progress φ^* (or φ^*_{ln}) on the sphere function for the (1,λ)-ES,

$$\lim_{d\to\infty} \varphi^*(\sigma^*) = c_{1,\lambda}\sigma^* - \frac{{\sigma^*}^2}{2} \,, \tag{10.26}$$

where $c_{1,\lambda}$ is the expected value of the maximum of λ standard normal distributions and usually is in the order of one.

10.4.3. *Linear Convergence of Adaptive ESs*

We have seen that convergence of adaptive ESs is at most linear and have proven, on spherical functions, the linear convergence *in expectation* of the artificial scale-invariant step-size (1+1)-ES where the step-size is scaled to the distance to the optimum. In this section, we explain how the linear convergence of real adaptation schemes can be analyzed. We assume that f is a spherical function, $f(x) = g(\|x\|)$ for g in $\mathcal{M}$. We will first present the proof of *almost sure* convergence of the (1+1)-ES with scale-invariant step-size and will illustrate afterwards that the convergence of real step-size adaptation schemes is the natural extension of this result.

10.4.3.1. *Almost Sure Linear Convergence of the (1+1)-ES Scale-invariant Step-size*

We remind that for the (1+1)-ES with scale-invariant step-size on $g(\|x\|), g \in \mathcal{M}$, for each $n \in \mathbb{N}$ we have $\sigma_n = \sigma\|\mathrm{X}_n\|$ with $\sigma \geq 0$. A new probe point $\mathrm{X}_n + \sigma\|\mathrm{X}_n\|N_n$ is accepted if $\|\mathrm{X}_n + \sigma\|\mathrm{X}_n\|N_n\| \leq \|\mathrm{X}_n\|$ or normalizing by $\|\mathrm{X}_n\|$ if $\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma N_n\right\| \leq 1$. Therefore $\|\mathrm{X}_{n+1}\|/\|\mathrm{X}_n\|$ satisfies

$$\frac{\|\mathrm{X}_{n+1}\|}{\|\mathrm{X}_n\|} = \left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma N_n 1_{\{\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma N_n\|\leq 1\}}\right\| \quad , \tag{10.27}$$

where $1_{\{\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|}+\sigma N_n\|\leq 1\}} = 1$ if $\left\|\frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \sigma N_n\right\| \leq 1$ and zero otherwise.

We connect now linear convergence and law of large numbers by stating the following technical lemma.

Lemma 10.21. *For $n \geq 2$, the following holds*

$$\frac{1}{n}\ln\frac{\|\mathrm{X}_n\|}{\|X_0\|} = \frac{1}{n}\sum_{k=0}^{n-1}\ln\frac{\|\mathrm{X}_{k+1}\|}{\|\mathrm{X}_k\|} \quad , a.s. \tag{10.28}$$

The proof of the lemma is trivial: using the property $\ln(a) + \ln(b) = \ln(ab)$ for all $a, b > 0$ we find that both sides equal $n^{-1}\ln\prod_{k=0}^{n-1}\|\mathrm{X}_{k+1}\|/\|\mathrm{X}_k\|$. Linear convergence defined in Equation (10.10) means that the left-hand side (and thus the RHS) of Equation (10.28) converges to a constant. In order to prove linear convergence, we exploit the fact that the right-hand side is the sum of n random variables divided by n, suggesting the use of a Law of Large Numbers (LLN):

Lemma 10.22 (LLN for independent random variables).
Let $(Y_n)_{n\in\mathbb{N}}$ be a sequence of independent, identically distributed, integrable ($E[|Y_0|] < +\infty$) random variables. Then

$$\frac{1}{n}\sum_{k=0}^{n-1} Y_k \xrightarrow[n\to\infty]{} E[Y_0]\, a.s.$$

In order to apply Lemma 10.22 to the (1+1)-ES with scale-invariant step-size, it remains to be shown that the summands in the right-hand side of Equation (10.28) are i.i.d. and integrable random variables:

Proposition 10.23. *For the (1+1)-ES with scale-invariant step-size, $(\ln(\|\mathrm{X}_{n+1}\|/\|\mathrm{X}_n\|) : n \in \mathbb{N})$ are independent identically distributed as*

$\ln^-(\|e_1 + \sigma\mathcal{N}\|)$ *where* $\mathcal{N}$ *is a random vector following the distribution* $N(0, I_d)$.

