Linear Convergence of Comparison-based Step-size Adaptive Randomized Search via Stability of Markov Chains

Anne Auger, Nikolaus Hansen

LINEAR CONVERGENCE OF COMPARISON-BASED STEP-SIZE ADAPTIVE RANDOMIZED SEARCH VIA STABILITY OF MARKOV CHAINS

ANNE AUGER* AND NIKOLAUS HANSEN*

Abstract. In this paper, we consider *comparison-based* adaptive stochastic algorithms for solving numerical optimisation problems. We consider a specific subclass of algorithms called comparison-based step-size adaptive randomized search (CB-SARS), where the state variables at a given iteration are a vector of the search space and a positive parameter, the step-size, typically controlling the overall standard deviation of the underlying search distribution.

We investigate the linear convergence of CB-SARS on *scaling-invariant* objective functions. Scaling-invariant functions preserve the ordering of points with respect to their function value when the points are scaled with the same positive parameter (the scaling is done w.r.t. a fixed reference point). This class of functions includes norms composed with strictly increasing functions as well as *non quasi-convex* and *non-continuous* functions. On scaling-invariant functions, we show the existence of a homogeneous Markov chain, as a consequence of natural invariance properties of CB-SARS (essentially scale-invariance and invariance to strictly increasing transformation of the objective function). We then derive sufficient conditions for asymptotic *global linear convergence* of CB-SARS, expressed in terms of different stability conditions of the normalised homogeneous Markov chain (irreducibility, positivity, Harris recurrence, geometric ergodicity) and thus define a general methodology for proving global linear convergence of CB-SARS algorithms on scaling-invariant functions.

Key words. stochastic algorithms, numerical optimisation, Markov chains, Monte Carlo Markov Chains, comparison-based, linear convergence, invariance, adaptive randomized search, adaptive algorithms

1. Introduction. We consider the problem of minimizing an objective function $f : \mathbb{R}^n \mapsto \mathbb{R}$ where the search cost is defined as the number of calls to the function f. We investigate *comparison-based* search algorithms that use the f-values only through *comparisons*. Because the f-values are totally ordered, from pair-wise comparisons a ranking of f-values can be derived and we can equivalently refer to our scenario as *comparison-* or *ranking-based*. In allusion to the term *derivative-free optimization*, we might speak of *function-value-free optimization* in this case. Well-known derivative-free methods are comparison-based algorithms, for instance the pattern search method by Hooke and Jeeves [13] and the simplex method by Nelder and Mead [18] and we believe that their success is to some extend due to their comparison-based property.

From the fact that the methods only use the comparison information follows invariance of the algorithms to composing the objective function (to the left) by a strictly increasing function $g : \mathbb{R} \to \mathbb{R}$. This invariance property provides robustness because an error on the objective function value–that can stem from various sources of noise–has an impact only if it changes the result of a comparison, i.e., if it changes the *f-ordering* of the candidate solutions under consideration. This invariance provides robustness also in that very small or very large f-values can only have a limited impact. The invariance also facilitates predictability, because the sequence of solutions generated on f and on $g \circ f$ are indistinguishable. At the same time it arguably makes convergence proofs harder to tackle, as one has a weaker control on the objective function decrease.

In this context, this paper investigates the linear convergence of a class of *adaptive stochastic* comparison-based algorithms, namely comparison-based *step-size adaptive* randomised search, CB-SARS. Formally, a SARS is a stochastically recursive sequence

*Inria, LRI, Bât 660, University Paris Sud, 91405, Orsay, France (`first.lastname_at_inria.fr`).

on the state space $\Omega = \mathbb{R}^n \times \mathbb{R}^+_>$. Given $(\mathbf{X}_0, \sigma_0) \in \mathbb{R}^n \times \mathbb{R}^+_>$, the sequence is iteratively defined as

$$(\mathbf{X}_{t+1}, \sigma_{t+1}) = \mathcal{F}((\mathbf{X}_t, \sigma_t), \mathbf{U}_t) \tag{1.1}$$

where $\mathbf{X}_t \in \mathbb{R}^n$ represents the favorite or incumbent solution at iteration t, $\sigma_t \in \mathbb{R}^+_>$ is the so-called step-size, $\mathcal{F}$ is a measurable function and $(\mathbf{U}_t)_{t\in\mathbb{N}}$ is an independent identically distributed (i.i.d.) sequence of random vectors. Often, the step-size σ_t represents the overall standard-deviation of an underlying sampling distribution. The objective function f must be available to the *transition function* $\mathcal{F}$. While for SARS, the transition function can use the *f-values* of candidate solutions, the transition function of CB-SARS uses only f-comparisons. A formal definition will be given in Definition 2.3.

The definition via (1.1) is general and abstract, however, often, SARS and CB-SARS take a specific form where the connexion with gradient methods becomes clear while the methods are *derivative* and even *function-value* free. Indeed, the update of the incumbent solution generally writes

$$\mathbf{X}_{t+1} = \mathbf{X}_t + \kappa\sigma_t \mathbf{Y}_t \tag{1.2}$$

where $\mathbf{Y}_t$ is a combination of selected random directions that can be seen as an approximation of a gradient direction and κ is a learning rate. This connexion can be pushed further for some specific algorithms where $\theta_t = (\mathbf{X}_t, \sigma_t)$ encodes the mean vector and standard deviation of a Gaussian distribution and a joint optimization criterion formulated on the manifold defined by the family of Gaussian distributions P_θ can be defined. Applying a gradient update step with respect to θ to this joint criterion and taking a Monte Carlo approximation of the gradient[1] defines a comparison-based step-size adaptive randomized search whose update equations are given in (2.22) and (2.25) [1, 20]. Note that the learning rate κ (and κ_m, κ_σ in (2.22) and (2.25)) correspond to the step-size of the underlying *stochastic approximation algorithm* (here we however reserve the step-size name for σ_t unless explicitly specified).

Several random optimization methods akin to the update in (1.2) were recently studied. First Nesterov proved complexity bounds for some gradient-free algorithms using oracles for directional derivatives (that use Gaussian random directions) [19]. Later on, Stich et al. analyzed the simple Random Pursuit (RP) where $\mathbf{Y}_t$ is a random direction and σ_t is the result of a line-search in the $\mathbf{Y}_t$ direction. Assuming exact or approximate line search, they prove the linear convergence of RP for strongly convex functions. They experimentally compared RP to an accelerated version of RP, to Nesterov's schemes and to a classical CB-SARS [23, 21, 9]. The accelerated RP and Nesterov's schemes need as input some constant related to the function (i.e., they are not tested in a black-box setting). Concluding their observations on the performance of the CB-SARS, the authors emphasize "*that the performance of the adaptive step-size ES scheme [the classical CB-SARS] is remarkable given the fact that it does not need any function-specific parametrization. A comparison to the RP shows that it needs four times fewer function evaluations on functions $f_2 - f_4$.*" [25]. The main reason are the saved expenses due to the omitted line searchs. The theoretical analysis in [25] heavily relies on the control of the f-decrease at each iteration with the assumption of exact line search (or with a controlled error in the case of approximate line search).

[1] The gradient is taken wrt the Fisher Information metric, it is also called natural gradient.

We believe that the author's analysis however cannot be generalized to many CB-SARS. We resort thus to a different approach that can prove in particular the linear convergence of the CB-SARS algorithm experimented in the aforementioned paper (for which the authors stress the remarkable performance) [4].

While the previously mentioned papers analyze the algorithms on strongly convex and convex functions, we consider here the class of *scaling-invariant functions*, natural in the context of *comparison-based* algorithms. We call a function f scaling-invariant with respect to $\mathbf{x}^\star$ if for all $\rho > 0$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$f(\rho(\mathbf{x} - \mathbf{x}^\star)) \leq f(\rho(\mathbf{y} - \mathbf{x}^\star)) \Leftrightarrow f(\mathbf{x} - \mathbf{x}^\star) \leq f(\mathbf{y} - \mathbf{x}^\star) \ .$$

This class includes all norms and all functions that are the composition of norm functions by increasing transformations–having hence convex sublevel sets–but also non quasi-convex functions, i.e., functions with non-convex sublevel sets. Non-constant scaling-invariant functions do not admit strict local optima nor plateaus and are thus essentially unimodal.

We prove that if a CB-SARS is scale-invariant (informally has no intrinsic notion of scale), then, on a scaling-invariant function (where without loss of generality $\mathbf{x}^\star = 0$), the normalised process $\mathbf{X}_t/\sigma_t$ is an homogeneous Markov chain. In addition, stability properties of this Markov chain imply asymptotic linear convergence of the original algorithm independently of the starting point. We then formulate different *sufficient conditions* expressed as stability conditions on $\mathbf{X}_t/\sigma_t$, inducing *global* linear convergence almost surely and in expectation. We also formulate conditions for proving that the empirical estimate of the convergence rate converges geometrically to the theoretical one. We hence define a general methodology to prove linear convergence of CB-SARS algorithms. Our methodology generalizes previous works, restricted to a specific CB-SARS on the sphere function [6, 3], to a broader class of algorithms and a much broader class of functions. In a companion manuscript, the methodology has been applied to another comparison-based step-size adaptive randomized search algorithm [4].

The rest of this paper is organized as follows. We define in Section 2.1 CB-SARS algorithms. In Section 2.2, we formalize different invariance properties commonly associated to CB-SARS. In Section 2.3 we present several examples of existing methods that follow our general definition of CB-SARS and study their invariance properties. In Section 3 we define the class of scaling-invariant functions. In Section 4, we prove that for certain translation and scale-invariant CB-SARS algorithms optimizing scaling-invariant functions, $\mathbf{X}_t/\sigma_t$ is a homogeneous Markov chain. In Section 5, we give sufficient conditions to linear convergence expressed in terms of stability of the Markov chain exhibited in Section 4. A discussion is finally provided in Section 6.

Notations and definitions. The set of nonnegative real numbers is denoted $\mathbb{R}^+$ and $\mathbb{R}^+_>$ denotes elements of $\mathbb{R}^+$ excluding 0, $\mathbb{N}$ is the set of natural numbers including zero while $\mathbb{N}_>$ excludes zero. The Euclidian norm of a vector $\mathbf{x}$ of $\mathbb{R}^n$ is denoted $\|\mathbf{x}\|$. A Gaussian vector or multivariate normal distribution with mean vector $\mathbf{m}$ and covariance matrix $\mathbf{C}$ is denoted $\mathcal{N}(\mathbf{m}, \mathbf{C})$. The identity matrix in $\mathbb{R}^{n\times n}$ is denoted $\mathbf{I}_n$ such that a standard multivariate normal distribution, i.e. with mean vector zero and identity covariance matrix is denoted $\mathcal{N}(0, \mathbf{I}_n)$. The density of a standard multivariate normal distribution (in any dimension) is denoted $p_\mathcal{N}$. The set of strictly increasing functions g from $\mathbb{R}$ to $\mathbb{R}$ or from a subset $I \subset \mathbb{R}$ to $\mathbb{R}$ is denoted $\mathcal{M}$.

2. Comparison Based Step-size Adaptive Randomized Search (CB-SARS). In this section, we present a formal definition of CB-SARS algorithms. We

then define invariance properties generally associated to those algorithms and finish by giving several concrete examples of CB-SARS algorithms as well as analyzing their invariance properties.

2.1. Algorithm Definitions. We consider a SARS as defined in (1.1) and consider that each vector $\mathbf{U}_t$ belongs to a space $\mathbb{U}^p = \mathbb{U} \times \ldots \times \mathbb{U}$ and has p coordinates $\mathbf{U}_t^i$ belonging to $\mathbb{U}$. The probability distribution of the vector $\mathbf{U}_t$ is denoted $p_{\mathbf{U}}$. From the definition (1.1) follows that $((\mathbf{X}_t, \sigma_t))_{t\in\mathbb{N}}$ is a time homogeneous Markov chain. We call $\mathcal{F}$ the *transition function* of the algorithm. The objective function f is also an input argument to the transition function $\mathcal{F}$ as the update of $(\mathbf{X}_t, \sigma_t)$ depends on f, however we omit this dependence in general for the sake of simplicity in the notations. If there is an ambiguity we add the function f as upper-script, i.e. $\mathcal{F}^{f(\mathbf{x})}$ or $\mathcal{F}^f$.

A CB-SARS is a SARS where the transition function $\mathcal{F}$ depends on f only through comparison of candidate solutions and is the composition of several functions that we specify in the sequel. The p coordinates of $\mathbf{U}_t$ are in a first time used to create new candidate solutions according to a $\mathcal{S}ol$ function:

$$\mathbf{X}_t^i = \mathcal{S}ol((\mathbf{X}_t, \sigma_t), \mathbf{U}_t^i), i = 1, \ldots, p \ .$$

(For instance in the case where $\mathbb{U} = \mathbb{R}^n$ the $\mathcal{S}ol$ function can equal $\mathcal{S}ol((\mathbf{x}, \sigma), \mathbf{u}^i) = \mathbf{x} + \sigma\mathbf{u}^i$.) The p candidate solutions are then evaluated on f and ordered according to their objective function value. The permutation corresponding to the ordered objective function values $f(\mathbf{X}_t^i)$ is denoted $\mathcal{S} \in \mathfrak{S}(p)$ where we denote $\mathfrak{S}(p)$ the set of permutations of p elements. Formally $\mathcal{S}$ is the output of the $\mathcal{O}rd$ function defined below. It is then used to order the coordinates of the vector $\mathbf{U}_t$ accordingly. More formally the permutation acts on the coordinates of $\mathbf{U}_t$ via the following function:

$$\begin{aligned} \mathfrak{S}(p) \times \mathbb{U}^p &\to \mathbb{U}^p \\ (\mathcal{S}, \mathbf{U}_t) &\mapsto \mathcal{S} * \mathbf{U}_t = \left(\mathbf{U}_t^{\mathcal{S}(1)}, \ldots, \mathbf{U}_t^{\mathcal{S}(p)}\right) \end{aligned} \tag{2.1}$$

where the previous equation implicitly defines the operator $*$.

The update of $(\mathbf{X}_t, \sigma_t)$ is achieved using the current state $(\mathbf{X}_t, \sigma_t)$ and the ranked coordinates of $\mathbf{U}_t$. More precisely let us consider a measurable function $\mathcal{G}$ called update function that maps $\Omega \times \mathbb{U}^p$ onto Ω, the update of $(\mathbf{X}_t, \sigma_t)$ reads

$$(\mathbf{X}_{t+1}, \sigma_{t+1}) = \mathcal{G}((\mathbf{X}_t, \sigma_t), \mathcal{S} * \mathbf{U}_t) = \mathcal{G}((\mathbf{X}_t, \sigma_t), \mathbf{Y}_t) \ , \tag{2.2}$$

where $\mathbf{Y}_t$ denotes the ordered coordinates of $\mathbf{U}_t$, i.e. $\mathbf{Y}_t = (\mathbf{U}_t^{\mathcal{S}(1)}, \ldots, \mathbf{U}_t^{\mathcal{S}(p)})$. We formalize the definition of a CB-SARS below after introducing a definition for the function $\mathcal{S}ol$ for generating solutions as well as for the ordering function.

