Journal of Global Optimization manuscript No.
(will be inserted by the editor)

Global linear convergence of Evolution Strategies with recombination on scaling-invariant functions

Cheikh Toure · Anne Auger · Nikolaus Hansen

Received: date / Accepted: date

Abstract Evolution Strategies (ESs) are stochastic derivative-free optimization algorithms whose most prominent representative, the CMA-ES algorithm, is widely used to solve difficult numerical optimization problems. We provide the first rigorous investigation of the linear convergence of step-size adaptive ESs involving a population and recombination, two ingredients crucially important in practice to be robust to local irregularities or multimodality.

We investigate the convergence of step-size adaptive ESs with weighted recombination on composites of strictly increasing functions with continuously differentiable scaling-invariant functions with a global optimum. This function class includes functions with non-convex sublevel sets and discontinuous functions. We prove the existence of a constant r such that the logarithm of the distance to the optimum divided by the number of iterations converges to r. The constant is given as an expectation with respect to the stationary distribution of a Markov chain—its sign allows to infer linear convergence or divergence of the ES and is found numerically.

Our main condition for convergence is the increase of the expected log step-size on linear functions. In contrast to previous results, our condition is equivalent to the almost sure geometric divergence of the step-size on linear functions.

Keywords Evolution Strategies; Linear Convergence; CMA-ES; Scaling-invariant functions; Foster-Lyapunov drift conditions.

1 Introduction

Evolution Strategies (ESs) are stochastic numerical optimization algorithms introduced in the 70's [54, 55, 60, 61]. They aim at optimizing an objective function $f : \mathbb{R}^n \to \mathbb{R}$ in a so-called zero-order black-box scenario where gradients are not available and only *comparisons* between f-values of candidate solutions are used to update the state of the algorithm. ESs sample candidate solutions from a multivariate normal distribution parametrized by a mean vector and a covariance matrix.

Inria and CMAP, Ecole Polytechnique, IP Paris, France
cheikh.toure@polytechnique.edu
firstname.lastname@inria.fr

The mean vector represents the incumbent or current favorite solution while the covariance matrix determines the geometric shape of the sampling probability distribution. In adaptive ESs, not only the mean vector but also a step-size or the covariance matrix is adapted in each iteration. Covariance matrices can be seen as encoding a metric such that Evolution Strategies that adapt a full covariance matrix are variable metric algorithms [64].

In the domain of Evolutionary Computation, the covariance-matrix-adaptation ES (CMA-ES) [29, 36] is nowadays recognized as state-of-the-art to solve difficult numerical optimization problems that can typically be non-convex, non-linear, ill-conditioned, non-separable, rugged or multi-modal[1] [18, 23, 24, 32] [56, Fig. 20]. Other relevant algorithms to solve ill-structured, non-convex, multi-modal, non-differentiable problems are also often population based like Estimation of Distribution algorithms notably AMaLGaM [17], Differential Evolution [22, 62], and Particle Swarm Optimization (PSO) [47]. PSO methods however exploit separability and are inefficient to solve non-separable ill-conditioned problems [37]. The CMA-ES algorithm is based upon several maximum likelihood updates [31], can be interpreted as a natural gradient descent [5, 25, 53] and has been tightly linked to the EM-algorithm [7]. Adaptation of the full covariance matrix is crucial to solve general ill-conditioned, non-separable problems. Up to a multiplicative factor that converges to zero, the covariance matrix in CMA-ES becomes on strictly convex quadratic objective functions close to the inverse Hessian of the function [27].

The CMA-ES algorithm follows a $(\mu/\mu_w, \lambda)$-ES algorithmic scheme where from the offspring population of λ candidate solutions sampled at each iteration, the $\mu \approx \lambda/2$ best solutions—the new parent population—are recombined as a weighted sum to define the new mean vector of the multivariate normal distribution. On a unimodal spherical function, the optimal step-size, i.e. the standard deviation that should be used to sample each coordinate of the candidate solutions, depends monotonously on μ [55]. Hence, increasing the population size makes the search less local while preserving a close-to-optimal convergence rate per function evaluation as long as λ remains moderately large [8, 9, 30]. This remarkable theoretical property implies robustness and partly explains why on many multi-modal test functions increasing λ empirically increases the probability to converge to the global optimum [34]. The robustness when increasing λ and the inherent parallel nature of ESs are two key features behind their success for tackling difficult black-box optimization problems.