For the technical details of the proof we refer to (Jebalia *et al.*, 2009, Lemma 7) where the result was proven in a slightly different setting where the objective function include noises. A weaker result stating that the random variables are orthogonal was proven in (Jebalia *et al.*, 2008). Moreover, we have seen in Lemma 10.15 that $\ln^- \|e_1 + \sigma\mathcal{N}\|$ is integrable such that we can apply Lemma 10.22 and together with Lemma 10.21 obtain the almost sure linear convergence of $\frac{1}{n} \ln \|\mathrm{X}_n\|/\|X_0\|$:

Theorem 10.24. *The (1+1)-ES with scale-invariant step-size converges linearly almost surely on the sphere function:*

$$\lim_{n\to\infty} \frac{1}{n} \ln \frac{\|\mathrm{X}_n\|}{\|X_0\|} = E[\ln^- \|e_1 + \sigma\mathcal{N}\|] = F_{(1+1)}(\sigma), a.s.$$

The idea of using Laws of Large Numbers for analyzing the convergence of evolution strategies was introduced in (Bienvenüe and François, 2003) and used for analyzing ESs with scale-invariant step-size in (Auger and Hansen, 2006; Jebalia *et al.*, 2008, 2009).

10.4.3.2. *How to Analyze Linear Convergence of Real Step-size Adaptation Schemes?*

The linear convergence of real adaptation schemes will also follow from applying a Law of Large Numbers. However, contrary to the scale-invariant step-size case, the sequence $(\ln(\|\mathrm{X}_{n+1}\|/\|\mathrm{X}_n\|), n \in \mathbb{N})$ is not independent and a different LLN needs thus to be apply. We will illustrate the different steps of the analysis exemplary for the one-fifth success rule that we define now precisely on spherical functions $g(\|x\|)$, $g \in \mathcal{M}$. Assume (X_n, σ_n) are given, the next iterate $(\mathrm{X}_{n+1}, \sigma_{n+1})$ is constructed in the following manner:

$$\mathrm{X}_{n+1} = \mathrm{X}_n + \sigma_n N_n 1_{\{\|\mathrm{X}_n + \sigma_n N_n\| \leq \|\mathrm{X}_n\|\}}, \tag{10.29}$$

$$\sigma_{n+1} = \sigma_n \left(\alpha^{-1/4} + (\alpha - \alpha^{-1/4}) 1_{\{\|\mathrm{X}_n + \sigma_n N_n\| \leq \|\mathrm{X}_n\|\}}\right) \ , \tag{10.30}$$

where $\alpha > 1$, i.e., the step-size is multiplied by α in case of success and divided by $\alpha^{1/4}$ otherwise such that Equation (10.30) can be rewritten

$$\sigma_{n+1} = \begin{cases} \alpha\sigma_n & \text{if } \|\mathrm{X}_n + \sigma_n N_n\| \leq \|\mathrm{X}_n\| \ , \\ \alpha^{-1/4}\sigma_n & \text{otherwise,} \end{cases} \tag{10.31}$$

and $(X_0, \sigma_0) \in \mathbb{R}^d \times \mathbb{R}^+$. The equivalent of Equation (10.27) is now

$$\frac{\|\mathrm{X}_{n+1}\|}{\|\mathrm{X}_n\|} = \left\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \frac{\sigma_n}{\|\mathrm{X}_n\|} N_n 1_{\{\| \frac{\mathrm{X}_n}{\|\mathrm{X}_n\|} + \frac{\sigma_n}{\|\mathrm{X}_n\|} N_n \| \leq 1\}} \right\| , \tag{10.32}$$