Definition 2.1 ($\mathcal{S}ol$ function). *Given $\mathbb{U}$, the state space for the sampling coordinates of $\mathbf{U}_t$, a $\mathcal{S}ol$ function used to create candidate solutions is a measurable function mapping $\Omega \times \mathbb{U}$ into $\mathbb{R}^n$, i.e.*

$$\mathcal{S}ol : \Omega \times \mathbb{U} \mapsto \mathbb{R}^n \ .$$

We now define the ordering function that returns a permutation based on the objective function values.

Definition 2.2 ($\mathcal{O}rd$ function). *The ordering function $\mathcal{O}rd$ maps $\mathbb{R}^p$ to $\mathfrak{S}(p)$, the set of permutations with p elements and returns for any set of real values $(f_1, \ldots, f_p)$ the permutation of ordered indexes. That is $\mathcal{S} = \mathcal{O}rd(f_1, \ldots, f_p) \in \mathfrak{S}(p)$ where*

$$f_{\mathcal{S}(1)} \leq \ldots \leq f_{\mathcal{S}(p)} \ .$$

When more convenient we might denote $\mathcal{O}rd((f_i)_{i=1,\dots,p})$ *instead of* $\mathcal{O}rd(f_1,\dots,f_p)$. *When needed for the sake of clarity we might use the notations* $\mathcal{O}rd^f$ *or* $\mathcal{S}^f$ *to emphasize the dependency in* f.

We are now ready to give a formal definition of a comparison-based step-size adaptive randomized search.

DEFINITION 2.3 (CB-SARS minimizing $f : \mathbb{R}^n \to \mathbb{R}$). *Let* $p \in \mathbb{N}_>$ *and* $\mathbb{U}^p = \mathbb{U} \times \ldots \times \mathbb{U}$ *where* $\mathbb{U}$ *is a subset of* $\mathbb{R}^m$. *Let* $p_{\mathbf{U}}$ *be a probability distribution defined on* $\mathbb{U}^p$ *where each* $\mathbf{U}$ *distributed according to* $p_{\mathbf{U}}$ *has a representation* $(\mathbf{U}^1,\dots,\mathbf{U}^p)$ *(each* $\mathbf{U}^i \in \mathbb{U}$*). Let* $\mathcal{S}ol$ *be a solution function as in Definition 2.1. Let* $\mathcal{G}_1 : \Omega \times \mathbb{U}^p \mapsto \mathbb{R}^n$ *and* $\mathcal{G}_2 : \mathbb{R}^+ \times \mathbb{U}^p \mapsto \mathbb{R}^+$ *be two mesurable mappings and let denote* $\mathcal{G} = (\mathcal{G}_1, \mathcal{G}_2)$.

A CB-SARS is determined by the quadruplet $(\mathcal{S}ol, \mathcal{G}, \mathbb{U}^p, p_{\mathbf{U}})$ *from which the recursive sequence* $(\mathbf{X}_t, \sigma_t) \in \Omega$ *is defined via* $(\mathbf{X}_0, \sigma_0) \in \Omega$ *and for all* t*:*

$$\mathbf{X}_t^i = \mathcal{S}ol((\mathbf{X}_t, \sigma_t), \mathbf{U}_t^i), i = 1,\dots,p \tag{2.3}$$

$$\mathcal{S} = \mathcal{O}rd(f(\mathbf{X}_t^1),\dots,f(\mathbf{X}_t^p)) \in \mathfrak{S}(p) \tag{2.4}$$

$$\mathbf{X}_{t+1} = \mathcal{G}_1\left((\mathbf{X}_t, \sigma_t), \mathcal{S} * \mathbf{U}_t\right) \tag{2.5}$$

$$\sigma_{t+1} = \mathcal{G}_2\left(\sigma_t, \mathcal{S} * \mathbf{U}_t\right) \tag{2.6}$$

where $(\mathbf{U}_t)_{t\in\mathbb{N}}$ *is an i.i.d. sequence of random vectors on* $\mathbb{U}^p$ *distributed according to* $p_{\mathbf{U}}$, $\mathcal{O}rd$ *is the ordering function as in Definition 2.2.*

The previous definition illustrates the *function-value-free* property as we see that the update of the state $(\mathbf{X}_t, \sigma_t)$ is performed using solely the information given by the permutation that contains the order of the candidate solutions according to f. For a comparison-based step-size adaptive randomized search, the function $\mathcal{F}$ introduced in (1.1) is the composition of the update function $\mathcal{G}$, the solution operator $\mathcal{S}ol$ and the ordering function, more precisely

$$\boxed{\mathcal{F}^f((\mathbf{x},\sigma),\mathbf{u}) = \mathcal{G}((\mathbf{x},\sigma), \mathcal{O}rd(f(\mathcal{S}ol((\mathbf{x},\sigma),\mathbf{u}^i))_{i=1,\dots,p}) * \mathbf{u})} \ . \tag{2.7}$$

Note that for the update of the step-size (equation (2.6)), we assume that $\mathbf{X}_t$ does not come into play. Examples of CB-SARS are given in Section 2.3.

2.2. Invariance Properties. Invariance is an important principle in science in general and in optimization. Invariance generalizes properties that are true on a single function to a whole class of functions. Some invariances are taken for granted in optimization like translation invariance while others are less common or less recognized. In the sequel we start by formalizing that CB-SARS are invariant to strictly monotonic transformations of f. We focus then in Section 2.2.2 on invariance in search space and formalize translation invariance and scale invariance. We also derive sufficient conditions for a CB-SARS to be translation and scale invariant.

2.2.1. Invariance to Strictly Monotonic Transformations of f. Invariance to strictly monotonic transformations of f of a CB-SARS algorithm is a direct consequence of the algorithm definition. It stems from the fact that the objective function only comes into play through the ranking of the solutions via the ordering function (see (2.4), (2.5) and (2.6)). This ordering function does output the same result on f or any strictly monotonic transformation of f. More formally let us define $\mathcal{M}_I$ the set of strictly monotonic mappings $g : I \to \mathbb{R}$, where I is a subset of $\mathbb{R}$ i.e. if for all x and y in I such that $x < y$ we have $g(x) < g(y)$ and define $\mathcal{M} = \cup_I \mathcal{M}_I$. The

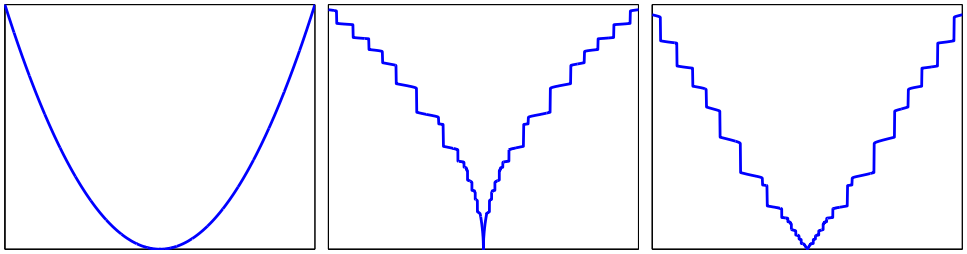

FIG. 2.1. *Illustration of invariance to strictly increasing transformations. Representation of three instances of functions belonging to the invariance (w.r.t. strictly increasing transformations) class of $f(\mathbf{x}) = \|\mathbf{x}\|^2$ in dimension 1. On the left the sphere function and middle and right functions $g \circ f$ for two different $g \in \mathcal{M}$. On those three functions, a comparison-based step-size adaptive randomized search will generate the same sequence $(\mathbf{X}_t, \sigma_t)$ (see Proposition 2.4) and consequently convergence will take place at the same rate.*

elements of $\mathcal{M}$ preserve the ordering. The invariance to composite of f by a function in $\mathcal{M}$ is stated in the following proposition.

PROPOSITION 2.4. *[Invariance to strictly monotonic transformations] Consider $(\mathcal{S}ol, \mathcal{G}, \mathbb{U}^p, p_{\mathbf{U}})$ a CB-SARS as defined in Definition 2.3 optimizing $f : \mathbb{R}^n \to \mathbb{R}$ and let $(\mathbf{X}_t, \sigma_t)$ be the Markov chain sequence defined as $(\mathbf{X}_0, \sigma_0) \in \mathbb{R}^n \times \mathbb{R}_>^+$ and*

$$(\mathbf{X}_{t+1}, \sigma_{t+1}) = \mathcal{G}((\mathbf{X}_t, \sigma_t), \mathcal{S}^f * \mathbf{U}_t)$$

where $(\mathbf{U}_t)_{t\in\mathbb{N}}$ is an i.i.d. sequence of random vectors on $\mathbb{U}^p$ distributed according to $p_{\mathbf{U}}$ and $\mathcal{S}^f = \mathcal{O}rd(f(\mathcal{S}ol((\mathbf{X}_t, \sigma_t), \mathbf{U}_t^i))_{1\leq i\leq p})$. Let $g : f(\mathbb{R}^n) \to \mathbb{R}$ (where $f(\mathbb{R}^n)$ is the image of f) in $\mathcal{M}$ be a strictly monotonic mapping and $(\mathbf{X}'_t, \sigma'_t)$ be the Markov chain obtained when optimizing $g \circ f$ using the same sequence $(\mathbf{U}_t)_{t\in\mathbb{N}}$ and same initial state $(\mathbf{X}_0{}', \sigma_0{}') = (\mathbf{X}_0, \sigma_0)$. Then almost surely for all t

$$(\mathbf{X}_t, \sigma_t) = (\mathbf{X}'_t, \sigma'_t) \ .$$

Proof. The proof is immediate, by induction. Assume $(\mathbf{X}_t, \sigma_t) = (\mathbf{X}'_t, \sigma'_t)$ and let denote $\mathbf{X}_t^i = \mathcal{S}ol((\mathbf{X}_t, \sigma_t), \mathbf{U}_t^i)$. Because $\mathcal{O}rd(f(\mathbf{X}_t^1), \ldots, f(\mathbf{X}_t^p)) = \mathcal{O}rd(g\circ f(\mathbf{X}_t^1), \ldots, g\circ f(\mathbf{X}_t^p)) = \mathcal{S}$, then

$$(\mathbf{X}_{t+1}, \sigma_{t+1}) = \mathcal{G}((\mathbf{X}_t, \sigma_t), \mathcal{S} * \mathbf{U}_t) = \mathcal{G}((\mathbf{X}'_t, \sigma'_t), \mathcal{S} * \mathbf{U}_t) = (\mathbf{X}'_{t+1}, \sigma'_{t+1}) \ . \qquad \square$$

Consequently, on the three functions depicted in Figure 2.1, a comparison-based step-size adaptive randomized search will produce the same sequence $(\mathbf{X}_t, \sigma_t)_{t\in\mathbb{N}}$. Hence if convergence takes place on one of those functions, it will take place on the two others and at the same convergence rate.

2.2.2. Invariances in the Search Space: Translation and Scale-Invariance. We consider now invariance of comparison-based step-size adaptive randomized search related to transformations in the search space. We use a classical approach to formalize invariance using homomorphisms transforming state variables via a group action and visualize invariances with a commutative diagram. We start by translation invariance, usually taken for granted in optimization.

Translation invariance. Most optimization algorithms are translation invariant, which implies the same performance when optimizing $\mathbf{x} \mapsto f(\mathbf{x})$ or $\mathbf{x} \mapsto f(\mathbf{x} - \mathbf{x}_0)$

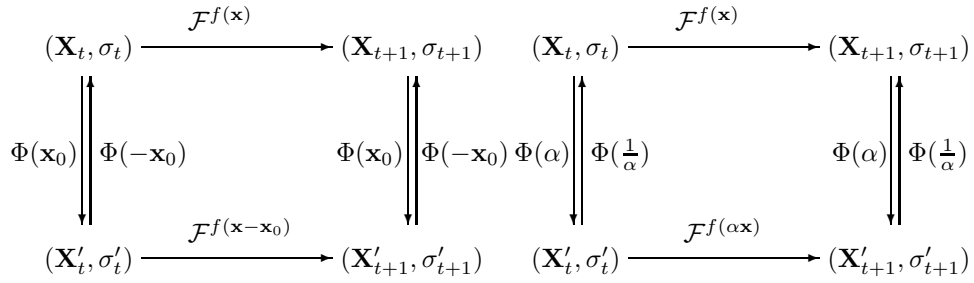

FIG. 2.2. *Left: Commutative diagram for the translation invariance property applied to one iteration of a step-size adaptive algorithm* ($\Phi(-\mathbf{x}_0) = [\Phi(\mathbf{x}_0)]^{-1}$). *Right: Commutative diagram for the scale-invariance property applied to one iteration of a step-size adaptive algorithm* ($\Phi(1/\alpha) = [\Phi(\alpha)]^{-1}$). *The homomorphisms* Φ *(different on the left and right) define for any* $\mathbf{x}_0$ *(resp.* α*) a search space transformation* $\Phi(\mathbf{x}_0)$ *(resp.* $\Phi(\alpha)$*).*

for all $\mathbf{x}_0$ provided a respective initialization of the algorithm. More precisely, let us consider $\mathbb{R}^n$ endowed with the addition operation $+$ as a group and consider $\mathcal{A}(\Omega)$ the set of invertible mappings from the state space Ω to itself, that is, the set of all (invertible) state space transformations. Endowed with the function composition $\circ$, $(\mathcal{A}(\Omega), \circ)$ yields also a group structure. We remind the definition of a group morphism.

DEFINITION 2.5 (Group homomorphism). *Let* $(G_1, .)$ *and* $(G_2, *)$ *be two groups. A mapping* $\Phi : G_1 \to G_2$ *is called group homomorphism if for all* $x, y \in G_1$ *we have* $\Phi(x.y) = \Phi(x) * \Phi(y)$.

From the definition follows that for any $x \in G_1$, $\Phi(x^{-1}) = [\Phi(x)]^{-1}$ where x^{-1} (resp.$[\Phi(x)]^{-1}$) denotes the inverse of x (resp. of $[\Phi(x)]$). Note that in case x belongs to an additive group, the inverse is denoted $-x$. Let $\mathrm{Homo}((\mathbb{R}^n, +), (\mathcal{A}(\Omega), \circ))$ be the set of group homomorphisms from $(\mathbb{R}^n, +)$ to $(\mathcal{A}(\Omega), \circ)$. For instance, consider $\Phi \in \mathrm{Homo}((\mathbb{R}^n, +), (\mathcal{A}(\Omega), \circ))$, i.e. $\Phi : \mathbf{y} \in (\mathbb{R}^n, +) \mapsto \Phi(\mathbf{y})$ where $\Phi(\mathbf{y})$ is a state space transformation such that for all $(\mathbf{x}, \sigma) \in \mathbb{R}^n \times \mathbb{R}^+_>$, $\Phi(\mathbf{y})((\mathbf{x}, \sigma)) = (\mathbf{x} + \mathbf{y}, \sigma)$. We are now ready to state a definition of translation invariance.