Convergence is a central question in optimization. For comparison-based algorithms like ESs, linear convergence (where the distance to the optimum decreases geometrically) is the fastest possible convergence [43, 65]. Gradient methods also converge linearly on strongly convex functions [52, Theorem 2.1.15]. We have ample empirical evidence that adaptive ESs converge linearly on wide classes of functions [30, 37, 38, 57]. Yet, establishing proofs is known to be difficult. So far, linear convergence could be proven only for step-size adaptive algorithms where the covariance matrix equals a scalar times the identity [11, 13, 39–42] or a scalar

[1] The `cmaes` and the `pycma` Python modules that implement the algorithm are downloaded more than 300,000 and 30,000 times per week, respectively, from PyPI as of September 2022. Both modules implement the main ideas of CMA-ES [36] and further enhancements published over the years, notably the rank-μ update [35], a better setting for step-size damping and the weights [34], an active covariance matrix update [44], and restart mechanisms with increasing population size [12, 28].

times a covariance matrix with eigenvalues upper bounded and bounded away from zero [2]. In addition, these proofs require the parent population size to be one.

In this context, we analyze here for the first time the linear convergence of a step-size adaptive ES with a parent population size greater than one and recombination, following a $(\mu/\mu_w, \lambda)$-ES framework. As a second novelty, we model the step-size update by a generic function and thereby also encompass the step-size updates in the CMA-ES algorithm [29] (however with a specific parameter setting which leads to a reduced state-space) and in the xNES algorithm [25].

Our proofs hold on composites of strictly increasing functions with either continuously differentiable scaling-invariant functions with a unique argmin or nontrivial linear functions. This class of functions includes discontinuous functions, functions with infinite many critical points, and functions with non-convex sublevel sets. It does not include functions with more than one (local or global) optimum.

In this paper, we use a methodology based on analyzing the stochastic process defined as the difference between the mean vector and a reference point (often the optimum of the problem), normalized by the step-size [14]. This construct is a viable model of the underlying (translation and scale-invariant) algorithm when optimizing scaling-invariant functions, in which case the stochastic process is also a Markov chain and here referred to as *σ-normalized Markov chain*. This chain is *homogeneous* as a consequence of three crucial invariance properties of the ES algorithms: translation invariance, scale invariance, and invariance to strictly increasing transformations of the objective function. Proving *stability* of the σ-normalized Markov chain (φ-irreducibility, Harris recurrence, positivity) is key to obtain almost sure *linear behavior* of the algorithm [14]. The sign and value of the convergence or divergence rate can however only be obtained from elementary Monte Carlo simulations. The technically challenging part in the proof methodology is the stability analysis. It was not carried out by Auger and Hansen [14] who presented the methodology and some algorithm classes that can be addressed by the methodology but assumed stability of the algorithms without proof. We prove in the following the stability for some algorithms belonging to the $(\mu/\mu_w, \lambda)$-ES framework and thus formally prove linear behavior of these algorithms.

Relation to previous works: In contrast to our study, most theoretical analyses of linear convergence concern the so-called (1+1)-ES where a single candidate solution is sampled ($\lambda = 1$) and the new mean is the best among the current mean and the sampled solution and in addition the one-fifth success rule is used to adapt the step-size [48, 54]. Jägersküpper established lower-bounds and upper-bounds on the hitting time to reduce the distance to the optimum related to linear convergence on spherical functions [39, 42] and on some convex-quadratic functions [40, 41]. Remarkably, these studies derive the dependency of the hitting time bounds on dimension and condition number, an aspect which is not covered with our approach. The underlying methodology used for the proofs was later unveiled as connected to drift analysis where an overall Lyapunov function of the state of the algorithm (mean and step-size) is used to prove upper and lower bounds on the hitting time of an epsilon neighborhood of the optimum [1]. With this drift analysis, Akimoto et al. [1] provide lower and upper bounds on the hitting time of an ϵ-ball pertaining to linear convergence (coming as well with dependency in the dimension) for the the (1+1)-ES with one-fifth success rule on spherical functions. The analysis was

later generalized for classes of functions including strongly convex functions with Lipschitz gradient as well as positively homogeneous functions [2, 51].