where σ is replaced by $\sigma_n/\|\mathrm{X}_n\|$. Because Equation (10.32) depends on the random variable $\sigma_n/\|\mathrm{X}_n\|$, the random sequence $(\ln(\|\mathrm{X}_{n+1}\|/\|\mathrm{X}_n\|), n \in \mathbb{N})$ will not be independent and, thus, the LLN for independent random variables cannot be applied. However, $\|\mathrm{X}_n\|/\sigma_n$ is a Markov chain whose distribution can be defined in a simple way. Let $Z_0 = \|X_0\|/\sigma_0$, and define

$$Z_{n+1} = \frac{1}{\alpha^*} \| Z_n e_1 + N_n 1_{\{\|Z_n e_1 + N_n\| \leq Z_n\}} \| , \tag{10.33}$$

where $\alpha^* = \alpha^{-1/4} + (\alpha - \alpha^{-1/4}) 1_{\{\|Z_n e_1 + N_n\| \leq Z_n\}}$, i.e., corresponding to the multiplicative factor in Equation (10.30). Then it is clear that Z_n is a Markov Chain and not difficult to show that Z_n follows the same distribution as $\|\mathrm{X}_n\|/\sigma_n$. The Markov chain Z_n can be exploited to prove linear convergence thanks to the following lemma.

Lemma 10.25. *The following equality holds in distribution*

$$\frac{1}{n} \ln \frac{\|\mathrm{X}_n\|}{\|X_0\|} = \frac{1}{n} \sum_{k=0}^{n-1} \ln \frac{\left\| Z_k e_1 + N_k 1_{\{\|Z_k e_1 + N_k\| \leq Z_k\}} \right\|}{Z_k} \ . \tag{10.34}$$

The summands of the right-hand side of Equation(10.34) correspond to replacing in the right-hand side of Equation (10.32), $\sigma_n/\|\mathrm{X}_n\|$ by $1/Z_n$ and $\mathrm{X}_n/\|\mathrm{X}_n\|$ by e_1. The proof of this lemma is similar to the proof of Lemma 3 in (Jebalia *et al.*, 2009). Its main ingredients are the isotropy of the sampling distribution and of spherical functions. In addition, with Equation (10.33) we have $\alpha^* Z_{k+1} = \left\| Z_k e_1 + N_k 1_{\{\|Z_k e_1 + N_k\| \leq Z_k\}} \right\|$ and thus in distribution

$$\frac{1}{n} \ln \frac{\|\mathrm{X}_n\|}{\|X_0\|} = \frac{1}{n} \sum_{k=0}^{n-1} \ln \frac{\alpha^* Z_{k+1}}{Z_k} \ . \tag{10.35}$$

Since (Z_n) is a Markov chain, Equations (10.34) or (10.35) suggest to apply a LLN for Markov chains. However, not all Markov chains satisfy a LLN. The properties needed to satisfy a LLN are so-called stability properties, namely φ-irreducibility, Harris recurrence and positivity that are explained in the next following paragraphs.

Given a homogeneous Markov chain $(Z_n)_n \subset \mathbb{R}$, with transition kernel $P(.,.)$ and denoting $\mathcal{B}(\mathbb{R})$ the Borel sigma-algebra on $\mathbb{R}$, $(Z_n)_n$ is φ-irreducible if there exists a measure φ such that:

$$\forall (x, A) \in \mathbb{R} \times \mathcal{B}(\mathbb{R}), \varphi(A) > 0, \exists\ n_0 \geq 0 \text{ such that } P^{n_0}(x, A) > 0\,, \quad (10.36)$$

where $P^{n_0}(x, A)$ equals $\Pr[Z_{n_0} \in A | Z_0 = x]$. Another equivalent definition for the φ-irreducibility of the Markov chain $(Z_n)_n$ is: for all $x \in \mathbb{R}$ and for all $A \in \mathcal{B}(\mathbb{R})$ such that $\varphi(A) > 0$, $\Pr[\tau_A < +\infty | Z_0 = x] > 0$, where τ_A is the hitting time of Z_n on A, i.e.,

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

If the last term of Equation (10.36) is equal to one, the chain is *recurrent*. A φ-irreducible chain $(Z_n)_n$ is *Harris recurrent* if:

$$\forall A \in \mathcal{B}(\mathbb{R}) \text{ such that } \varphi(A) > 0; \Pr[\eta_A = \infty | Z_0 = x] = 1, x \in \mathbb{R}\,,$$

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

$$\pi(A) = \int_{\mathbb{R}} P(x, A) \mathrm{d}\pi(x), A \in \mathcal{B}(\mathbb{R})\,.$$

If in addition this measure is a probability measure, the chain is called *positive*. Positive, Harris-recurrent chains satisfy a LLN as stated in (Meyn and Tweedie, 1993, Theorem 17.0.1) and recalled here.