DEFINITION 2.6 (Translation Invariance). *A SARS with transition function* $\mathcal{F}$ *is translation invariant if there exists a group homomorphism* $\Phi \in \mathrm{Homo}((\mathbb{R}^n, +), (\mathcal{A}(\Omega), \circ))$ *such that for any objective function* f*, for any* $\mathbf{x}_0 \in \mathbb{R}^n$*, for any* $(\mathbf{x}, \sigma) \in \Omega$ *and for any* $\mathbf{u} \in \mathbb{U}^p$

$$\mathcal{F}^{f(\mathbf{x})}((\mathbf{x}, \sigma), \mathbf{u}) = \underbrace{\Phi(\mathbf{x}_0)^{-1}}_{\Phi(-\mathbf{x}_0)} \left(\mathcal{F}^{f(\mathbf{x}-\mathbf{x}_0)}(\Phi(\mathbf{x}_0)(\mathbf{x}, \sigma), \mathbf{u}) \right) \quad , \tag{2.8}$$

where the function to be optimized is shown as upper-script of the transition function $\mathcal{F}$. The previous definition means that a SARS algorithm is translation invariant, if we can find an homomorphism Φ (that depends on the algorithm) that defines for any translation $\mathbf{x}_0$, a search space transformation $\Phi(\mathbf{x}_0)$, such that: (i) if we start from $(\mathbf{X}_t, \sigma_t)$ and apply one iteration of the algorithm to optimize $f(\mathbf{x})$ or (ii) apply the state space transformation $\Phi(\mathbf{x}_0)$ to the state of the algorithm, apply one iteration of the algorithm on $\mathbf{x} \mapsto f(\mathbf{x} - \mathbf{x}_0)$ and transform back the state of the algorithm via $\Phi(-\mathbf{x}_0)$, then we recover $(\mathbf{X}_{t+1}, \sigma_{t+1})$. This property is pictured via a double-commutative diagram (see Figure 2.2).

We consider in the next proposition some specific properties of $\mathcal{S}ol$ and $\mathcal{G}$ that render a comparison-based step-size adaptive randomized search translation invariant. These properties are satisfied for algorithms presented in Section 2.3.

PROPOSITION 2.7. *Let $(\mathcal{S}ol, \mathcal{G}, \mathbb{U}^p, p_{\mathbf{U}})$ be a CB-SARS according to Definition 2.3. If the following conditions are satisfied:*
(i) for all $\mathbf{x}, \mathbf{x}_0 \in \mathbb{R}^n$ for all $\sigma > 0$, for all $\mathbf{u}^i \in \mathbb{U}$

$$\mathcal{S}ol((\mathbf{x}+\mathbf{x}_0, \sigma), \mathbf{u}^i) = \mathcal{S}ol((\mathbf{x}, \sigma), \mathbf{u}^i) + \mathbf{x}_0 \tag{2.9}$$

(ii) for all $\mathbf{x}, \mathbf{x}_0 \in \mathbb{R}^n$ for all $\sigma > 0$, for all $\mathbf{y} \in \mathbb{U}^p$

$$\mathcal{G}_1((\mathbf{x}+\mathbf{x}_0, \sigma), \mathbf{y}) = \mathcal{G}_1((\mathbf{x}, \sigma), \mathbf{y}) + \mathbf{x}_0 \tag{2.10}$$

then $(\mathcal{S}ol, \mathcal{G}, \mathbb{U}^p, p_{\mathbf{U}})$ is translation invariant and the associated group homomorphism Φ is defined by

$$\Phi(\mathbf{x}_0)(\mathbf{x}, \sigma) = (\mathbf{x}+\mathbf{x}_0, \sigma) \text{ for all } \mathbf{x}_0, \mathbf{x}, \sigma \ . \tag{2.11}$$

In addition, assuming that the $\mathcal{S}ol$ function satisfies property (2.9), then if $(\mathcal{S}ol, \mathcal{G}, \mathbb{U}^p, p_{\mathbf{U}})$ is translation invariant with (2.11) as homomorphism, then (2.10) is satisfied.

Proof. Consider the homomorphism defined in (2.11), then (2.10) writes

$$\mathcal{G}(\Phi(\mathbf{x}_0)(\mathbf{x}, \sigma), \mathbf{y}) = \Phi(\mathbf{x}_0)\left(\mathcal{G}((\mathbf{x}, \sigma), \mathbf{y})\right) \ , \tag{2.12}$$

and (2.9) writes $\mathcal{S}ol(\Phi(\mathbf{x}_0)(\mathbf{x}, \sigma), \mathbf{u}^i) - \mathbf{x}_0 = \mathcal{S}ol((\mathbf{x}, \sigma), \mathbf{u}^i)$. This latter equation implies that the same permutation $\mathcal{S}$ will result from ordering solutions generated by the $\mathcal{S}ol$ function on f from $(\mathbf{x}, \sigma)$ or on $f(\mathbf{x}-\mathbf{x}_0)$ starting from $\Phi(\mathbf{x}_0)(\mathbf{x}, \sigma)$. Using (2.12) we hence have $\mathcal{G}(\Phi(\mathbf{x}_0)(\mathbf{x}, \sigma), \mathcal{S}^{f(\mathbf{x}-\mathbf{x}_0)}_{\Phi(\mathbf{x}_0)(\mathbf{x},\sigma)} * \mathbf{u}) = \Phi(\mathbf{x}_0)\left(\mathcal{G}((\mathbf{x}, \sigma), \mathcal{S}^f_{(\mathbf{x},\sigma)} * \mathbf{u})\right)$ which turns out to coincide with (2.8). The inverse is immediate. □

Scale-invariance property. The scale invariance property translates the fact that the algorithm has no intrinsic notion of scale. It can be defined similarly to translation invariance by considering the set of group homomorphisms from the group $(\mathbb{R}^+_>, .)$ (where . denotes the multiplication between two real numbers) to the group $(\mathcal{A}(\Omega), \circ)$. We denote this set $\text{Homo}((\mathbb{R}^+_>, .), (\mathcal{A}(\Omega), \circ))$.

DEFINITION 2.8 (Scale-invariance). *A SARS with transition function $\mathcal{F}$ is* scale-invariant *if there exists an homomorphism $\Phi \in \text{Homo}((\mathbb{R}^+_>, .), (\mathcal{A}(\Omega), \circ))$ such that for any f, for any $\alpha > 0$, for any $(\mathbf{x}, \sigma) \in \mathbb{R}^n \times \mathbb{R}^+_>$ and for any $\mathbf{u} \in \mathbb{U}^p$*

$$\mathcal{F}^{f(\mathbf{x})}((\mathbf{x}, \sigma), \mathbf{u}) = \Phi(1/\alpha)\left(\mathcal{F}^{f(\alpha\mathbf{x})}(\Phi(\alpha)(\mathbf{x}, \sigma), \mathbf{u})\right) \ , \tag{2.13}$$

where the function optimized is shown as upper-script of the transition function $\mathcal{F}$. In the previous definition we have used the fact that for any element α of the multiplicative group $(\mathbb{R}^+_>, .)$ its inverse is $1/\alpha$. The scale-invariance property can be pictured via a double-commutative diagram (see Figure 2.2).

We derive in the next proposition some conditions for a CB-SARS to be scale-invariant that will be useful in the sequel to prove that the algorithms presented in Section 2.3 are scale-invariant.

PROPOSITION 2.9. *Let $(\mathcal{S}ol, \mathcal{G}, \mathbb{U}^p, p_{\mathbf{U}})$ be a CB-SARS according to Definition 2.3. If for all $\alpha > 0$ the following three conditions are satisfied: (i) for all $\mathbf{u}^i \in \mathbb{U}, (\mathbf{x}, \sigma) \in \mathbb{R}^n \times \mathbb{R}^+_>$,*

$$\mathcal{S}ol((\mathbf{x}, \sigma), \mathbf{u}^i) = \alpha \mathcal{S}ol\left(\left(\frac{\mathbf{x}}{\alpha}, \frac{\sigma}{\alpha}\right), \mathbf{u}^i\right) \tag{2.14}$$

(ii) for all $\mathbf{y} \in \mathbb{U}^p, (\mathbf{x}, \sigma) \in \mathbb{R}^n \times \mathbb{R}^+_>$

$$\mathcal{G}_1((\mathbf{x}, \sigma), \mathbf{y}) = \alpha \mathcal{G}_1\left(\left(\frac{\mathbf{x}}{\alpha}, \frac{\sigma}{\alpha}\right), \mathbf{y}\right) \tag{2.15}$$

and (iii) for all $\mathbf{y} \in \mathbb{U}^p, \sigma \in \mathbb{R}^+_>$

$$\mathcal{G}_2(\sigma, \mathbf{y}) = \alpha \mathcal{G}_2 \left(\frac{\sigma}{\alpha}, \mathbf{y} \right) \quad , \tag{2.16}$$

then it is scale invariant and the associated homomorphism is $\Phi : \alpha \in \mathbb{R}^+_> \mapsto \Phi(\alpha)$ *where for all* $(\mathbf{x}, \sigma) \in \mathbb{R}^n \times \mathbb{R}^+_>$,

$$\Phi(\alpha)(\mathbf{x}, \sigma) = (\mathbf{x}/\alpha, \sigma/\alpha) \quad . \tag{2.17}$$

Inversely, assuming that the $\mathcal{S}ol$ *function satisfies* (2.14)*, if* $(\mathcal{S}ol, \mathcal{G}, \mathbb{U}^p, p_{\mathbf{U}})$ *is scale-invariant with the homomorphism defined in* (2.17)*, then* (2.15) *and* (2.16) *are satisfied.*

Proof. From (i) we deduce that for any $(\mathbf{x}, \sigma)$ in $\mathbb{R}^n \times \mathbb{R}^+_>$ and any $\mathbf{u}^i \in \mathbb{U}$,

$$f(\mathcal{S}ol((\mathbf{x}, \sigma), \mathbf{u}^i)) = f\left(\alpha \mathcal{S}ol\left(\left(\frac{\mathbf{x}}{\alpha}, \frac{\sigma}{\alpha} \right), \mathbf{u}^i \right) \right)$$

which implies that the same permutation $\mathcal{S}$ will result from ordering solutions (with $\mathcal{O}rd$) on f starting from $(\mathbf{x}, \sigma)$ or on $f(\alpha \mathbf{x})$ starting from $(\mathbf{x}/\alpha, \sigma/\alpha)$, i.e. with some obvious notations $\mathcal{S}^{f(\mathbf{x})}_{(\mathbf{x},\sigma)} = \mathcal{S}^{f(\alpha \mathbf{x})}_{(\frac{\mathbf{x}}{\alpha}, \frac{\sigma}{\alpha})}$. On the other hand using (2.7) the following holds

$$\mathcal{F}^{f(\mathbf{x})}((\mathbf{x}, \sigma), \mathbf{u}) = \mathcal{G}((\mathbf{x}, \sigma), \mathcal{S}^{f(\mathbf{x})}_{(\mathbf{x},\sigma)} * \mathbf{u}) \tag{2.18}$$

$$\mathcal{F}^{f(\alpha \mathbf{x})}\left(\left(\frac{\mathbf{x}}{\alpha}, \frac{\sigma}{\alpha} \right), \mathbf{u} \right) = \mathcal{G}\left(\left(\frac{\mathbf{x}}{\alpha}, \frac{\sigma}{\alpha} \right), \mathcal{S}^{f(\alpha \mathbf{x})}_{(\frac{\mathbf{x}}{\alpha}, \frac{\sigma}{\alpha})} * \mathbf{u} \right) \quad . \tag{2.19}$$

Assuming (ii) and (iii) we find that $\mathcal{F}^{f(\mathbf{x})}((\mathbf{x}, \sigma), \mathbf{u}) = \alpha \mathcal{F}^{f(\alpha \mathbf{x})}\left(\left(\frac{\mathbf{x}}{\alpha}, \frac{\sigma}{\alpha} \right), \mathbf{u} \right)$. Using the homomorphism defined in (2.17) the previous equation reads

$$\mathcal{F}^{f(\mathbf{x})}((\mathbf{x}, \sigma), \mathbf{u}) = \Phi(1/\alpha) \mathcal{F}^{f(\alpha \mathbf{x})}\left((\Phi(\alpha)(\mathbf{x}, \sigma)), \mathbf{u} \right)$$

which is (2.13). Hence we have found an homomorphism such that (2.13) holds, which is the definition of scale-invariance. The inverse is immediate. ∎

Remark that given a CB-SARS that satisfies the conditions (i), (ii) and (iii) from the previous proposition, we can reparametrize the state of the algorithm by $\tilde{\sigma}_t = \sigma_t^2$ (if the underlying sampling distribution is Gaussian, this means parametrize by variance instead of standard deviation) leaving unchanged the parametrization for the mean vector. Then the conditions of the previous proposition are not anymore valid for the new parametrization. Yet the algorithm is still scale-invariant but a different morphism needs to be considered, namely

$$\Phi(\alpha)(\mathbf{x}, \tilde{\sigma}) = (\mathbf{x}/\alpha, \tilde{\sigma}/\alpha^2) \quad . \tag{2.20}$$

Hence the sufficient conditions derived are not general at all, however they cover typical settings for CB-SARS. Adapting however Proposition 2.9 for other parametrizations is usually easy.

2.3. Examples of CB-SARS. In order to illustrate the CB-SARS framework introduced, we present in this section several examples of CB-SARS algorithms and analyze their invariance properties.

2.3.1. Non-elitist Step-size Adaptive Evolution Strategies (ES). We consider two examples of algorithms following Definition 2.3 that were introduced under the name Evolution Strategies (ES). They all share the same sampling space $\mathbb{U}^p = \mathbb{R}^{n\times p}$. A vector $\mathbf{U}_t \in \mathbb{U}^p = \mathbb{R}^{n\times p}$ is composed of p i.i.d. standard multivariate normal distributions, i.e. $\mathbf{U}_t^i \sim \mathcal{N}(0, \mathbf{I}_n) \in \mathbb{R}^n$ and thus the joint density $p_{\mathbf{U}}(\mathbf{u}^1, \ldots, \mathbf{u}^p)$ is the product $p_{\mathcal{N}}(\mathbf{u}^1) \ldots p_{\mathcal{N}}(\mathbf{u}^p)$ where $p_{\mathcal{N}}(\mathbf{x}) = \frac{1}{(2\pi)^{n/2}} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{x}\right)$. The solution operator to sample new solutions is given by:

$$\mathcal{S}ol((\mathbf{X}_t, \sigma_t), \mathbf{U}_t^i)(= \mathbf{X}_t^i) = \mathbf{X}_t + \sigma_t \mathbf{U}_t^i \,,\; i = 1, \ldots, p \;\;, \tag{2.21}$$

and hence each candidate solution $\mathbf{X}_t^i$ follows the distribution $\mathcal{N}(\mathbf{X}_t, \sigma_t^2 \mathbf{I}_n)$.