Using the same methodology as in this paper, the linear convergence of the $(1+1)$-ES with step-size adapted via the one-fifth success rule is proven on increasing transformations of C^1 positively homogeneous functions p with a unique global argmin and upper bounds on the degree of p and on the norm of the gradient $\|\nabla p\|$ [13].

While most theoretical studies of linear convergence concern a (1+1)-ES, the $(1, \lambda)$-ES with self-adaptation has been analyzed on the sphere function [11] and more recently an ODE method has been developed and applied to a $(\mu/\mu, \lambda)$-ES with a specific step-size adaptation concluding linear convergence on the sphere function when the learning rate is small enough [4]. Our analysis holds for wider classes of functions and does not impose a small learning rate. However it does not allow to obtain the sign of the convergence rate.

A few studies attempt to analyze ESs with covariance matrix adaptation: A variant of CMA-ES, modified to ensure a sufficient decrease, globally converges (but not provably linearly) [21]. Provided the eigenvalues of the covariance matrix stay upper bounded and bounded away from zero (which is not the case in the affine-invariant CMA-ES), a (1+1)-CMA-ES with any covariance matrix update and proper step-size adaptation converges linearly [2]. When convergence occurs on a twice continuously differentiable function for CMA-ES without step-size adaptation, the limit point is a local (or global) optimum [6].

This paper is organized as follows. We present in Section 2 the algorithm framework, the assumptions on the algorithm and the class of objective functions where the convergence analysis is carried out. In Section 3 we present the main proof idea to prove a linear behavior and present the ensuing proof structure. In Section 4, we present different Markov chain notions and tools needed for our analysis. In Section 5, we establish different stability properties on the σ-normalized Markov chain. We state and prove the main results in Section 6. Notations are summarized in Table 1.

2 Algorithm framework and class of functions studied

We present in this section the step-size adaptive algorithm framework analyzed, the assumptions on the algorithm and the function class considered as well as preliminary results. In the following, we consider an abstract measurable space $(\Omega, \mathcal{F})$ and a probability measure P so that $(\Omega, \mathcal{F}, P)$ is a measure space.

2.1 The $(\mu/\mu_w, \lambda)$-ES algorithm framework

We introduce step-size adaptive ESs with recombination, referred to as step-size adaptive $(\mu/\mu_w, \lambda)$-ES. Given a positive integer n and a function $f : \mathbb{R}^n \to \mathbb{R}$ to be minimized, the sequence of states of the algorithm is represented by $\{(X_k, \sigma_k)\,; k \in \mathbb{N}\}$ where at iteration k, $X_k \in \mathbb{R}^n$ is the incumbent (the favorite solution considered as current estimate of the optimum) and the positive scalar σ_k is the step-size. We fix positive integers λ and μ such that $\mu \leq \lambda$.

Table 1: Notations

$[\cdot]_i$	is the i^{th} vector of a sequence of vectors
$\|\cdot\|$	is the Euclidean norm
$\|\cdot\|_\infty$	is the infinity norm on a space of bounded functions
$\|\nu\|_h$	$= \sup_{\|g\|\le h} \|\mathbb{E}_\nu(g)\|$ is for a positive function h the norm of the signed measure ν
$\pi_1 \times \pi_2$	is the product measure from two measure spaces $(\mathcal{Z}_i, \mathcal{B}(\mathcal{Z}_i), \pi_i)$, $i = 1, 2$, on the product measurable space $(\mathcal{Z}_1 \times \mathcal{Z}_2,\ \mathcal{B}(\mathcal{Z}_1) \otimes \mathcal{B}(\mathcal{Z}_2))$ where $\otimes$ is the tensor product
A^c	is the complement of a set A
$A^\top$	is the transpose of a matrix A
$\mathbf{B}(x, \rho)$	$= \{y \in \mathbb{R}^n; \|x - y\| < \rho\}$ is the open ball around $x \in \mathbb{R}^n$ with radius $\rho > 0$ and $\overline{\mathbf{B}(x, \rho)}$ is its closure
$\mathcal{B}(\mathcal{Z})$	is the Borel sigma-field of the topological space $\mathcal{Z}$
$\mathbb{E}_\nu(g)$	$= \int g(z)\, \nu(\mathrm{d}z)$ for any real-valued function g and a signed measure ν
$\mathcal{L}_{f,z}$	$= \{y \in \mathbb{R}^n\,;\, f(y) = f(z)\}$ is the level set for an objective function $f : \mathbb{R}^n \to \mathbb{R}$ and an element $z \in \mathbb{R}^n$
$\mathbb{N}$	is the set of non-negative integers
$\mathcal{N}$	is the standard normal distribution
$\mathcal{N}_m$	is the standard multivariate normal distribution in dimension m
$\mathcal{N}(x, C)$	is the multivariate normal distribution with mean $x \in \mathbb{R}^m$ and covariance matrix C
$p_{\mathcal{N}_m}$	is the probability density function of $\mathcal{N}_m$
$\mathbb{R}_+$	is the set of non-negative real numbers
u	$= (u^1, \ldots, u^m) \in \mathbb{R}^{pm}$ where $u^i \in \mathbb{R}^p$ for $i = 1, \ldots, m$ and $p \in \mathbb{N} \setminus \{0\}$ and we write $u = (u^1) = u^1$ if $m = 1$
$w^\top u$	$= \sum_{i=1}^m w_i u^i$ for $w \in \mathbb{R}^m$ and $u \in \mathbb{R}^{pm}$