Theorem 10.26 (LLN for Harris positive chains). *Suppose that $(Z_n)_n$ is a positive Harris chain with invariant probability measure π, then for any function G, satisfying $\pi(|G|) := \int |G(x)| \mathrm{d}\pi(x) < \infty$, holds*

$$\lim_{n \to \infty} \frac{1}{n} \sum_{k=0}^{n-1} G(Z_k) = \pi(G)\,. \quad (10.37)$$

Therefore, in order to prove linear convergence of the (1+1)-ES with one-fifth success rule, one can investigate the stability of Z_n and prove that Theorem 10.26 applies to the right-hand side of Equation (10.34) deducing thus the convergence of $\frac{1}{n} \ln (\|\mathrm{X}_n\| / \|X_0\|)$[d].

[d] In fact, the right-hand side of Equation (10.34) can be written $\frac{1}{n} \sum_k G(Z_k, N_k)$ with $G(z, N_n) = \ln \|z e_1 + N_n 1_{\{\|z e_1 + N_n\| \leq z\}}\| / z$ such that one needs to study the stability of the Markov chain (Z_k, N_k). However, the stability of (Z_k, N_k) is a direct corollary of the stability of Z_k since N_k is independent of Z_k.

Theorem 10.27. *If Z_n is φ-irreducible, Harris-recurrent and positive with invariant probability measure π and*

$$\int_{\mathbb{R}} E\left[|\ln \|ze_1 + N_0 1_{\{\|ze_1+N_0\|\leq z\}}\|/z|\right] d\pi(z) < \infty \ ,$$

then the (1+1)-ES with one-fifth success rule converges linearly, more precisely

$$\frac{1}{n} \ln \frac{\|\mathrm{X}_n\|}{\|X_0\|} \xrightarrow[n\to\infty]{} \int_{\mathbb{R}} E\left[\ln \|ze_1 + N_0 1_{\{\|ze_1+N_0\|\leq z\}}\|/z\right] d\pi(z) \ . \quad (10.38)$$

Assuming that $\|X_0\|/\sigma_0$ is distributed according to π, we can formulate a non-asymptotic linear convergence result (the equivalent of Equation (10.24)):

Theorem 10.28. *If Z_n is φ-irreducible, Harris-recurrent and positive with invariant probability measure π,*

$$\int_{\mathbb{R}} E[|\ln \|ze_1 + N_0 1_{\{\|ze_1+N_0\|\leq z\}}\|/z|] d\pi(z) < \infty,$$

and $Z_0 \sim \pi$, then for all $n_0 \in \mathbb{N}$ and for $n > n_0$

$$\frac{1}{n-n_0} E\left[\ln \frac{\|\mathrm{X}_n\|}{\|X_{n_0}\|}\right] = \int_{\mathbb{R}} E\left[\ln \|ze_1 + N_0 1_{\{\|ze_1+N_0\|\leq z\}}\|/z\right] d\pi(z).$$