Given the vector of ordered samples $\mathbf{Y}_t = \mathcal{S} * \mathbf{U}_t = (\mathbf{U}_t^{\mathcal{S}(1)}, \ldots, \mathbf{U}_t^{\mathcal{S}(p)})$ where $\mathcal{S}$ is the permutation resulting from the ranking of objective function values of the solutions (see (2.4)), the update equation for the mean vector $\mathbf{X}_t$ that defines the function $\mathcal{G}_1$ is given by

$$\mathbf{X}_{t+1} = \mathcal{G}_1((\mathbf{X}_t, \sigma_t), \mathbf{Y}_t) := \mathbf{X}_t + \kappa_m \sigma_t \sum_{i=1}^{p} w_i \mathbf{Y}_t^i \tag{2.22}$$

where $\kappa_m \in \mathbb{R}^+$ is usually called the learning rate and is often set to 1 and $w_i \in \mathbb{R}$ are weights that satisfy $w_1 \geq \ldots \geq w_p$ and $\sum_{i=1}^p |w_i| = 1$.

Recently, an interesting interpretation of the meaning of the vector $\sigma_t \sum_{i=1}^p w_i \mathbf{Y}_t^i$ was given: it is an approximation of the (n first coordinates) of the natural gradient of a joint criterion defined on the manifold of the family of gaussian probability distribution [1, 20].

Several step-size updates have been used with the update of the mean vector $\mathbf{X}_t$ in (2.22). First of all, consider the update derived from the cumulative step-size adaptation or path-length control without cumulation [12] that reads

$$\sigma_{t+1} = \mathcal{G}_2(\sigma_t, \mathbf{Y}_t) = \sigma_t \exp\left(\kappa_\sigma \left(\frac{\sqrt{\mu_{\mathrm{w}}} \| \sum_{i=1}^p w_i \mathbf{Y}_t^i \|}{E[\|\mathcal{N}(0, \mathbf{I}_n)\|]} - 1\right)\right) \tag{2.23}$$

where $\kappa_\sigma > 0$ is the learning rate for the step-size update usually set close to one and $\mu_{\mathrm{w}} = 1/\sum w_i^2$. The value $1/\kappa_\sigma$ is often considered as a damping parameter. The ruling principle for the update is to compare the length of the recombined step $\sum_{i=1}^p w_i \mathbf{Y}_t^i$ to its expected length if the objective function would return independent random values. Indeed if the signal given by the objective function is random, the step-size should stay constant. It is not difficult to see that in such condition, a random ordering takes place and hence the distribution of the vector $\mathbf{Y}_t$ is the same as the distribution of the vector $\mathbf{U}_t$, finally it follows that $\sqrt{\mu_{\mathrm{w}}} \sum_{i=1}^p w_i \mathbf{Y}_t^i$ is distributed according to a standard multivariate normal distribution. Hence (2.23) implements to increase the step-size if the observed length of $\sqrt{\mu_{\mathrm{w}}} \sum_{i=1}^p w_i \mathbf{Y}_t^i$ is larger than the expected length under random selection and decrease it otherwise. Overall, the update function associated to the CSA without cumulation reads

$$\mathcal{G}_{\mathrm{CSAw/o}}((\mathbf{x}, \sigma), \mathbf{y}) = \begin{pmatrix} \mathbf{x} + \sigma\kappa_m \sum_{i=1}^p w_i \mathbf{y}^i \\ \sigma \exp\left(\kappa_\sigma \left(\frac{\sqrt{\mu_{\mathrm{w}}} \| \sum_{i=1}^p w_i \mathbf{y}^i \|}{E[\|\mathcal{N}(0, \mathbf{I}_n)\|]} - 1\right)\right) \end{pmatrix} \;\;.$$

In practice the step-size update CSA is used in combination with cumulation and in the update, the norm of the selected step is replaced by the norm of a path that

cumulates steps of previous generation [11]. The CSA with cumulation is the default step-size adaptation mechanism used in the CMA-ES algorithm.

The second example we present corresponds to the natural gradient update for the step-size with exponential parametrization [10] that writes

$$\sigma_{t+1} = \sigma_t \exp\left(\frac{\kappa_\sigma}{2n}\mathrm{Tr}\left(\sum_{i=1}^{p} w_i \mathbf{Y}_t^i (\mathbf{Y}_t^i)^T - \mathbf{I}_n\right)\right) \tag{2.24}$$

$$= \sigma_t \exp\left(\frac{\kappa_\sigma}{2n}\sum_{i=1}^{p} w_i (\|\mathbf{Y}_t^i\|^2 - n)\right). \tag{2.25}$$

We then define the update function for the step-size update in xNES as

$$\mathcal{G}_{\text{xNES}}((\mathbf{x},\sigma),\mathbf{y}) = \begin{pmatrix} \mathbf{x}+\sigma\kappa_m \sum_{i=1}^{p} w_i \mathbf{y}^i \\ \sigma \exp\left(\frac{\kappa_\sigma}{2n}\sum_{i=1}^{p} w_i (\|\mathbf{y}^i\|^2 - n)\right) \end{pmatrix} . \tag{2.26}$$

Here, when κ_m and κ_σ are equal, they coincide with the step-size of the (natural) gradient step of a joint criterion defined on the manifold of Gaussian distributions with covariance matrices equal to a scalar times identity [10].

With those algorithms, it is not guaranteed that the best solution at iteration $t+1$ has a smaller objective function value than the best solution at iteration t. In the case where only positive weights are used, a compact notation for the algorithms described above is $(\mu/\mu_w, \lambda)$-ES where $\lambda = p$ and μ equals the number of non-zero weights.

Invariance properties. The two different comparison-based step-size adaptive randomized search algorithms presented in this section are translation invariant and scale-invariant. They indeed satisfy the sufficient conditions derived in Proposition 2.7 and Proposition 2.9.

2.3.2. Evolution Strategy with Self-adaptation. Another type of algorithms included in the comparison-based step-size adaptive randomized search definition are the so-called self-adaptive step-size ES. The idea of self-adaptation dates back from the 70's and consists in adding the parameters to be adapted (step-size, covariance matrix, ...) to the vector that undergo variations (mutations and recombinations) and let the selection (through the ordering function) adjusts the parameters [21, 24]. In the case where one single step-size is adapted, the step-size undergoes first a mutation: it is multiplied by a random variable following a log-normal distribution $\mathrm{Logn}(0,\tau^2)$ where $\tau \approx 1/\sqrt{n}$. The mutated step-size is then used as overall standard deviation for the multivariate normal distribution $\mathcal{N}(0,\mathbf{I}_n)$. In this case, the space $\mathbb{U}^p$ equals $\mathbb{R}^{(n+1)\times p}$. The n first coordinates of an element $\mathbf{U}_t^i \in \mathbb{U} = \mathbb{R}^{n+1}$ denoted $[\mathbf{U}_t^i]_{1\ldots n}$ $(\in \mathbb{R}^n)$ correspond to the sampled standard multivariate normal distribution vector and the last coordinate denoted $[\mathbf{U}_t^i]_{n+1}$ to the sampled normal distribution for sampling the log-normal distribution used to mutate the step-size. The solution function is defined as

$$\mathcal{S}ol((\mathbf{X}_t,\sigma_t),\mathbf{U}_t^i) = \mathbf{X}_t^i = \mathbf{X}_t + \sigma_t \exp\left(\tau[\mathbf{U}_t^i]_{n+1}\right)[\mathbf{U}_t^i]_{1\ldots n} \tag{2.27}$$

where $[\mathbf{U}_t^i]_{1\ldots n} \sim \mathcal{N}(0,\mathbf{I}_n)$ and $[\mathbf{U}_t^i]_{n+1} \sim \mathcal{N}(0,1)$. The distribution $p_{\mathbf{U}}$ admits thus a density that equals $p_{\mathbf{U}}(\mathbf{u}^1,\ldots,\mathbf{u}^p) = p_{\mathcal{N}}(\mathbf{u}^1)\ldots p_{\mathcal{N}}(\mathbf{u}^p)\,, \mathbf{u}^i \in \mathbb{R}^{n+1}$. Remark that the ordering function selects the couple multivariate normal distribution and log-normal distribution used to mutate the step-size at the same time. Assuming that only the

best solution plays a role in the update of $\mathbf{X}_t$ (i.e. it corresponds to a single non-zero weight in the recombination equation (2.22)), the update for the mean vector reads

$$\mathbf{X}_t = \mathbf{X}_t + \sigma_t \exp(\tau[\mathbf{Y}_t^1]_{n+1})[\mathbf{Y}_t^1]_{1\ldots n} \tag{2.28}$$

and the update for the step-size is

$$\sigma_{t+1} = \sigma_t \exp(\tau[\mathbf{Y}_t^1]_{n+1}) \ . \tag{2.29}$$

A step-size adaptive Evolution Strategy satisfying (2.27),(2.28) and (2.29) is called $(1, p)$ self-adaptive step-size ES ($(1, p)$-SA). The $(1, p)$ refers to the fact that a single solution is selected out of the p. The update function $\mathcal{G}$ for the $(1, p)$-SA reads

$$\mathcal{G}_{(1,p)-\mathrm{SA}}((\mathbf{x}, \sigma), \mathbf{y}) = \begin{pmatrix} \mathbf{x}+\sigma \exp(\tau[\mathbf{y}^1]_{n+1})[\mathbf{y}^1]_{1\ldots n} \\ \sigma \exp(\tau[\mathbf{y}^1]_{n+1}) \end{pmatrix} \ .$$

We see thus that the step-size is adapted by the selection that occurs through the ordering. The rationale behind the method being that unadapted step-size cannot successfully give good solutions and that selection will adapt (for free) the step-size (explaining thus the terminology "self-adaptation"). Self-adaptive algorithms have been popular in the 90's certainly due to the fact that its idea is simple and attractive. However self-adaptation leads in general to too small step-size. Different variants of self-adaptation using multiple parents and recombinations exist, we refer to the review paper [5] for further readings and references.

Invariances. In virtue of Proposition 2.7 and Proposition 2.9 the $(1, p)$-SA is translation and scale-invariant.

The linear convergence of the self-adaptive ES algorithm described in this section in dimension 1 was proven in [3] on spherical functions using the Markov chain approach presented here.

2.3.3. Step-size Random Search or Compound Random Search or (1+1)-ES with $1/5$ Success Rule. The last example presented is an algorithm where the sequence $f(\mathbf{X}_t)$ is decreasing, i.e. updates that only improve or leave $\mathbf{X}_t$ unchanged are performed. At each iteration a single new solution is sampled from $\mathbf{X}_t$, i.e.

$$\mathbf{X}_t^1 = \mathbf{X}_t + \sigma_t \mathbf{U}_t^1$$

where $\mathbf{U}_t^1 \in \mathbb{R}^n$ follows a standard multivariate normal distribution, and hence $\mathbf{X}_t^1$ follows the distribution $\mathcal{N}(\mathbf{X}_t, \sigma_t^2 \mathbf{I}_n)$. The step $\mathbf{U}_t^1$ is accepted if the candidate solution is better than the current one and rejected otherwise. Let us denote $\mathbf{U}_t^2 = 0 \in \mathbb{R}^n$ the zero vector and take $\mathbf{U}_t = (\mathbf{U}_t^1, \mathbf{U}_t^2)$. Hence $\mathbb{U}^p = \mathbb{R}^{n\times 2}$ and the probability distribution of $\mathbf{U}$ equals $p_{\mathbf{U}}(\mathbf{u}^1, \mathbf{u}^2) = p_{\mathcal{N}}(\mathbf{u}^1)\delta_0(\mathbf{u}^2)$ where δ_0 is the Dirac delta function. The $\mathcal{S}ol$ function corresponds then to the function in (2.21).

The update equation for $\mathbf{X}_t$ is similar to (2.22) with weights $(w_1, w_2) = (1, 0)$. Remark that contrary to the algorithms presented before, the sampled step $\mathbf{U}_t$, the selected step $\mathbf{Y}_t$ and the new mean $\mathbf{X}_{t+1}$ have a singular part w.r.t. the Lebesgue measure. An algorithm following such an update is often referred as $(1 + 1)$-ES but was also introduced under the name Markov monotonous search [26], step-size random search [23] or compound random search [9].

The adaptation of the step-size idea starts from the observation that if the step-size is very small, the probability of success (i.e. to sample a better solution) is approximately one-half but the improvements are small because the step is small. On

the opposite if the step-size is too large, the probability of success will be small, typically the optimum will be overshoot and the improvement will also be very small. In between lies an optimal step-size associated to an optimal probability of success [23, 21, 9]. A proposed adaptive step-size algorithm consists in trying to maintain a probability of success (i.e. probability to sample a better solution) to a certain target value p_{target}, increase the step-size in case the probability of success is larger than p_{target} and decrease it otherwise [9, 21, 22]. The optimal probability of success, i.e. allowing to obtain an optimal convergence rate has been computed on the sphere function $f(\mathbf{x}) = \|\mathbf{x}\|^2$ for dimension of the search problem going to infinity and is roughly equal to 0.27 [23, 21]. Another function where the asymptotic optimal probability of success was computed is the corridor function[2] where it is equal to $1/(2e)$ [22]. As a trade-off between the probability of success on the sphere and on the corridor, the target probability is often taken equal to $1/5 = 0.20$ and gave the name one-fifth success rule to the step-size adaptive algorithm. We call the algorithm with p_{target} as target success probability the *generalized one-fifth success rule*.[3]

Several implementations of the generalized one-fifth success rule exist. In some implementations, the probability of success is estimated by fixing a step-size for a few iterations, counting the number of successful solutions and deducing an estimation of the probability of success. The step-size is then increased if the probability of success is larger than p_{target} and decreased otherwise [21, 22]. A somehow simpler implementation consists in estimating at each iteration the probability of success as $1_{\{f(\mathbf{X}_t^1)<f(\mathbf{X}_t)\}} = 1_{\{\mathbf{Y}_t^1\neq 0\}}$[4]: this indicator function being equal to one in case of success and zero otherwise. Consequently the algorithm will increase the step-size after a successful step and decrease it otherwise as proposed in [9, 15]. The update rule for the step-size reads

$$\sigma_{t+1} = \sigma_t \exp\left(\kappa_\sigma \frac{1_{\{\mathbf{Y}_t^1\neq 0\}} - p_{\text{target}}}{1 - p_{\text{target}}}\right)$$

where $\kappa_\sigma > 0$ is a learning rate coefficient. Denoting $\gamma = \exp(\kappa_\sigma)$ and $q = \frac{1-p_{\text{target}}}{p_{\text{target}}}$ (for a target success probability set to $1/5$, the parameter $q = 4$) yields

$$\sigma_{t+1} = \sigma_t \left(\gamma 1_{\{\mathbf{Y}_t\neq 0\}} + \gamma^{-1/q} 1_{\{\mathbf{Y}_t^1=0\}}\right) = \sigma_t \left((\gamma - \gamma^{-1/q}) 1_{\{\mathbf{Y}_t^1\neq 0\}} + \gamma^{-1/q}\right) \quad . \tag{2.30}$$

Overall, the update transformation for the (1+1)-ES with generalized one-fifth success rule is

$$\mathcal{G}_{(1+1)_{1/5}}((\mathbf{x},\sigma),\mathbf{y}) = \begin{pmatrix} \mathbf{x}+\sigma\mathbf{y}^1 \\ \sigma\left((\gamma-\gamma^{-1/q})1_{\{\mathbf{y}^1\neq 0\}}+\gamma^{-1/q}\right) \end{pmatrix} \quad .$$

[2] The corridor function is defined as $f(\mathbf{x}) = \mathbf{x}_1$ for $-b < \mathbf{x}_2 < b, \ldots - b < \mathbf{x}_n < b$, for $b > 0$ otherwise $+\infty$.