Let $(X_0, \sigma_0) \in \mathbb{R}^n \times (0, \infty)$ and $U = \{U_{k+1} = (U_{k+1}^1, \ldots, U_{k+1}^\lambda)\,; k \in \mathbb{N}\}$ be a sequence of independent and identically distributed (i.i.d.) random inputs independent from (X_0, σ_0), where for all $k \in \mathbb{N}$, $U_{k+1} = (U_{k+1}^1, \ldots, U_{k+1}^\lambda)$ is composed of λ independent random vectors following a standard multivariate normal distribution $\mathcal{N}_n$. Given (X_k, σ_k) for $k \in \mathbb{N}$, we consider the following iterative update. First, we define λ candidate solutions as

$$X_{k+1}^i = X_k + \sigma_k\, U_{k+1}^i \quad \text{for } i = 1, \ldots, \lambda. \tag{1}$$

Second, we evaluate the candidate solutions on f. We then denote an f-sorted permutation of $\left(X_{k+1}^1, \ldots, X_{k+1}^\lambda\right)$ as $\left(X_{k+1}^{1:\lambda}, \ldots, X_{k+1}^{\lambda:\lambda}\right)$ such that

$$f(X_{k+1}^{1:\lambda}) \le \cdots \le f(X_{k+1}^{\lambda:\lambda}) \tag{2}$$

and thereby define the indices $i\!:\!\lambda$. To break possible ties, we require that $i\!:\!\lambda < j\!:\!\lambda$ if $f(X_{k+1}^i) = f(X_{k+1}^j)$ and $i < j$. The sorting indices $i\!:\!\lambda$ are also used for the σ-normalized difference vectors U_{k+1}^i in that $U_{k+1}^{i:\lambda} = \frac{X_{k+1}^{i:\lambda} - X_k}{\sigma_k}$. Accordingly, we define the *selection function* α_f of $z \in \mathbb{R}^n$ and $u = (u^1, \ldots, u^\lambda) \in \mathbb{R}^{n\lambda}$ to yield the sorted sequence of the difference vectors as

$$\alpha_f(z, u) = (u^{1:\lambda}, \ldots, u^{\mu:\lambda}) \in \mathbb{R}^{n\mu}, \tag{3}$$

with $f(z + u^{1:\lambda}) \le \cdots \le f(z + u^{\lambda:\lambda})$ and the above tie breaking. For $\lambda = 2$ and $\mu = 1$, the selection function has the simple expression $\alpha_f(z, (u^1, u^2)) =$

$(u^1 - u^2)\mathbb{1}_{\{f(z+u^1)\leq f(z+u^2)\}} + u^2$. By definition, for $k \in \mathbb{N}$, $\alpha_f(X_k, \sigma_k U_{k+1}) = \left(\sigma_k U_{k+1}^{1:\lambda}, \ldots, \sigma_k U_{k+1}^{\mu:\lambda}\right)$ so that

$$\frac{\alpha_f(X_k, \sigma_k U_{k+1})}{\sigma_k} = \left(U_{k+1}^{1:\lambda}, \ldots, U_{k+1}^{\mu:\lambda}\right). \tag{4}$$

However, α_f is not a homogeneous function in general, because the indices $i:\lambda$ in (4) depend on f and hence on α_f and hence on σ_k.