Proving the stability is in general the most difficult part in the analysis and has been achieved for the (1,λ)-ES with self-adaptation for $d = 1$ using drift conditions (Auger, 2005). The convergence rate in Equation (10.38) is expressed implicitly by means of the invariant distribution of the chain Z_n. However, it is also possible to derive a central limit theorem to characterize the convergence of Equation (10.38) and, then derive confidence intervals for a Monte Carlo estimation of the convergence rate. For this, a stronger stability property needs to be satisfied, namely the geometric ergodicity. Moreover, using the Fubini theorem, it is possible to prove that $\frac{1}{n}\ln\sigma_n$ converges to the same limit than $\frac{1}{n}\ln\frac{\|\mathrm{X}_n\|}{\|X_0\|}$ and to prove an alternative expression for the convergence rate, namely

$$\int_{\mathbb{R}} E\left[\ln \|ze_1 + N_0 1_{\{\|ze_1+N_0\|\leq z\}}\|/z\right] d\pi(z) = \int_{\mathbb{R}} E[\ln(\alpha^*(z))] d\pi(z)$$

where $\alpha^*(z)$ is the multiplicative factor for the step-size change, i.e., $\alpha^*(z) = \alpha^{-1/4} + (\alpha - \alpha^{-1/4}) 1_{\{\|ze_1+N_0\|\leq z\}}$ (Auger, 2005). The fact that both σ_n and $\|\mathrm{X}_n\|$ converge (log)-linearly at the same rate can be observed on the

left plot in Figure 10.2 where we observe the same rate for the decrease of $\ln \|\mathrm{X}_n\|$ and $\ln \sigma_n$.

The link between stability of the normalized chain $\|\mathrm{X}_n\|/\sigma_n$ and linear convergence or divergence of ESs was first pointed out in Bienvenüe and François (2003) and exploited in Auger (2005). Beyond the fact that stability of Z_n implies linear convergence, it is interesting to note that the stability is a natural generalization of the scale-invariant step-size update rule. Indeed, stability implies that after a transition phase, the distribution of $\|\mathrm{X}_n\|/\sigma_n$ will be close to the invariant distribution π, i.e., $\|\mathrm{X}_n\|/\sigma_n \approx \pi$ whereas for the algorithm with scale-invariant step-size we have $\|\mathrm{X}_n\|/\sigma_n = \sigma$. In other words, the scale-invariant step-size rule approximates π by a constant. The benefit of this simplification is the possibility to derive explicit formulae for the convergence rates for d to infinity. The transition phase is illustrated in Figure 10.2 where the experiment was started in the tail of the invariant distribution: the step-size was chosen very small equal to 10^{-9} such that Z_0 is very large. The adaptation stage lasts up to the iteration 150. Afterwards $Z_n = \|\mathrm{X}_n\|/\sigma_n$ "looks" stable as both $\|\mathrm{X}_n\|$ and σ_n decrease at the same rate.

Other approaches to investigate linear convergence have been used on the sphere and certain convex quadratic functions by Jägersküpper (2007, 2005) who derives lower and upper bounds on the time needed to halve the distance to the optimum for a special one-fifth success rule algorithm. With such an approach, it is possible to derive the dependence in the dimension of the convergence rate. However, the approach seems less general in terms of step-size update rules that can be tackled (Jägersküpper and Preuss, 2008).

10.5. Discussion and Conclusion

Stochastic optimization algorithms for numerical optimization are studied in different communities taking different viewpoints. Showing (or disproving) global convergence on a broad class of functions is often a comparatively easy task. In contrast, proving an associated *rate* of convergence, or convergence speed, is often much more intricate. In particular fast, i.e., linear convergence, with running times proportional to $d \log 1/\epsilon$, can only be proven on comparatively restricted classes of functions or in the vicinity of a well-shaped optimum. Here, d is the search space dimension and ϵ is a precision to reach. Linear convergence is a general lower bound: rank-based

algorithms (and thus ESs) cannot be faster than linear with a convergence rate decreasing like $1/d$ (see also Chapter 11).

We believe that global convergence is per se rather meaningless in practice. The (1+1)-ES with fixed step-size as well as the pure random search converge with probability one to the global optimum of functions belonging to a broad class, where the main assumption is that a neighbourhood of the global optimum should be reachable by the search distribution with a positive probability. However, the convergence rate is sub-linear with degree $1/d$, therefore, the running time is proportional to $(1/\epsilon)^d$. Even for moderate dimension, e.g., $d = 10$, this is prohibitively slow in practice. More promising upper bounds for the convergence rate can be achieved for non-adaptive algorithms, when the sampling distribution admits a singularity. For a sampling distribution chosen depending on ϵ, the bound is $O(\ln^2(1/\epsilon))$, and it is slightly worse if we relax the dependency in ϵ. Corresponding practical algorithms have yet to be implemented.