[3] Note that p_{target} does not correspond to the optimal probability of success as indeed if the probability of success equals the target probability, the step-size is kept constant. Hence if convergence occurs the achieved probability of success is smaller than the target probability. Therefore, on the sphere, if convergence occurs, $p_{\text{target}} = 0.20$ corresponds to an achieved probability of success smaller than 0.20, hence a probability of success smaller than optimal which will consequently favor larger step-sizes as the probability of success decreases with increasing step-sizes [4].

[4] This equality is true only almost everywhere.

Invariance. Using again Proposition 2.7 and Proposition 2.9, the $(1+1)$-ES with generalized one-fifth success rule is translation and scale-invariant.

REMARK 1. *In all the examples presented, the p components $(\mathbf{U}_t^i)_{1\leq i\leq p}$ of the vectors $\mathbf{U}_t$ are independent. It is however not a requirement of our theoretical setting.*

3. Scaling-Invariant Functions. In this section we define the class of scaling-invariant functions that preserve the f-ordering of two points centered with respect to a reference point $\mathbf{x}^\star$ when they are scaled by any given factor.

DEFINITION 3.1 (Scaling-invariant function). *A function $f : \mathbb{R}^n \mapsto \mathbb{R}$ is scaling-invariant with respect to $\mathbf{x}^\star \in \mathbb{R}^n$, if for all $\rho > 0$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$*

$$f(\rho(\mathbf{x}-\mathbf{x}^\star)) \leq f(\rho(\mathbf{y}-\mathbf{x}^\star)) \Leftrightarrow f(\mathbf{x}-\mathbf{x}^\star) \leq f(\mathbf{y}-\mathbf{x}^\star) \ . \tag{3.1}$$

This definition implies that two points $\mathbf{x}-\mathbf{x}^\star$ and $\mathbf{y}-\mathbf{x}^\star$ belong to the same level set if and only if for all $\rho > 0$ also $\rho(\mathbf{x}-\mathbf{x}^\star)$ and $\rho(\mathbf{y}-\mathbf{x}^\star)$ belong to the same level set, i.e.

$$f(\mathbf{x}-\mathbf{x}^\star) = f(\mathbf{y}-\mathbf{x}^\star) \Leftrightarrow f(\rho(\mathbf{x}-\mathbf{x}^\star)) = f(\rho(\mathbf{y}-\mathbf{x}^\star)) \ .$$

Hence, scaling-invariance can be equivalently defined with strict inequalities in (3.1). Remark that if f is scaling-invariant, then for any g strictly increasing the composite $g \circ f$ is also scaling-invariant.

PROPOSITION 3.2. *Let f be a scaling-invariant function, then, f cannot admit any strict local optima except $\mathbf{x}^\star$. In addition, on a line crossing $\mathbf{x}^\star$ a scaling invariant function is either constant equal to $f(\mathbf{x}^\star)$ or cannot admit a local plateau, i.e. a ball where the function is locally constant.*

Proof. We can assume w.l.o.g. scaling-invariance with respect to $\mathbf{x}^\star = 0$. Assume to get a contradiction that f admits a strict local maximum different from $\mathbf{x}^\star$, i.e. there exist $\mathbf{x}_0$ and $\epsilon > 0$ such that for all $\mathbf{x} \in B(\mathbf{x}_0, \epsilon)$ (open ball of center $\mathbf{x}_0$ and radius ϵ), $f(\mathbf{x}) < f(\mathbf{x}_0)$. We now consider a point $\mathbf{x}_1$ belonging to $B(\mathbf{x}_0, \epsilon)$ and to the line $(0, \mathbf{x}_0)$ such that $\|\mathbf{x}_1\| > \|\mathbf{x}_0\|$. Then $f(\mathbf{x}_1) < f(\mathbf{x}_0)$ as $\mathbf{x}_0$ is a strict local maximum and $\mathbf{x}_1$ can be written $\mathbf{x}_1 = \theta\mathbf{x}_0$ with $\theta > 1$ as $\mathbf{x}_1 \in (0, \mathbf{x}_0)$ and has a larger norm than $\mathbf{x}_0$. Hence $f(\mathbf{x}_0) > f(\mathbf{x}_1) = f(\theta\mathbf{x}_0)$ which is by the scaling-invariance property equivalent to $f(\mathbf{x}_0/\theta) > f(\mathbf{x}_0)$. However, $\mathbf{x}_0/\theta \in B(\mathbf{x}_0, \epsilon)$ as $\|\mathbf{x}_0/\theta - \mathbf{x}_0\| = |1-1/\theta|\|\mathbf{x}_0\| = (\theta-1)\|\mathbf{x}_0\|/\theta = \|\mathbf{x}_1 - \mathbf{x}_0\|/\theta < \epsilon/\theta < \epsilon$. Then we have found a point $\mathbf{x}_0/\theta \in B(\mathbf{x}_0, \epsilon)$ that has a function value strictly larger than $f(\mathbf{x}_0)$ which contradicts the fact that $\mathbf{x}_0$ is a strict local maximum. The same reasoning holds to prove that the function has no strict local minimum.

The fact that the function is constant on a line crossing $\mathbf{x}^\star$ or cannot admit a local plateau, comes from the fact that if the function is non-constant on a line and admits a local plateau, then we can find two points from the plateau $\mathbf{x}$ and $\mathbf{y}$ with equal function value such that the point $\mathbf{x}$ is at the extremity of the local plateau, then we just scale $\mathbf{x}$ and $\mathbf{y}$ such that $\mathbf{x}$ is outside the plateau and $\mathbf{y}$ stays on the plateau. By the scaling invariant property, the scaled points should still have an equal function value which is impossible as we have scaled $\mathbf{x}$ to be outside the plateau. ∎

Examples of scaling-invariant functions include linear functions or composite of norm functions by functions in $\mathcal{M}$, i.e. $f(\mathbf{x}) = g(\|\mathbf{x}\|)$ where $\|.\|$ is a norm on $\mathbb{R}^n$ and $g \in \mathcal{M}$. Thus the famous sphere function $f(\mathbf{x}) = \sum_{i=1}^n \mathbf{x}_i^2$ which is the square of the Euclidian norm or more generally any convex quadratic function $f(\mathbf{x}) = (\mathbf{x}-\mathbf{x}^\star)^T H(\mathbf{x}-\mathbf{x}^\star)$ with $H \in \mathbb{R}^{n\times n}$ positive definite symmetric are scaling-invariant

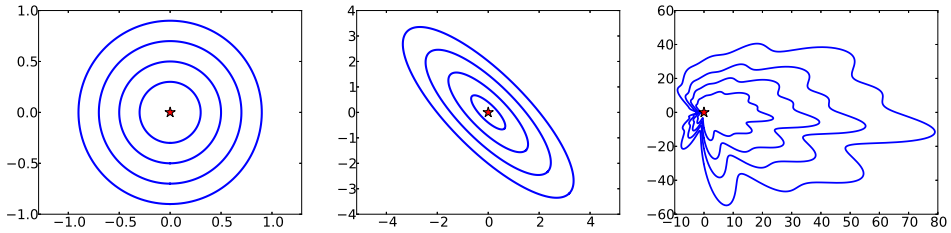

FIG. 3.1. *Illustration of scaling-invariant functions w.r.t. the point* $\mathbf{x}^\star$ *in the middle depicted with a star. The three functions are composite of* $g \in \mathcal{M}$ *by* $f(\mathbf{x} - \mathbf{x}^\star)$ *where* f *is a positively homogeneous function (see Definition 3.3). Left: composite of* $g \in \mathcal{M}$ *and* $f(\mathbf{x}) = \|\mathbf{x} - \mathbf{x}^\star\|$. *Middle: composite of* $g \in \mathcal{M}$ *and* $f(\mathbf{x}) = (\mathbf{x} - \mathbf{x}^\star)^T \mathbf{A}(\mathbf{x} - \mathbf{x}^\star)$ *for* $\mathbf{A}$ *symmetric positive definite. Both functions on the left have convex sublevel sets contrary to the one on the right.*

functions with respect to $\mathbf{x}^\star$. The sublevel sets defined as the sets $\{\mathbf{x} \in \mathbb{R}^n, f(\mathbf{x}) \leq c\}$ for $c \in \mathbb{R}$ for those previous examples are convex sets, i.e. the functions are *quasi-convex*. However, functions with non-convex sublevel sets can also be scaling-invariant (see Figure 3.1).

A particular class of scaling-invariant functions are positively homogeneous functions whose definition is reminded below.

DEFINITION 3.3 (Positively homogeneous functions). *A function* $f : \mathbb{R}^n \mapsto \mathbb{R}$ *is said positively homogeneous with degree* α *if for all* $\rho > 0$ *and for all* $\mathbf{x} \in \mathbb{R}^n$, $f(\rho \mathbf{x}) = \rho^\alpha f(\mathbf{x})$. From this definition it follows that if a function $\hat{f}$ is positively homogeneous with degree α then $\hat{f}(\mathbf{x} - \mathbf{x}^\star)$ is scaling-invariant with respect to $\mathbf{x}^\star$ for any $\mathbf{x}^\star \in \mathbb{R}^n$. Remark that positive homogeneity is not always preserved if f is composed by a strictly increasing transformation.

Examples of positively homogeneous functions are linear functions that are positively homogeneous functions with degree 1. Also, every function deriving from a norm is positively homogeneous with degree 1. Examples of scaling-invariant functions deriving from positively homogenous functions are depicted in Figure 3.1.

In the paper [4], stability of the normalized Markov chain is studied on functions $h = g \circ f$ where f is positive homogeneous and $g \in \mathcal{M}$.

4. Joint Markov chains on Scaling-Invariant Functions. We consider CB-SARS algorithms that are translation invariant and scale-invariant satisfying the properties (2.14), (2.15) and (2.16) in Proposition 2.9. The functions considered are scaling-invariant with $\mathbf{x}^\star = 0$. This can be assumed w.l.o.g. because of the translation invariance of the algorithms. We prove under those conditions that $\mathbf{X}_t/\sigma_t$ is a homogeneous Markov chain.

PROPOSITION 4.1. *Consider a scaling-invariant (in zero) objective function* f *optimized by* $(\mathcal{S}ol, (\mathcal{G}_1, \mathcal{G}_2), \mathbb{U}^p, p_{\mathbf{U}})$, *a CB-SARS algorithm assumed to be translation-invariant and scale-invariant satisfying* (2.14), (2.15) *and* (2.16). *Let* $(\mathbf{X}_t, \sigma_t)_{t \in \mathbb{N}}$ *be the Markov chain associated to this CB-SARS. Let* $\mathbf{Z}_t = \frac{\mathbf{X}_t}{\sigma_t}$ *for all* $t \in \mathbb{N}$. *Then* $(\mathbf{Z}_t)_{t \in \mathbb{N}}$ *is a homogeneous Markov chain that can be defined independently of* $(\mathbf{X}_t, \sigma_t)$, *provided* $\mathbf{Z}_0 = \mathbf{X}_0/\sigma_0$ *via*

$$\mathbf{Z}_t^i = \mathcal{S}ol((\mathbf{Z}_t, 1), \mathbf{U}_t^i), i = 1, \ldots, p \tag{4.1}$$

$$\mathcal{S} = \mathcal{O}rd(f(\mathbf{Z}_t^1), \ldots, f(\mathbf{Z}_t^p)) \tag{4.2}$$

$$\mathbf{Z}_{t+1} = G(\mathbf{Z}_t, \mathcal{S} * \mathbf{U}_t) \tag{4.3}$$

where the function G equals for all $\mathbf{z} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{U}^p$

$$G(\mathbf{z}, \mathbf{y}) = \frac{\mathcal{G}_1((\mathbf{z}, 1), \mathbf{y})}{\mathcal{G}_2(1, \mathbf{y})} \quad . \tag{4.4}$$

Translated in words, the normalized homogeneous Markov chain $(\mathbf{Z}_t)_{t\in\mathbb{N}}$ of the previous definition is generated independently of $(\mathbf{X}_t, \sigma_t)$ by (i) sampling candidate solutions with the $\mathcal{S}ol$ function starting from $(\mathbf{Z}_t, 1)$ (i.e. with step-size 1) (ii) ordering the candidate solutions (iii) using the ranking of the candidate solutions to compute $\mathbf{Z}_{t+1}$ as the ratio of $\mathcal{G}_1((\mathbf{Z}_t, 1), \mathcal{S} * \mathbf{U}_t)$ (i.e. the mean update equation but with step-size 1 and starting from $\mathbf{Z}_t$) divided by the multiplicative update for the step-size taken in $\sigma = 1$.