The update of the state of the algorithm uses the objective function only through the above selection function which is invariant to strictly increasing transformations of the objective function. Indeed, the selection is determined through the ranking of candidate solutions in (2) which is the same when on $g \circ f$ or f given that g is strictly increasing. We talk about comparison-based algorithms. Formally:

Lemma 1 *Let $f = \varphi \circ g$ where $g : \mathbb{R}^n \to \mathbb{R}$ and φ is strictly increasing. Then $\alpha_f = \alpha_g$.*

To update the mean vector X_k, we consider a weighted average of the $\mu \leq \lambda$ best solutions $\sum_{i=1}^{\mu} w_i X_{k+1}^{i:\lambda}$ where $w = (w_1, \ldots, w_\mu)$ is a non-zero vector. When only positive weights summing to one are used, this weighted average is situated in the convex hull of the μ best points. The next incumbent X_{k+1} is constructed by combining X_k and $\sum_{i=1}^{\mu} w_i X_{k+1}^{i:\lambda}$

$$X_{k+1} = \left(1 - \sum_{i=1}^{\mu} w_i\right) X_k + \sum_{i=1}^{\mu} w_i X_{k+1}^{i:\lambda} = X_k + \sigma_k \sum_{i=1}^{\mu} w_i U_{k+1}^{i:\lambda} \ . \tag{5}$$

Positive weights with small indices move the new mean vector towards the better solutions, hence these weights should generally be large. In ESs, the weights are always non-increasing in i. With the notable exception of Natural Evolution Strategies ([25] and related works), all weights are positive. In practice, $\sum_{i=1}^{\mu} w_i$ is often set to 1 such that the new mean vector is the weighted average of the μ best solutions. Proposition 12 describes (generally weak) explicit conditions for the weights under which our results hold. We write the step-size update in an abstract manner as

$$\sigma_{k+1} = \sigma_k \, \Gamma\left(U_{k+1}^{1:\lambda}, \ldots, U_{k+1}^{\mu:\lambda}\right) \tag{6}$$

where $\Gamma : \mathbb{R}^{n\mu} \to \mathbb{R}_+ \setminus \{0\}$ is a measurable function. This generic step-size update is by construction scale-invariant, which is key for our analysis [14, Proposition 2.9]. The update of the mean vector and of the step-size are both functions of the f-sorted sampled vectors $(U_{k+1}^{1:\lambda}, \ldots, U_{k+1}^{\mu:\lambda})$.

Using (4), we rewrite the algorithm framework (5) and (6) for all k as:

$$X_{k+1} = X_k + \sum_{i=1}^{\mu} w_i \left[\alpha_f(X_k, \sigma_k U_{k+1})\right]_i = X_k + w^\top \alpha_f(X_k, \sigma_k U_{k+1}) \tag{7}$$

$$\sigma_{k+1} = \sigma_k \, \Gamma\left(\frac{\alpha_f(X_k, \sigma_k U_{k+1})}{\sigma_k}\right) \tag{8}$$

with $U = \{U_{k+1} \,; k \in \mathbb{N}\}$ the sequence of identically distributed random inputs and $w \in \mathbb{R}^\mu \setminus \{0\}$. In (7), we use the notation $[\ ]_i$ to denote the i^{th} vector of the μ vectors composing $\alpha_f(X_k, \sigma_k U_{k+1})$.

2.2 Algorithms encompassed

The generic update in (6) or equivalently (8) encompasses the step-size update of the cumulative step-size adaptation evolution strategy ($(\mu/\mu_w, \lambda)$-CSA-ES) [14,36] with cumulation factor set to 1 where for $d_\sigma > 0$, $w \in \mathbb{R}^\mu \backslash \{0\}$ and $u = (u^1, \ldots, u^\mu) \in \mathbb{R}^{n\mu}$,

$$\Gamma^0_{\mathrm{CSA1}}(u^1, \ldots, u^\mu) = \exp\left(\frac{1}{d_\sigma}\left(\frac{\|\sum_{i=1}^\mu w_i u^i\|}{\|w\|\,\mathbb{E}\left[\|\mathcal{N}_n\|\right]} - 1\right)\right) \quad . \tag{9}$$

The acronym CSA1 emphasizes that we only consider a particular case here: in the original CSA algorithm, the sum in (9) is an exponentially fading average of these sums from the past iterations with a smoothing factor of $1 - c_\sigma$. Equation (9) only holds when the cumulation factor c_σ is equal to 1, whereas in practice, $1/c_\sigma$ is between $\sqrt{n}/2$ and $n + 2$ (see [29] for more details). The damping parameter $d_\sigma \approx 1$ scales the change magnitude of $\log(\sigma_k)$.