Adaptive ESs, however, do not converge to the global optimum with probability one on such broad classes of functions, because they might never recover from converging to a local optimum (Rudolph, 2001). Instead, adaptive ESs have been shown to achieve linear convergence on restricted function classes. For example, Jägersküpper (2007) lower and upper bounds the time to halve the distance to the optimum with the (1+1)-ES with one-fifth success rule on special ellipsoidal functions. Linear convergence can also be investigated using the Markov chain $\|\mathrm{X}_n\|/\sigma_n$. We have illustrated the corresponding proof techniques in the context of evolution strategies.

One might argue that linear convergence results on convex quadratic functions are weak results because the class of functions is rather limited. However, much slower convergence results are rather irrelevant in practice and linear convergence is not limited to convex quadratic functions: (1) as pointed out in this chapter, the invariance of ESs to strictly monotonic transformations implies the generalization of the result to the class of functions $\{g \circ f, g \in \mathcal{M}, f \text{ convex quadratic}\}$, that contains non-convex, non-smooth functions; (2) linear convergence with a positive probability (on a large class of functions) will imply linear convergence with probability one of a restart version of the algorithm, where a constant distribution is sampled simultaneously and the restart is conducted when a superior solution is sampled[e]; (3) robustness of the linear convergence in presence of noise has been proven when using a scale-invariant-constant step-size

[e]This idea was suggested to the authors first by Günter Rudolph.

(Jebalia *et al.*, 2009); (4) adaptation has been recognized as a key of the success of evolution strategies also in practice.

10.6. Appendix

10.6.1. *Proof of Theorem 10.17*

We prove now Theorem 10.17 that was stated page 311.

Proof. We fix n and assume that we are at the iteration n of a (1+1)-ES with adaptive step-size such that (X_n, σ_n) is known.

Maximal progress towards x^* in one step: The next iterate X_{n+1} either equals the sampled offspring $\mathrm{X}_n + \sigma_n N_n$ or the parent X_n (depending on what is the best according to f) and thus the distance between X_{n+1} and y^* is always larger or equal than the minimum between the distance between the offspring and y^* and the parent and y^*:

$$\|\mathrm{X}_{n+1} - y^*\| \geq \min\{\|\mathrm{X}_n - y^*\|, \|\mathrm{X}_n + \sigma_n N_n - y^*\|\} \ . \tag{10.39}$$

If $a > 0$, the minimum of (a, b) equals $a \min(1, b/a)$ such that

$$\|\mathrm{X}_{n+1} - y^*\| \geq \|\mathrm{X}_n - y^*\| \min\{1, \|\frac{\mathrm{X}_n - y^*}{\|\mathrm{X}_n - y^*\|} + \frac{\sigma_n}{\|\mathrm{X}_n - y^*\|} N_n\|\} \ . \tag{10.40}$$

Taking the logarithm of the previous equation we obtain

$$\begin{aligned} \ln \|\mathrm{X}_{n+1} - y^*\| \geq \ln \|\mathrm{X}_n - y^*\| + \\ \ln\left[\min\left\{1, \left\| \frac{\mathrm{X}_n - y^*}{\|\mathrm{X}_n - y^*\|} + \frac{\sigma_n}{\|\mathrm{X}_n - y^*\|} N_n \right\| \right\}\right] \ . \end{aligned} \tag{10.41}$$

We assume that $E[\ln \|\mathrm{X}_n - y^*\|] < +\infty$ for all n such that we can take the expectation in Equation (10.41) condition to (X_n, σ_n). We use the notation $\ln^-(x) = \max(0, -\ln(x))$ such that $\ln(\min(1, h(x))) = -\ln^-(h(x))$. By linearity of the expectation we obtain that

$$\begin{aligned} E[\ln \|\mathrm{X}_{n+1} - y^*\| | \mathrm{X}_n, \sigma_n] \geq \ln \|\mathrm{X}_n - y^*\| - \\ E\left[\ln^- \left\| \frac{\mathrm{X}_n - y^*}{\|\mathrm{X}_n - y^*\|} + \frac{\sigma_n}{\|\mathrm{X}_n - y^*\|} N_n \right\| \Big| \mathrm{X}_n, \sigma_n \right] \ . \end{aligned} \tag{10.42}$$