REMARK 2. *If the function is scale-invariant in $\mathbf{x}^\star$ with $\mathbf{x}^\star$ being not necessarily zero, then the normalized Markov chain to consider in the previous proposition is $\mathbf{Z}_t = \frac{\mathbf{X}_t - \mathbf{x}^\star}{\sigma_t}$.*

REMARK 3. *The previous proposition assumes that scale-invariance is satisfied via the conditions specified in Propositions 2.9. We believe however that when a CB-SARS is scale-invariant under different conditions, a normalized homogeneous Markov chain can be found. For instance when the parametrization $(\mathbf{X}_t, \tilde{\sigma}_t) = (\mathbf{X}_t, \sigma_t^2)$ is used (see discussion around (2.20)) the normalized Markov chain is $\mathbf{X}_t / \sqrt{\tilde{\sigma}_t}$.*

Proof. (of Proposition 4.1) Consider a scaling-invariant function in zero, f. Candidate solutions sampled according to the $\mathcal{S}ol$ operator satisfy according to property (2.14) $\mathcal{S}ol((\mathbf{x}, \sigma), \mathbf{u}^i) = \sigma \mathcal{S}ol((\mathbf{x}/\sigma, 1), \mathbf{u}^i)$. However in a comparison-based step-size adaptive randomized search, the permutation $\mathcal{S}$ results from ordering the objective function of the candidate solutions, i.e. ordering $f(\mathcal{S}ol((\mathbf{x}, \sigma), \mathbf{u}^i))$ which is the same as ordering $f(\sigma \mathcal{S}ol((\mathbf{x}/\sigma, 1), \mathbf{u}^i))$ according to property (2.14). By the scaling-invariant property of the function f, we see that it is the same as ordering $f(\mathcal{S}ol((\mathbf{x}/\sigma, 1), \mathbf{u}^i))$. In other words, on a scaling-invariant function, $\mathcal{S} = \mathcal{S}^f_{(\mathbf{X}_t, \sigma_t)} = \mathcal{S}^f_{(\mathbf{X}_t/\sigma_t, 1)}$ (putting the initial state as lower subscript).
Let $\mathbf{X}_t, \sigma_t, \mathbf{U}_t$ be given and let $\mathbf{Z}_t = \mathbf{X}_t/\sigma_t$, then $\mathbf{Z}_{t+1} = \frac{\mathbf{X}_{t+1}}{\sigma_{t+1}} = \frac{\mathcal{G}_1((\mathbf{X}_t, \sigma_t), \mathcal{S} * \mathbf{U}_t)}{\mathcal{G}_2(\sigma_t, \mathcal{S} * \mathbf{U}_t)}$. Because of properties (2.15) and (2.16), $\mathbf{Z}_{t+1} = \frac{\sigma_t \mathcal{G}_1((\mathbf{X}_t/\sigma_t, 1), \mathcal{S}^f_{(\mathbf{X}_t, \sigma_t)} * \mathbf{U}_t)}{\sigma_t \mathcal{G}_2(1, \mathcal{S}^f_{(\mathbf{X}_t, \sigma_t)} * \mathbf{U}_t)}$ and thus $\mathbf{Z}_{t+1} = \frac{\mathcal{G}_1((\mathbf{Z}_t, 1), \mathcal{S}^f_{(\mathbf{Z}_t, 1)} * \mathbf{U}_t)}{\mathcal{G}_2(1, \mathcal{S}^f_{(\mathbf{Z}_t, 1)} * \mathbf{U}_t)}$. Since we have assume translation invariance of the algorithm, the same construction holds if $\mathbf{x}^\star \neq 0$ with $\mathbf{Z}_t = \frac{\mathbf{X}_t - \mathbf{x}^\star}{\sigma_t}$. □

Because we assume scale-invariance via the properties of Proposition 2.9, the step-size update has a specific shape. Indeed (2.16) implies that

$$\sigma_{t+1} = \sigma_t \mathcal{G}_2(1, \mathbf{Y}_t) \tag{4.5}$$

where $\mathbf{Y}_t = \mathcal{S} * \mathbf{U}_t$. Let us denote the multiplicative step-size update as $\eta^\star$, i.e.

$$\eta^\star(\mathbf{Y}_t) = \mathcal{G}_2(1, \mathbf{Y}_t) \quad . \tag{4.6}$$

As explained in the proof of the previous proposition, on a scaling-invariant function the ranking permutation is the same starting from $(\mathbf{X}_t, \sigma_t)$ or from $(\mathbf{Z}_t, 1)$ such that we find that on scaling-invariant functions

$$\eta^\star(\mathcal{S}_{(\mathbf{X}_t, \sigma_t)} * \mathbf{U}_t) = \eta^\star(\mathcal{S}_{(\mathbf{Z}_t, 1)} * \mathbf{U}_t) \tag{4.7}$$

where $\mathcal{S}_{(\mathbf{X}_t, \sigma_t)}$ is the permutation giving the ranking starting from the state $(\mathbf{X}_t, \sigma_t)$ that equals $\mathcal{S}_{(\mathbf{Z}_t, 1)}$ the ranking permutation starting from $(\mathbf{Z}_t, 1)$.

Remark that the construction of the homogeneous Markov chain in the previous proposition only requires that the function is scaling-invariant. We do not assume here that the function has a unique global optimum. Hence the function could be the linear function $f(\mathbf{x}) = [\mathbf{x}]_1$. We will now explicit the transition functions G associated to the different comparison-based step-size adaptive randomized search examples described above.

Non-elitist ES with CSAw/o, xNES *step-size adaptation.* Given a vector $\mathbf{y} \in \mathbb{U}^p = \mathbb{R}^{n\times p}$, the update functions for the normalized Markov chains $\mathbf{Z}$ associated to the different algorithms given in Section 2.3.1 read:

$$G_{\text{CSAw/o}}(\mathbf{z},\mathbf{y}) = \frac{\mathbf{z} + \kappa_m \sum_{i=1}^p w_i \mathbf{y}^i}{\exp\left(\kappa_\sigma \left(\frac{\sqrt{\mu_\text{w}}\|\sum_{i=1}^p w_i \mathbf{y}^i\|}{E[\|\mathcal{N}(0,\mathbf{I}_n)\|]} - 1\right)\right)} \tag{4.8}$$

$$G_{\text{xNES}}(\mathbf{z},\mathbf{y}) = \frac{\mathbf{z} + \kappa_m \sum_{i=1}^p w_i \mathbf{y}^i}{\exp\left(\frac{\kappa_\sigma}{2n}\left(\sum_{i=1}^p w_i(\|\mathbf{y}^i\|^2 - n)\right)\right)} \ . \tag{4.9}$$

$(1,p)$*-SA-ES.* For the $(1,p)$-SA-ES described in Section 2.3.2, the transition function G reads for $\mathbf{y} \in \mathbb{U}^p$

$$G_{\text{SA}}(\mathbf{z},\mathbf{y}) = \frac{\mathbf{z} + \exp(\tau[\mathbf{y}^1]_{n+1})[\mathbf{y}^1]_{1\ldots n}}{\exp(\tau[\mathbf{y}^1]_{n+1})} \tag{4.10}$$

$(1+1)$*-ES with generalized* $1/5$ *success rule.* The transition function G for the normalized Markov chain of the $(1+1)$-ES with generalized one-fifth success rule reads, for all $\mathbf{z} \in \mathbb{R}^n$, for all $\mathbf{y}$ in $\mathbb{R}^{n\times 2}$

$$G_{(1+1)}(\mathbf{z},\mathbf{y}) = \frac{\mathbf{z} + \mathbf{y}^1}{\left((\gamma - \gamma^{-1/q})1_{\{\mathbf{y}^1 \neq 0\}} + \gamma^{-1/q}\right)} \ . \tag{4.11}$$

5. Sufficient Conditions for Linear Convergence of CB-SARS on Scaling-Invariant Functions. We consider throughout this section that $(\mathbf{X}_t, \sigma_t)_{t\in\mathbb{N}}$ is a Markov chain resulting from a CB-SARS (as defined in Definition 2.3) that is translation invariant and scale-invariant satisfying the conditions of Proposition 2.9. The function optimized is a scaling-invariant function f in zero. In this context, let $(\mathbf{Z}_t = \frac{\mathbf{X}_t}{\sigma_t})_{t\in\mathbb{N}}$ be the homogeneous Markov chain defined in Proposition 4.1.

For proving linear convergence, we investigate the log-progress $\ln \|\mathbf{X}_{t+1}\|/\|\mathbf{X}_t\|$. The chains $(\mathbf{X}_t, \sigma_t)_{t\in\mathbb{N}}$ and $(\mathbf{Z}_t)_{t\in\mathbb{N}}$ being connected by the relation $\mathbf{Z}_t = \mathbf{X}_t/\sigma_t$, the log-progress can be expressed as

$$\ln \frac{\|\mathbf{X}_{t+1}\|}{\|\mathbf{X}_t\|} = \ln \frac{\|\mathbf{Z}_{t+1}\|\eta^\star(\mathbf{Y}(\mathbf{Z}_t,\mathbf{U}_t))}{\|\mathbf{Z}_t\|} \tag{5.1}$$

where the ordered vector $\mathcal{S}_{(\mathbf{Z}_t,1)} * \mathbf{U}_t$ is denoted $\mathbf{Y}(\mathbf{Z}_t,\mathbf{U}_t)$ to signify its dependency in $\mathbf{Z}_t$ and $\mathbf{U}_t$, i.e.

$$\mathbf{Y}(\mathbf{z},\mathbf{u}) = \mathcal{S}_{(\mathbf{z},1)} * \mathbf{u} = \mathcal{O}rd(f(\mathcal{S}ol((\mathbf{z},1),\mathbf{u}^i)_{i=1,\ldots,p})) * \mathbf{u} \ . \tag{5.2}$$

For (5.1) we have used the fact that the step-size change starting from $(\mathbf{X}_t,\sigma_t)$ equals the step-size change starting from $(\mathbf{Z}_t,1) = (\mathbf{X}_t/\sigma_t,1)$ (see (4.7)). Using the property of the logarithm, we express $\frac{1}{t}\ln\frac{\|\mathbf{X}_t\|}{\|\mathbf{X}_0\|}$ as

$$\frac{1}{t}\ln\frac{\|\mathbf{X}_t\|}{\|\mathbf{X}_0\|} = \frac{1}{t}\sum_{k=0}^{t-1}\ln\frac{\|\mathbf{X}_{k+1}\|}{\|\mathbf{X}_k\|} = \frac{1}{t}\sum_{k=0}^{t-1}\ln\frac{\|\mathbf{Z}_{k+1}\|}{\|\mathbf{Z}_k\|}\eta^\star(\mathbf{Y}(\mathbf{Z}_t,\mathbf{U}_t)) \ . \tag{5.3}$$

Let us define for $\mathbf{z} \in \mathcal{Z}$, $\mathcal{R}(\mathbf{z})$ the expectation of the logarithm of $\eta^\star(\mathbf{Y}(\mathbf{z}, \mathbf{U}))$ for $\mathbf{U} \sim p_{\mathbf{U}}$, i.e.

$$\mathcal{R}(\mathbf{z}) = E[\ln(\eta^\star(\mathbf{Y}(\mathbf{z}, \mathbf{U}))] \tag{5.4}$$

$$= \int \ln\left(\eta^\star\left(\mathcal{O}rd(f(\mathcal{S}ol((\mathbf{z}, 1), \mathbf{u}^i))_{i=1,\ldots,p})\right) * \mathbf{u}\right) p_{\mathbf{U}}(\mathbf{u}) d\mathbf{u} \ . \tag{5.5}$$

Linear convergence. Almost sure linear convergence can be proven by exploiting (5.3) that suggests the application of a Law of Large Numbers (LLN) for Markov chains. Sufficient conditions for proving a LLN for Markov chains are φ-irreducibility, Harris recurrence and positivity whose definitions are briefly reviewed, see however Meyn and Tweedie for more background [17].

Let $\mathbf{Z} = (\mathbf{Z}_t)_{t\in\mathbb{N}}$ be a Markov chain defined on a state space $\mathcal{Z}$ equipped with the Borel sigma-algebra $\mathcal{B}(\mathcal{Z})$. We denote $P^t(\mathbf{z}, A)$, $t \in \mathbb{N}$, $\mathbf{z} \in \mathcal{Z}$ and $A \in \mathcal{B}(\mathcal{Z})$ the transition probabilities of the chain

$$P^t(\mathbf{z}, A) = P_{\mathbf{z}}(\mathbf{Z}_t \in A) \tag{5.6}$$

where $P_{\mathbf{z}}$ and $E_{\mathbf{z}}$ denote the probability law and expectation of the chain under the initial condition $\mathbf{Z}_0 = \mathbf{z}$. If a probability μ on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ is the initial distribution of the chain, the corresponding quantities are denoted P_μ and E_μ. For $t = 1$, the transition probability in Eq. (5.6) is denoted $P(\mathbf{z}, A)$. The chain $\mathbf{Z}$ is φ-irreducible if there exists a non-zero measure φ such that for all $A \in \mathcal{B}(\mathcal{Z})$ with $\varphi(A) > 0$, for all $\mathbf{z}_0 \in \mathcal{Z}$, the chain started at $\mathbf{z}_0$ has a positive probability to hit A, that is there exists $t \in \mathbb{N}_>$ such that $P^t(\mathbf{z}_0, A) > 0$. A σ-finite measure π on $\mathcal{B}(\mathcal{Z})$ is said invariant if it satisfies

$$\pi(A) = \int \pi(d\mathbf{z}) P(\mathbf{z}, A), \ \ A \in \mathcal{B}(\mathcal{Z}) \ .$$

If the chain $\mathbf{Z}$ is φ-irreducible and admits an invariant probability measure then it is called *positive*. A small set is a set C such that for some $\delta > 0$ and $t > 0$ and some non trivial probability measure ν_t,

$$P^t(\mathbf{z}, .) \geq \delta \nu_t(.), \mathbf{z} \in C \ .$$

The set C is then called a ν_t-small set. Consider a small set C satisfying the previous equation with $\nu_t(C) > 0$ and denote $\nu_t = \nu$. The chain is called aperiodic if the g.c.d. of the set

$$E_C = \{k \geq 1 : C \text{ is a } \nu_k\text{-small set with } \nu_k = \alpha_k \nu \text{ for some } \alpha_k > 0\}$$

is one for some (and then for every) small set C.

A φ-irreducible Markov chain is *Harris-recurrent* if for all $A \subset \mathcal{Z}$ with $\varphi(A) > 0$, and for all $\mathbf{z} \in \mathcal{Z}$, the chain will eventually reach A with probability 1 starting from $\mathbf{z}$, formally if $P_{\mathbf{z}}(\eta_A = \infty) = 1$ where η_A be the *occupation time* of A, i.e. $\eta_A = \sum_{t=1}^{\infty} 1_{\mathbf{Z}_t \in A}$. An (Harris-)recurrent chain admits an unique (up to a constant multiple) invariant measure [17, Theorem 10.0.4].

Typical sufficient conditions for a Law of Large Numbers to hold are φ-irreducibility, positivity and Harris-recurrence:

THEOREM 5.1. *[[17] Theorem 17.0.1] Assume that $\mathbf{Z}$ is a positive Harris-recurrent chain with invariant probability π. Then the LLN holds for any g with $\pi(|g|) = \int |g(\mathbf{x})| \pi(d\mathbf{x}) < \infty$, that is for any initial state $\mathbf{Z}_0$, $\lim_{t\to\infty} \frac{1}{t} \sum_{k=0}^{t-1} g(\mathbf{Z}_k) = \pi(g)$ a.s.*

This theorem allows to state sufficient conditions for the almost sure linear convergence of scale-invariant CB-SARS satisfying the assumptions of Proposition 4.1. However, before stating those sufficient conditions, let us remark that as a consequence of (5.1), assuming positivity of $\mathbf{Z}$ and denoting π its invariant probability measure, and assuming that (i) $\mathbf{Z}_0 \sim \pi$, (ii) $\int \ln \|\mathbf{z}\| \pi(d\mathbf{z}) < \infty$ and (iii) $\int \mathcal{R}(\mathbf{z})\pi(d\mathbf{z}) < \infty$, then for all $t \geq 0$

$$E_\pi \left[\ln \frac{\|\mathbf{X}_{t+1}\|}{\|\mathbf{X}_t\|} \right] = \int E_{\mathbf{U}\sim p_\mathbf{U}}[\ln(\eta^\star(\mathbf{Y}(\mathbf{z},\mathbf{U})))]\pi(d\mathbf{z}) = \int \mathcal{R}(\mathbf{z})\pi(d\mathbf{z}) \ . \tag{5.7}$$

We define the convergence rate CR as the opposite of the RHS of the previous equation, i.e.

$$\mathrm{CR} = -\int E_{\mathbf{U}\sim p_\mathbf{U}}[\ln(\eta^\star(\mathbf{Y}(\mathbf{z},\mathbf{U})))]\pi(d\mathbf{z}) = -\int \mathcal{R}(\mathbf{z})\pi(d\mathbf{z}) \ . \tag{5.8}$$

We now state sufficient conditions such that linear convergence at the rate CR holds almost surely independently of the initial state.