Equation (9) increases the step-size if and only if the length of $\sum_{i=1}^\mu w_i U^{i:\lambda}_{k+1}$ is larger than the expected length of $\sum_{i=1}^\mu w_i U^i_{k+1}$ under random selection which is equal to $\|w\|\,\mathbb{E}\left[\|\mathcal{N}_n\|\right]$. Since the function Γ^0_{CSA1} is not continuously differentiable (an assumption needed in our analysis) we consider a version of the $(\mu/\mu_w, \lambda)$-CSA1-ES [10] that compares the square length of $\sum_{i=1}^\mu w_i U^{i:\lambda}_{k+1}$ to the expected square length of $\sum_{i=1}^\mu w_i U^i_{k+1}$ which is $n\|w\|^2$. Hence, the step-size update we consider and that satisfies our assumptions is defined for $d_\sigma > 0$, $w \in \mathbb{R}^\mu \setminus \{0\}$ and $u = (u^1, \ldots, u^\mu) \in \mathbb{R}^{n\mu}$ as

$$\Gamma_{\mathrm{CSA1}}(u^1, \ldots, u^\mu) = \exp\left(\frac{1}{2d_\sigma n}\left(\frac{\|\sum_{i=1}^\mu w_i u^i\|^2}{\|w\|^2} - n\right)\right). \tag{10}$$

Another step-size update encompassed with (4) is given by the Exponential Natural Evolution Strategy (xNES) [14, 25, 53, 58] and defined for $d_\sigma > 0$, $w \in \mathbb{R}^\mu \setminus \{0\}$ and $u = (u^1, \ldots, u^\mu) \in \mathbb{R}^{n\mu}$ as

$$\Gamma_{\mathrm{xNES}}(u^1, \ldots, u^\mu) = \exp\left(\frac{1}{2d_\sigma n}\left(\sum_{i=1}^\mu \frac{w_i}{\sum_{j=1}^\mu |w_j|}\left(\|u^i\|^2 - n\right)\right)\right). \tag{11}$$

Both equations (10) and (11) correlate the step-size increment with the vector lengths of the μ best solutions. While (10) takes the squared norm of the weighted sum of the vectors, (11) takes the weighted sum of squared norms. Hence, correlations between the directions u^i affect only (10). Both equations are offset to become unbiased such that $\log \circ \Gamma$ is zero in expectation when $u^i \sim \mathcal{N}_n$ for all $1 \le i \le \lambda$, are i.i.d. random vectors.

2.3 Assumptions on the algorithm framework

We pose some assumptions on the algorithm (7) and (8) starting with assumptions on the step-size update function Γ.

A1. The function $\Gamma : \mathbb{R}^{n\mu} \to \mathbb{R}_+ \backslash \{0\}$ is continuously differentiable (C^1).
A2. Γ is *invariant under rotation* in the following sense: for all $n \times n$ orthogonal matrices T, for all $u = (u_1, \ldots, u_\mu) \in \mathbb{R}^{n\mu}$, $\Gamma(Tu_1, \ldots, Tu_\mu) = \Gamma(u)$.

A3. The function Γ is lower-bounded by a constant $m_\Gamma > 0$, that is for all $x \in \mathbb{R}^{n\mu}$, $\Gamma(x) \geq m_\Gamma$.

A4. $\log \circ \Gamma$ is $\mathcal{N}_{n\mu}$-integrable, that is, $\int |\log(\Gamma(u))| \, p_{\mathcal{N}_{n\mu}}(u) \mathrm{d}u < \infty$.

We can easily verify that Assumptions A1–A4 are satisfied for the $(\mu/\mu_w, \lambda)$-CSA1 and $(\mu/\mu_w, \lambda)$-xNES updates given in (10) and (11). More precisely, the following lemma holds.