The offspring distribution $N(0, I_d)$ being spherical, i.e., the direction of $N(0, I_d)$ is uniformly distributed, it does not matter where the parent inducing the offspring is located on the unit hypersphere and thus

$$E\left[\ln^- \left\| \frac{\mathrm{X}_n - y^*}{\|\mathrm{X}_n - y^*\|} + \frac{\sigma_n}{\|\mathrm{X}_n - y^*\|} N_n \right\| \Big| \mathrm{X}_n, \sigma_n \right] = F_{(1+1)}\left(\frac{\sigma_n}{\|\mathrm{X}_n - y^*\|} \right), \tag{10.43}$$

where $F_{(1+1)}$ is defined in Lemma 10.15. Using the same lemma, we know that $F_{(1+1)}$ is continuous, the supremum $\tau := \sup F_{(1+1)}([0, +\infty])$ is reached and the step-size σ_F such that $F_{(1+1)}(\sigma_F) = \tau$ exists. Injecting this in Equation (10.42) we obtain $E[\ln \|\mathrm{X}_{n+1} - y^*\| | \mathrm{X}_n, \sigma_n] \geq \ln \|\mathrm{X}_n - y^*\| - \tau$ and consequently $E[\ln \|\mathrm{X}_{n+1} - y^*\|] \geq E[\ln \|\mathrm{X}_n - y^*\|] - \tau$. □

10.6.2. *Proof of Proposition 10.19 and Corollary 10.20*

We prove Proposition 10.19 and Corollary 10.20 stated page 312.

Proof. If $f(x) = g(\|x - x^*\|)$, Equation (10.42) with $y^* = x^*$ is an equality. If $\sigma_n = \sigma \|\mathrm{X}_n - x^*\|$, we obtain $E[\ln \|\mathrm{X}_{n+1} - x^*\|] = E[\ln \|\mathrm{X}_n - x^*\|] - F_{(1+1)}(\sigma)$ or $E[\ln \|\mathrm{X}_{n+1} - x^*\|] - E[\ln \|\mathrm{X}_n - x^*\|] = -F_{(1+1)}(\sigma)$. Summing from $n = n_0, \ldots, N$, we obtain that $E[\ln \|X_N - x^*\|] - E[\ln \|X_{n_0} - x^*\|] = -(N - n_0)F_{(1+1)}(\sigma)$. Dividing by N we obtain Equation (10.23). If $\sigma_n = \sigma_F \|\mathrm{X}_n - x^*\|$ where $F(\sigma_F) = \tau$, we obtain Equation (10.24). □

References

Auger, A. (2005). Convergence results for (1,λ)-SA-ES using the theory of φ-irreducible markov chains, *Theoretical Computer Science* **334**, pp. 35–69.

Auger, A. and Hansen, N. (2006). Reconsidering the progress rate theory for evolution strategies in finite dimensions, in A. Press (ed.), *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2006)*, pp. 445–452.

Baba, N. (1981). Convergence of random optimization methods for constrained optimization methods, *Journal of Optimization Theory and Applications*.

Beyer, H.-G. (2001). *The Theory of Evolution Strategies*, Natural Computing Series (Springer-Verlag).

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

Brooks, S. (1958). A discussion of random methods for seeking maxima, *The computer journal* **6**, 2.

Devroye, L. (1972). The compound random search, in *International Symposium on Systems Engineering and Analysis* (Purdue University), pp. 195–110.

Devroye, L. and Krzyżak, A. (2002). Random search under additive noise, in P. L. M. Dror and F. Szidarovsky (eds.), *Modeling Uncertainty* (Kluwer Academic Publishers), pp. 383–418.