THEOREM 5.2 (Almost sure linear convergence). *Let $(\mathbf{X}_t, \sigma_t)_{t\in\mathbb{N}}$ be the recursive sequence generated by a translation and scale-invariant CB-SARS satisfying the assumptions of Proposition 4.1 and optimizing a scaling-invariant function. Let $(\mathbf{Z}_t)_{t\in\mathbb{N}}$ be the homogeneous Markov chain defined in Proposition 4.1. Assume that $(\mathbf{Z}_t)_{t\in\mathbb{N}}$ is Harris-recurrent and positive with invariant probability measure π, that $E_\pi \ln \|\mathbf{z}\| < \infty$ and $E_\pi \mathcal{R}(\mathbf{z})d\mathbf{z} < \infty$. Then for all $\mathbf{X}_0$, for all σ_0, linear convergence holds asymptotically almost surely, i.e.*

$$\lim_{t\to\infty} \frac{1}{t} \ln \frac{\|\mathbf{X}_t\|}{\|\mathbf{X}_0\|} = -\mathrm{CR} \text{ and } \lim_{t\to\infty} \frac{1}{t} \ln \frac{\sigma_t}{\sigma_0} = -\mathrm{CR} \ .$$

Proof. Using (5.3) we obtain

$$\frac{1}{t} \ln \frac{\|\mathbf{X}_t\|}{\|\mathbf{X}_0\|} = \frac{1}{t}\sum_{k=0}^{t-1} \ln \|\mathbf{Z}_{k+1}\| - \frac{1}{t}\sum_{k=0}^{t-1} \ln \|\mathbf{Z}_k\| + \frac{1}{t}\sum_{k=0}^{t-1} \ln \eta^\star(\mathbf{Y}(\mathbf{Z}_t, \mathbf{U}_t)) \ .$$

We then apply Theorem 5.1 to each term of the RHS and find

$$\begin{aligned}\lim_{t\to\infty} \frac{1}{t} \ln \frac{\|\mathbf{X}_t\|}{\|\mathbf{X}_0\|} &= \int \ln \|\mathbf{z}\| \pi(d\mathbf{z}) - \int \ln \|\mathbf{z}\| \pi(d\mathbf{z}) + \int E[\ln \eta^\star(\mathbf{Y}(\mathbf{z},\mathbf{U})]\pi(d\mathbf{z}) \\ &= \int E[\ln \eta^\star(\mathbf{Y}(\mathbf{z},\mathbf{U})]\pi(d\mathbf{z}) = -\mathrm{CR} \ .\end{aligned}$$

Similarly since $\frac{1}{t} \ln \frac{\sigma_t}{\sigma_0} = \frac{1}{t}\sum_{k=0}^{t-1} \ln \eta^\star(\mathbf{Y}(\mathbf{Z}_t, \mathbf{U}_t))$, by applying Theorem 5.1, then $\lim_{t\to\infty} \frac{1}{t} \ln \frac{\sigma_t}{\sigma_0} = -\mathrm{CR}$. □

Positivity also guarantees convergence of $E_\mathbf{z}[h(\mathbf{Z}_t)]$ from "almost all" initial state $\mathbf{z}$ provided $\pi(|h|) < \infty$. More precisely from [17, Theorem 14.0.1] given a φ-irreducible and aperiodic chain $\mathbf{Z}$, for $h \geq 1$ a function on $\mathcal{Z}$, the following are equivalent: (i) The chain $\mathbf{Z}$ is positive (recurrent)[5] with invariant probability measure π and $\pi(h) := \int \pi(d\mathbf{z})h(\mathbf{z}) < \infty$. (ii) There exist some petite set C ([17, Section 5.5.2])

[5] Positive chains are recurrent according to Proposition 10.1.1 of [17] but the term positive recurrent is used to reinforce in the terminology the fact that they are recurrent (see [17] page 236).

and some extended-valued non-negative function V satisfying $V(\mathbf{z}_0) < \infty$ for some $\mathbf{z}_0$, and

$$\Delta V(\mathbf{z}) \leq -h(\mathbf{z}) + b1_C(\mathbf{z}), \quad \mathbf{z} \in \mathcal{Z}, \tag{5.9}$$

where Δ is the drift operator defined as

$$\Delta V(\mathbf{z}) = \int P(\mathbf{z}, d\mathbf{y})V(\mathbf{y}) - V(\mathbf{z}) = E_{\mathbf{z}}\left[V(\mathbf{Z}_1) - V(\mathbf{Z}_0)\right] \ . \tag{5.10}$$

Any of those two conditions imply that for any $\mathbf{z}$ in $S_V = \{\mathbf{z} : V(\mathbf{z}) < \infty\}$

$$\|P^t(\mathbf{z}, .) - \pi\|_h \xrightarrow[t\to\infty]{} 0 \ , \tag{5.11}$$

where $\|\nu\|_h := \sup_{g:|g|\leq h} |\nu(g)|$. Typically the function V will be finite everywhere such that the convergence in (5.11) will hold without any restrictions on the initial condition. The conditions (i) or (ii) for the chain $\mathbf{Z}$ with $h(\mathbf{z}) = |\ln \|\mathbf{z}\||+1$ imply the convergence in expectation of the log-progress independently of the starting point $\mathbf{z}$ taken into $S_V = \{\mathbf{z} : V(\mathbf{z}) < \infty\}$ where V is the function such that (5.9) is satisfied. More formally

THEOREM 5.3 (Linear convergence in expectation). *Let $(\mathbf{X}_t, \sigma_t)_{t\in\mathbb{N}}$ be the recursive sequence generated by a translation and scaling-invariant CB-SARS algorithm satisfying the assumptions of Proposition 4.1 optimizing a scaling-invariant function. Let $(\mathbf{Z}_t)_{t\in\mathbb{N}}$ be the homogeneous Markov chain defined in Proposition 4.1. Assume that $(\mathbf{Z}_t)_{t\in\mathbb{N}}$ is φ-irreducible and aperiodic and assume that either condition (i) or (ii) above are satisfied with $h(\mathbf{z}) = |\ln \|\mathbf{z}\||+1$. Assume also that there exists $\beta \geq 1$ such that*

$$\mathbf{y} \mapsto \mathcal{R}(\mathbf{y}) = \int \ln \eta^\star(\mathbf{Y}(\mathbf{y}, \mathbf{u}))p_{\mathbf{U}}(\mathbf{u})d\mathbf{u} \leq \beta(|\ln \|\mathbf{y}\||+1) \ . \tag{5.12}$$

Then for all initial condition $(\mathbf{X}_0, \sigma_0) = (\mathbf{x}, \sigma)$ such that $V(\mathbf{x}/\sigma) < \infty$ where V satisfies (5.9)

$$\lim_{t\to\infty} E_{\frac{\mathbf{x}}{\sigma}}\left[\ln \frac{\|\mathbf{X}_{t+1}\|}{\|\mathbf{X}_t\|}\right] = -\mathrm{CR} \text{ and } \lim_{t\to\infty} E_{\frac{\mathbf{x}}{\sigma}}\left[\ln \frac{\sigma_{t+1}}{\sigma_t}\right] = -\mathrm{CR} \ . \tag{5.13}$$

Proof. **Remark $\star$:** Note first that if (5.9) is satisfied for a function V for a given $h \geq 1$ then, for $\beta \geq 1$ the function βV will satisfy (5.9) for the function βh such that (5.11) will hold with βh.
Let us start by proving the RHS of (5.13) (we set $\mathbf{z} = \mathbf{x}/\sigma$)

$$\begin{aligned} E_{\frac{\mathbf{x}}{\sigma}}\left[\ln \frac{\sigma_{t+1}}{\sigma_t}\right] &= E_{\mathbf{z}}\left[\ln \eta^\star(\mathbf{Y}(\mathbf{Z}_t, \mathbf{U}_t))\right] \\ &= \int P^t(\mathbf{z}, d\mathbf{y}) \int \ln \eta^\star(\mathbf{Y}(\mathbf{y}, \mathbf{u}))p_{\mathbf{U}}(\mathbf{u})d\mathbf{u} = \int P^t(\mathbf{z}, d\mathbf{y})\mathcal{R}(\mathbf{y}) \ . \end{aligned}$$

Since $\mathcal{R}(\mathbf{y}) \leq \beta(|\ln \|\mathbf{y}\|| + 1)$ and $|\ln \|\mathbf{y}\|| + 1$ satisfies either (i) or (ii) we know from the remark $\star$ that $\lim_{t\to\infty} \|P^t(\mathbf{z}, .) - \pi\|_{\beta(\mathbf{y}\mapsto|\ln\|\mathbf{y}\||+1)} = 0$. Hence

$$|\int P^t(\mathbf{z}, d\mathbf{y})\mathcal{R}(\mathbf{y}) - \underbrace{\int \mathcal{R}(\mathbf{y})\pi(d\mathbf{y})}_{-\mathrm{CR}}| \leq \|P^t(\mathbf{z}, .) - \pi\|_{\beta(\mathbf{y}\mapsto|\ln\|\mathbf{y}\||+1)}$$

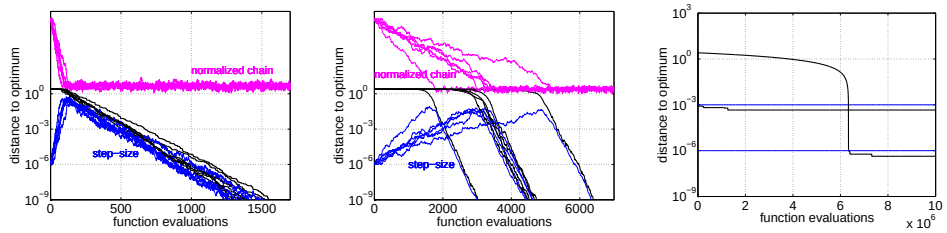

FIG. 5.1. *Convergence simulations on spherical functions* $f(\mathbf{x}) = g(\|\mathbf{x}\|)$ *for* $g \in \mathcal{M}$ *in dimension* $n = 10$. *Left: Simulation of the* $(1+1)$*-ES with one-fifth success rule (see Section 2.3.3, step-size update of (2.30) implemented with parameters* $p_{\text{target}} = 1/5$, $\kappa_\sigma = 1/3$ *were used)). Middle:* xNES*(see Section 2.3.1) using* $p = 4 + \lfloor 3 \ln n \rfloor$ *and* $\lfloor p/2 \rfloor$ *positive weights equals to* $w_i = \ln\left(\frac{\lambda}{2} + \frac{1}{2}\right) - \ln i$ *(default weights for the CMA-ES algorithm). Each plot is in log scale and depicts in black the distance to optimum, i.e.* $\|\mathbf{X}_t\|$, *in blue the respective step-size* σ_t *and in magenta the norm of the normalized chain* $\|\mathbf{Z}_t\|$. *The* x*-axis is the number of function evaluations corresponding thus to the iteration index* t *for the* $(1+1)$*-ES and to* $p \times t$ *for* xNES. *For both simulations* 6 *independent runs are conduced starting from* $\mathbf{X}_0 = (0.8, 0.8, \ldots, 0.8)$ *and* $\sigma_0 = 10^{-6}$. *Right: Simulation of a* $(1+1)$*-ES with constant step-size. Two runs conducted with a constant step-size equal to* 10^{-3} *and* 10^{-6}. *The distance to the optimum is depicted in black and the step-size in blue.*

converges to 0 when t goes to ∞ that proves the right limit in (5.13). To prove the left limit in (5.13), let us write

$$E_{\frac{\mathbf{x}}{\sigma}}\left[\ln \frac{\|\mathbf{X}_{t+1}\|}{\|\mathbf{X}_t\|}\right] = E_{\mathbf{z}}\left[\ln \frac{\eta^\star(\mathbf{Y}(\mathbf{Z}_t, \mathbf{U}_t))\|\mathbf{Z}_{t+1}\|}{\|\mathbf{Z}_t\|}\right]$$
$$= E_{\mathbf{z}}\left[\ln \eta^\star(\mathbf{Y}(\mathbf{Z}_t, \mathbf{U}_t))\right] + E_{\mathbf{z}}[\ln \|\mathbf{Z}_{t+1}\|] - E_{\mathbf{z}}[\ln \|\mathbf{Z}_t\|]$$

However $E_{\mathbf{z}}[\ln \|\mathbf{Z}_t\|] = \int P^t(\mathbf{z}, d\mathbf{y}) \ln \|\mathbf{y}\|$ that converges to $\int \ln \|\mathbf{y}\| \pi(d\mathbf{y})$ according to (5.11). This in turn implies that $E_{\mathbf{z}}[\ln \|\mathbf{Z}_{t+1}\|]$ converges to $\int \ln \|\mathbf{y}\| \pi(d\mathbf{y})$ and hence using the proven result for the right limit in (5.13), we obtain the left limit in (5.13).□

Stability like positivity and Harris-recurrence can be studied using drift conditions or Foster-Lyapunov criteria. A drift condition typically states that outside a set C, $\Delta V(\mathbf{z})$ is "negative". However "negativity" is declined in different forms. A drift condition for Harris recurrence of a φ-irreducible chain reads: if there exist a petite set C and a function V unbounded off petite sets such that

$$\Delta V(\mathbf{z}) \leq 0\,, \mathbf{z} \in C^c$$

holds, then the chain $\mathbf{Z}$ is Harris-recurrent [17, Theorem 9.1.8]. To ensure in addition positivity, a drift condition reads: if there exist a petite set C and V everywhere finite and bounded on C, a constant $b < \infty$ such that

$$\Delta V(\mathbf{z}) \leq -1 + b1_C(\mathbf{z}), \mathbf{z} \in \mathcal{Z}$$

holds, then $\mathbf{Z}$ is positive Harris-recurrent [17, Theorem 11.3.4].

Positivity and Harris-recurrence are typically proven using a stronger stability notion called *geometric ergodicity* [4, 3]. Geometric ergodicity characterizes that $P^t(\mathbf{z}, .)$ approaches the invariant probability measure π geometrically fast, at a rate $\rho < 1$ that is independent of the initial point $\mathbf{z}$. A drift condition for proving geometric ergodicity for a φ-irreducible and aperiodic chain reads: there exist a petite set C and constants $b < \infty$, $\beta > 0$ and a function $V \geq 1$ finite at some $\mathbf{z}_0 \in \mathcal{Z}$ satisfying

$$\Delta V(\mathbf{z}) \leq -\beta V(\mathbf{z}) + b1_C(\mathbf{z}), \mathbf{z} \in \mathcal{Z}\ . \tag{5.14}$$

This geometric drift condition implies that there exist constants $r > 1$ and $R < \infty$ such that for any starting point in the set $S_V = \{\mathbf{z} : V(\mathbf{z}) < \infty\}$

$$\sum_t r^t \|P^t(\mathbf{z}_0, .) - \pi\|_V \leq RV(\mathbf{z}_0) \tag{5.15}$$

where $\|\nu\|_V = \sup_{g:|g|\leq V} |\nu(g)|$ (see [17, Theorem 15.0.1]). This latter equation allows to have a stronger formulation for the linear convergence in expectation expressed in Theorem 5.3 as formalized in the next theorem.