Lemma 2 *The step-size update function Γ_{CSA1} defined in (10) satisfies Assumptions A1−A4. Endowed with non-negative weights $w_i \geq 0$ for all $i = 1, \ldots, \mu$, the step-size update function Γ_{xNES} defined in (11) satisfies Assumptions A1−A4.*

Proof. A1 and A4 are immediate to verify. For A2, the invariance under rotation comes from the norm-preserving property of orthogonal matrices. For all $u = \left(u^1, \ldots, u^\mu\right) \in \mathbb{R}^{n\mu}$, $\Gamma_{\mathrm{CSA1}}(u) \geq \exp\left(-\frac{1}{2d_\sigma}\right)$ such that Γ_{CSA1} satisfies A3. Similarly $\Gamma_{\mathrm{xNES}}(u) = \exp\left(-\frac{1}{2d_\sigma}\frac{\sum_{i=1}^\mu w_i}{\sum_{j=1}^\mu |w_j|} + \frac{1}{2d_\sigma n}\sum_{i=1}^\mu \frac{w_i}{\sum_{j=1}^\mu |w_j|}\|u^i\|^2\right)$. Since all the weights are non-negative, $\frac{1}{2d_\sigma n}\sum_{i=1}^\mu w_i\|u^i\|^2 \geq 0$. And then $-\frac{1}{2d_\sigma}\sum_{i=1}^\mu w_i + \frac{1}{2d_\sigma n}\sum_{i=1}^\mu w_i\|u^i\|^2 \geq -\frac{1}{2d_\sigma}\sum_{i=1}^\mu w_i$. Therefore $\Gamma_{\mathrm{xNES}}(u) \geq \exp\left(-\frac{1}{2d_\sigma}\right)$ which does not depend on u, such that Γ_{xNES} satisfies A3. □

Assumptions A1–A4 are also satisfied for a constant function Γ equal to a positive number. When the positive number is greater than 1, our main condition for a linear behavior is satisfied, as we will see later on. Yet, the step-size of this algorithm clearly diverges geometrically.

We formalize now the assumption on the source distribution used to sample candidate solutions, as it was already specified when defining the algorithm framework.

A5. $U = \{U_{k+1} = \left(U_{k+1}^1, \ldots, U_{k+1}^\lambda\right) \in \mathbb{R}^{n\lambda}; k \in \mathbb{N}\}$, see e.g. (1), is an i.i.d. sequence that is also independent from (X_0, σ_0), and for all natural integer k, U_{k+1} is an independent sample of λ standard multivariate normal distributions on $\mathbb{R}^n$ at time $k+1$.

The last assumption is natural as ESs use predominantly Gaussian distributions[2]. Yet, we can replace the multivariate normal distribution by a distribution with finite first and second moments and a probability density function of the form $x \mapsto \frac{1}{\sigma^n} g\left(\frac{\|x\|^2}{\sigma^2}\right)$ where $\sigma > 0$ and $g : \mathbb{R}_+ \to \mathbb{R}_+$ is C^1, non-increasing and submultiplicative in that there exists $K > 0$ such that for $t \in \mathbb{R}_+$ and $s \in \mathbb{R}_+$, $g(t+s) \leq Kg(t)g(s)$ (such that Proposition 11 holds).

[2] In Evolution Strategies, Gaussian distributions are mainly used for convenience: they are the natural choice to generate rotationally invariant random vectors. Several attempts have been made to replace Gaussian distributions by Cauchy distributions [46, 59, 68]. Yet, their implementations are typically not rotational invariant and steep performance gains are observed either in low dimensions or crucially based on the implicit exploitation of separability [33].

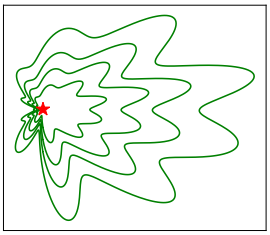

Fig. 1: Level sets of scaling-invariant functions with respect to the red star $x^\star$. A randomly generated scaling-invariant function from a "smoothly" randomly perturbed sphere function.

2.4 Assumptions on the objective function

Our main assumption on f to analyze the linear behavior of a step-size adaptive $(\mu/\mu_w, \lambda)$-ES is that it is scaling-invariant. We remind that f is scaling-invariant [14] with respect to a reference point $x^\star$ if for all $\rho > 0$, $x, y \in \mathbb{R}^n$

$$f(x^\star + x) \leq f(x^\star + y) \iff f\left(x^\star + \rho\, x\right) \leq f\left(x^\star + \rho\, y\right). \tag{12}$$

More precisely, we pose one of the following assumptions on f:

F1. The function f satisfies $f = \varphi \circ g$ where φ is a strictly increasing function and g is a C^1 scaling-invariant function with respect to $x^\star$ and has a unique global argmin (that is $x^\star$).