Greenwood, G. W. and Zhu, Q. J. (2001). Convergence in evolutionary programs with self-adaptation, *Evolutionary Computation* **9**, 2, pp. 147–157.

Hansen, N. and Ostermeier, A. (2001). Completely derandomized self-adaptation in evolution strategies, *Evolutionary Computation* **9**, 2, pp. 159–195.

Jägersküpper, J. (2005). Rigorous runtime analysis of the (1+1)-ES: 1/5-rule and ellipsoidal fitness landscapes, in LNCS (ed.), *Foundations of Genetic Algorithms: 8th International Workshop, FoGA 2005*, Vol. 3469, pp. 260–281.

Jägersküpper, J. (2007). Analysis of a simple evolutionary algorithm for minimization in euclidean spaces, *Theoretical Computer Science* **379**, 3, pp. 329–347.

Jägersküpper, J. (2008). Lower bounds for randomized direct search with isotropic sampling, *Operations research letters* **36**, 3, pp. 327–332.

Jägersküpper, J. and Preuss, M. (2008). Aiming for a theoretically tractable CSA variant by means of empirical investigations, in *Proceedings of the 2008 Conference on Genetic and Evolutionary Computation*, pp. 503–510.

Jebalia, M., Auger, A. and Hansen, N. (2009). Log-linear convergence and divergence of the scale-invariant (1+1)-ES in noisy environments, *Algorithmica* In press.

Jebalia, M., Auger, A. and Liardet, P. (2008). Log-linear convergence and optimal bounds for the (1+1)-ES, in N. Monmarché and al. (eds.), *Proceedings of Evolution Artificielle (EA'07)*, *LNCS*, Vol. 4926 (Springer), pp. 207–218.

Karr, A. F. (1993). *Probability*, Springer Texts in Statistics (Springer-Verlag).

Kern, S., Müller, S., Hansen, N., Büche, D., Ocenasek, J. and Koumoutsakos, P. (2004). Learning Probability Distributions in Continuous Evolutionary Algorithms - A Comparative Review, *Natural Computing* **3**, 1, pp. 77–112.

Matyas, J. (1965). Random optimization, *Automation and Remote control* **26**, 2.

Meyn, S. and Tweedie, R. (1993). *Markov Chains and Stochastic Stability* (Springer-Verlag, New York).

Nekrutkin, V. and Tikhomirov, A. (1993). Speed of convergence as a function of given accuracy for random search methods, *Acta Applicandae Mathematicae* **33**, pp. 89–108.

Rechenberg, I. (1973). *Evolutionstrategie: Optimierung Technisher Systeme nach Prinzipien des Biologischen Evolution* (Frommann-Holzboog Verlag, Stuttgart).

Rudolph, G. U. (2001). Self-adaptive mutations may lead to premature convergence, *IEEE Transactions on Evolutionary Computation* **5**, pp. 410–414.

Schumer, M. and Steiglitz, K. (1968). Adaptive step size random search, *Automatic Control, IEEE Transactions on* **13**, pp. 270–276.

Schwefel, H.-P. (1981). *Numerical Optimization of Computer Models* (John Wiley & Sons, Inc., New York, NY, USA).

Stewart, G. W. (1995). On sublinear convergence, Tech. Rep. CS-TR-3534, University of Maryland.

Teytaud, O. and Gelly, S. (2006). General lower bounds for evolutionary algorithms, in 10^{th} *International Conference on Parallel Problem Solving from Nature (PPSN 2006)*, Vol. 4193 (Springer), pp. 21–31.

Tikhomirov, A. S. (2006). On the markov homogeneous optimization method, *Computational Mathematics and Mathematical Physics* **46**, 3, pp. 361–375.

White, R. (1971). A survey of random methods for parameter optimization, *SIMULATION* **17**, pp. 197–205.

Zhigljavsky, A. A. (1991). *Theory of Global Random Search* (Kluwer Academic Publishers).

Zhigljavsky, A. A. and Zilinskas, A. (2008). *Stochastic global optimization*, *Springer Optimization and its applications*, Vol. 1 (Springer).