THEOREM 5.4. *Assume that* $\mathbf{Z}$ *is geometrically ergodic satisfying a drift condition with* V *as drift function. Let* $g(\mathbf{z}) = E\left[\ln[\|\mathcal{G}_1((\mathbf{z}, 1), \mathbf{Y}(\mathbf{z}, \mathbf{U}))\|/\|\mathbf{z}\|]\right]$ *and assume that* $|g| \leq \beta V$ *with* $\beta \geq 1$. *Then, there exist* $r > 1$ *and* $R < \infty$ *such that for any starting point* $(\mathbf{x}_0, \sigma_0)$

$$\sum_t r^t |E_{\frac{\mathbf{x}_0}{\sigma_0}} \ln \frac{\|\mathbf{X}_{t+1}\|}{\|\mathbf{X}_t\|} - (-\mathrm{CR})| \leq RV\left(\frac{\mathbf{x}_0}{\sigma_0}\right) \tag{5.16}$$

In particular, for any initial condition $(\mathbf{x}_0, \sigma_0)$ $\lim_{t\to\infty} |E_{\frac{\mathbf{x}_0}{\sigma_0}} \ln \frac{\|\mathbf{X}_{t+1}\|}{\|\mathbf{X}_t\|} - (-\mathrm{CR})| r^t = 0$ *where* r *is independent of the starting point. Or also for any initial condition* $\left|E_{\frac{\mathbf{x}_0}{\sigma_0}} \ln \frac{\|\mathbf{X}_{t+1}\|}{\|\mathbf{X}_t\|} - (-\mathrm{CR})\right| \leq \frac{RV(\mathbf{x}_0/\sigma_0)}{r^t}$. *Let* $\tilde{g}(\mathbf{z}) = E[\ln \eta^\star(\mathbf{z}, \mathbf{Y}(\mathbf{z}, \mathbf{U})]$. *If* $\tilde{g} \leq \beta V$ *for* $\beta \geq 1$. *Then there exist* $r > 1$ *and* $R < \infty$ *such that for any starting point* $(\mathbf{x}_0, \sigma_0)$

$$\sum_t r^t |E_{\frac{\mathbf{x}_0}{\sigma_0}} \ln \frac{\sigma_{t+1}}{\sigma_t} - (-\mathrm{CR})| \leq RV\left(\frac{\mathbf{x}_0}{\sigma_0}\right) \tag{5.17}$$

In particular, for any initial condition $(\mathbf{x}_0, \sigma_0)$ $\lim_{t\to\infty} |E_{\frac{\mathbf{x}_0}{\sigma_0}} \ln \frac{\sigma_{t+1}}{\sigma_t} - (-\mathrm{CR})| r^t = 0$ *where* r *is independent of the starting point. Or also for any initial condition* $\left|E_{\frac{\mathbf{x}_0}{\sigma_0}} \ln \frac{\sigma_{t+1}}{\sigma_t} - (-\mathrm{CR})\right| \leq \frac{RV(\mathbf{x}_0/\sigma_0)}{r^t}$.

Proof. See [4, Theorem 4.8]. □

Geometric ergodicity is also a sufficient condition for the existence of a Central Limit Theorem (see [17, Theorem 7.0.1]) that can characterize how fast $\frac{1}{t} \ln \sigma_t/\sigma_0$ or $\frac{1}{t} \ln \|\mathbf{X}_t\|/\|\mathbf{X}_0\|$ approach the limit $-\mathrm{CR}$. We refer to [17, Theorem 4.10] for the details.

Interpretation and Illustration. Figure 5.1 illustrates the theoretical results formalized above. On the two leftmost plots, six single runs of the $(1 + 1)$-ES with one-fifth success rule and of the xNES algorithm optimizing spherical functions $f(\mathbf{x}) = g(\|\mathbf{x}\|)$ for $g \in \mathcal{M}$ in dimension $n = 10$ are depicted (see caption for parameters used). The evolution of $\|\mathbf{X}_t\|$, σ_t and $\|\mathbf{Z}_t\|$ are displayed using a logarithmic scale. In order to be able to compare the convergence rate between both algorithms, the x-axis represents the number of function evaluations and not the iteration index (however for the $(1 + 1)$-ES both number of function evaluations and iteration index coincide). The runs are voluntarily started with a too small step-size (equal to 10^6) compared to the distance to the optimum in order to illustrate the adaptivity property of both algorithms. For the $(1 + 1)$-ES, we observe a low variance in the results: after 100 function evaluations all the runs reach a well adapted step-size and the linear convergence is observed for both the step-size and the norm. The slope of the linear decrease observed coincides with $-\mathrm{CR}$ the convergence rate associated to the $(1 + 1)$-ES (up to a factor because a base 10 is used for the display). As theoretically stated $\ln \sigma_t$ and $\ln \|\mathbf{X}_t\|$ converge at the same rate (same slope for the curves).

The norm of the normalized chain $\mathbf{Z}_t$ is depicted in magenta, we observe that the stationary regime or steady-state of the chain correspond to the moment where linear convergence starts as predicted by the theory.

For the xNES algorithm, we observe the same behavior for each single run, i.e. a first phase where the adaptation of the step-size is taking place, here it means that the step-size is increased and a second phase where linear convergence is observed. In terms of normalized chain it corresponds to a first phase where a "transient behavior" is observed and a second phase where the distribution of the chain is close from the stationary distribution. We however see a larger variance in the time needed to reach the stationary state for the normalized chain, i.e. in the time to adapt the step-size that we believe is related to the variance of the log of the step-size change on the linear function. Using a cumulation mechanism like in the CSA algorithm reduces this variance [8]. The slope after reaching a reasonable step-size corresponds to the convergence rate CR multiplied by p (up to the difference with the base 10 logarithm). Both convergence rates between the $(1+1)$-ES and xNES are comparable while of course the number of function evaluations to reach 10^{-9} starting from a step-size of 10^{-6} is much longer for xNES as the adaptation phase is much slower for xNES than for the $(1+1)$-ES. It illustrates that only comparing the convergence rate (per function evaluation) can be misleading as it does not reflect the adaptation time.

Convergence of each single run reflects the almost-sure convergence property. Theoretically, the geometric ergodicity ensures that the adaptation phase is "short" as the Markov chain reaches its stationary state geometrically fast, i.e. we can start from a bad initial step-size, this bad choice will be fast corrected by the algorithm that will then converge linearly. In terms of the Markov chain $\mathbf{Z}_t$ the bad choice is translated as starting far away from the stationary distribution and the correction means reaching the stationary measure. We see however that in those "fast" statements the constants are omitted as for the xNES we observe that the step-size increase can take up to more than 3 times more function evaluations than for decreasing the step-size.

The rightmost plot in Figure 5.1 depicts the convergence of a non step-size adaptive strategy, here a $(1+1)$-ES with constant step-size equal to 10^{-3} and 10^{-6}. Theoretically the algorithm converges with probability one, at the same rate than the pure random search algorithm though. The plots illustrate the necessity of a step-size adaptive method: a wrong choice of the initial parameter has a huge effect in terms of time needed to reach a given target value. Indeed starting from a step-size of 10^{-3}, 1000 function evaluations are needed to reach a target of 10^{-6} while with a step-size of 10^{-6} roughly 6.2×10^{6} function evaluations are needed to reach the same target (i.e. more than 3 orders of magnitude more). Also we see that starting from a step-size of 10^{-3}, the number of function evaluations to reach a target of 10^{-6} will be beyond what is feasible to compute on a computer.

This rightmost plot also illustrates the importance to study theoretically convergence rates, as convergence with probability one can be associated to an algorithm having very poor performance for practical purposes.

6. Discussion. This paper provides a general methodology to prove global linear convergence of some comparison-based step-size adaptive randomized search algorithms on scaling-invariant functions, a class of functions that includes in particular *non quasi-convex* and *non continuous* functions. The methodology exploits the invariance properties of the algorithms and turns the question of global linear convergence into the study of the stability of an underlying homogeneous normalized Markov chain. It generalizes previous works [6, 3] to a broader class of functions and a broader class

of algorithms.

Different notions of stability for a Markov chain exist. They imply different (non equivalent) formulations of linear convergence that give many insights on the dynamic of the algorithm: Positivity and Harris recurrence essentially imply the existence of a convergence rate CR such that for any initial state almost surely

$$\lim_{t\to\infty} \frac{1}{t} \ln \frac{\|\mathbf{X}_t\|}{\|\mathbf{X}_0\|} = -\mathrm{CR} \ ; \lim_{t\to\infty} \frac{1}{t} \ln \frac{\sigma_t}{\sigma_0} = -\mathrm{CR}$$

holds. Positivity essentially implies convergence in expectation. More precisely for any initial state $\mathbf{X}_0 = x, \sigma_0 = \sigma$

$$\lim_{t\to\infty} E_{\substack{x \\ \sigma}} \left[\ln \frac{\|\mathbf{X}_{t+1}\|}{\|\mathbf{X}_t\|} \right] = -\mathrm{CR} \ ; \lim_{t\to\infty} E_{\substack{x \\ \sigma}} \left[\ln \frac{\sigma_{t+1}}{\sigma_t} \right] = -\mathrm{CR} \ .$$

Geometric ergodicity then characterizes that the expected log-progress sequence converges geometrically fast to the convergence rate limit $-\mathrm{CR}$.

Linear convergence holds under any initial condition. This reflects the practical adaptivity property: the step-size parameter is adjusted on the fly and hence a bad choice of an initial parameter is not problematic. We have illustrated that the transition phase, formally how long it takes to be close to the invariant probability measure, relates to how long it takes to forget a bad initialization.

The methodology provides an exact formula for the convergence rate CR expressed in terms of expectation w.r.t. the invariant probability measure of the normalized Markov chain. Exploiting the exact expression for deducing properties on the convergence rate like dependency w.r.t. the dimension or dependency on function properties (like condition number of the hessian matrix if the function is convex quadratic) seems however to be quite challenging with this approach while it is feasible with ad-hoc techniques for specific algorithms (see [14]). Numerical simulations need then to be performed to investigate those properties. Nevertheless the Markov chain methodology proposed here provides a rigorous framework for performing those simulations: it proves that by essence Monte-Carlo simulation of the convergence rate is consistent and even provides through the Central Limit Theorem asymptotic confidence intervals for the simulations.

We have restricted for the sake of simplicity the CB-SARS framework to the update of a mean vector and a step-size. However some step-size adaptive algorithms like the cumulated step-size adaptation used in the CMA-ES algorithm include other state variables like an auxiliary vector (the path) used to update the step-size [11]. Adaptation of the present methodology to cases with more state variables seems however relatively straightforward.

The current approach exploits heavily invariance properties of the algorithms investigated together with invariance properties of the objective function. Hence, we expect that the methodology does not generalize directly to any unimodal function. However we believe that there is room for extension of the framework in some noisy context for instance (i.e. the objective function is stochastic).

Another possible approach to analyze the linear convergence of a comparison-based step-size adaptive randomized search consists in using stochastic approximation theory or the *method of ordinary differential equations* [16, 7]. We believe that linear convergence can then be proven on different function classes at the price of the assumption that the learning rates (κ_m, κ_σ in (2.22) and (2.25) for instance) are small

enough. A step needed in this analysis is the investigation of an ordinary differential equation obtained by suitable averaging. We believe that this can be done by extending results presented in [2].

REFERENCES

[1] Y. Akimoto, Y. Nagata, I. Ono, and S. Kobayashi. Bidirectional relation between CMA evolution strategies and natural evolution strategies. volume 6238 of *Lecture Notes in Computer Science*, pages 154–163. Springer Verlag, 2010.

[2] Youhei Akimoto, Anne Auger, and Nikolaus Hansen. Convergence of the continuous time trajectories of isotropic evolution strategies on monotonic $\mathcal{C}^2$-composite functions. In Carlos A. Coello Coello, Vincenzo Cutello, Kalyanmoy Deb, Stephanie Forrest, Giuseppe Nicosia, and Mario Pavone, editors, *Parallel Problem Solving from Nature - PPSN XII*, volume 7491 of *Lecture Notes in Computer Science*, pages 42–51. Springer, 2012.

[3] A. 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.

[4] A. Auger and N. Hansen. Linear convergence on positively homogeneous functions of a comparison based step-size adaptive randomized search: the (1+1) ES with generalized one-fifth success rule, 2013. ArXiv eprint.

[5] H.-G. Beyer and H.-P. Schwefel. Evolution strategies — a comprehensive introduction. *Natural computing*, 1(1):3–52, 2002.

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

[7] Vivek S Borkar. Stochastic approximation: a dynamical systems viewpoint. Cambridge University Press, 2008.

[8] A. Chotard, A. Auger, and N. Hansen. Cumulative step-size adaptation on linear functions. In *Parallel Problem Solving from Nature - PPSN XII*, pages 72–81. Springer, 2012.

[9] L. Devroye. The compound random search. In *International Symposium on Systems Engineering and Analysis*, pages 195–110. Purdue University, 1972.

[10] T. Glasmachers, T. Schaul, Y. Sun, D. Wierstra, and J. Schmidhuber. Exponential natural evolution strategies. In *Genetic and Evolutionary Computation Conference (GECCO 2010)*, pages 393–400. ACM Press, 2010.

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

[12] N. Hansen, A. Ostermeier, and A. Gawelczyk. On the adaptation of arbitrary normal mutation distributions in evolution strategies: The generating set adaptation. pages 57–64. Morgan Kaufmann, 1995.

[13] R. Hooke and T.A. Jeeves. "Direct Search" Solution of Numerical and Statistical Problems. *Journal of the ACM*, 8:212–229, 1961.

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

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

[16] Harold J. Kushner and G. George Yin. *Stochastic approximation and recursive algorithms and applications*. Springer Verlag, 2nd edition, 2003.

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

[18] John Ashworth Nelder and R Mead. A simplex method for function minimization. *The Computer Journal*, pages 308–313, 1965.

[19] Yurii Nesterov. Random gradient-free minimization of convex functions. CORE Discussion Papers 2011001, Universit catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2011.

[20] Y. Ollivier, L. Arnold, A. Auger, and N. Hansen. Information-geometric optimization algorithms: A unifying picture via invariance principles. *ArXiv e-prints*, June 2013.

[21] I. Rechenberg. *Evolutionstrategie: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution*. Frommann-Holzboog Verlag, Stuttgart, 1973.

[22] I. Rechenberg. *Evolutionsstrategie '94*. Frommann-Holzboog Verlag, 1994.

[23] M. Schumer and K. Steiglitz. Adaptive step size random search. *IEEE Transactions on Automatic Control*, 13(3):270–276, 1968.

[24] H.-P. Schwefel. *Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie*. Birkhäuser, 1977.
[25] S. Stich, C. Müller, and B. Gärtner. Optimization of convex functions with random pursuit. *SIAM Journal on Optimization*, 23(2):1284–1309, 2013.
[26] Anatoly A. Zhigljavsky and Antanas Zilinskas. *Stochastic global optimization*, volume 1 of *Springer Optimization and its applications*. Springer, 2008.