F2. The function f satisfies $f = \varphi \circ g$ where φ is a strictly increasing function and g is a nontrivial linear function.

Assumption F1 is our core assumption for studying convergence: we assume scaling invariance and continuous differentiability not on f but on g where $f = \varphi \circ g$ such that the function f can be discontinuous (we can include jumps in the function via the function φ). Because ESs are comparison-based algorithms and thus the selection function is identical on f or $g \circ f$ (see Lemma 1), our analysis is invariant if we carry it out on f or $g \circ f$. Strictly increasing transformations of strictly convex quadratic functions satisfy F1. Functions with non-convex sublevel sets can satisfy F1 (see Figure 1). More generally, strictly increasing transformations of C^1 positively homogeneous functions with a unique global argmin satisfy F1. Recall that a function p is positively homogeneous with degree $\alpha > 0$ and with respect to $x^\star$ if for all $x, y \in \mathbb{R}^n$, for all $\rho > 0$,

$$p(\rho(x - x^\star)) = \rho^\alpha p(x - x^\star) \ . \tag{13}$$

2.5 Preliminary results

If f is scaling-invariant with respect to $x^\star$, the composite of the selection function α_f with the translation $(z, u) \mapsto (x^\star + z, u)$ is positively homogeneous with degree 1. If in addition f is a measurable function with Lebesgue negligible level sets, then the explicit expression of the probability density function of $\alpha_f(x^\star + z, U_1)$ is known [19, Proposition 5.2] where U_1 follows the distribution of $\mathcal{N}_{n\lambda}$. These results are formalized in the next lemma.

Lemma 3 *If f is a scaling-invariant function with respect to $x^\star$, then the function $(z,u) \mapsto \alpha_f(x^\star + z, u)$ is positively homogeneous with degree 1. In other words, for all $z \in \mathbb{R}^n$, $\sigma > 0$ and $u = \left(u^1, \ldots, u^\lambda\right) \in \mathbb{R}^{n\lambda}$, $\alpha_f\left(x^\star + \sigma z, \sigma u\right) = \sigma \alpha_f\left(x^\star + z, u\right)$.*

If in addition f is a measurable function with Lebesgue negligible level sets and $U_1 = \left(U_1^1, \ldots, U_1^\lambda\right)$ is distributed according to $\mathcal{N}_{n\lambda}$, then for all $z \in \mathbb{R}^n$, the probability density function p_z^f of $\alpha_f(x^\star + z, U_1)$ exists and for all $u = (u^1, \ldots, u^\mu) \in \mathbb{R}^{n\mu}$,

$$p_z^f(u) = \frac{\lambda!}{(\lambda-\mu)!}(1 - Q_z^f(u^\mu))^{\lambda-\mu} \prod_{i=1}^{\mu-1} \mathbb{1}_{f(x^\star+z+u^i) < f(x^\star+z+u^{i+1})} \prod_{i=1}^{\mu} p_{\mathcal{N}_n}(u^i) \quad (14)$$

where $Q_z^f(w) = P\left(f\left(x^\star + z + \mathcal{N}_n\right) \leq f\left(x^\star + z + w\right)\right)$.

Proof. We have that $f(x^\star + z + u^{1:\lambda}) \leq \cdots \leq f(x^\star + z + u^{\lambda:\lambda})$ if and only if $f(x^\star + \sigma(z + u^{1:\lambda})) \leq \cdots \leq f(x^\star + \sigma(z + u^{\lambda:\lambda}))$. Therefore $\alpha_f\left(x^\star + \sigma z, \sigma u\right) = \sigma\left(u^{1:\lambda}, \ldots, u^{\mu:\lambda}\right) = \sigma\alpha_f\left(x^\star + z, u\right)$. Equation (14) holds whenever f has Lebesgue negligible level sets [19, Proposition 5.2]. □

On a linear function f, the selection function α_f defined in (3) is independent of the current state of the algorithm and is positively homogeneous with degree 1. We provide a simple formalism and proof of this result while it is already known and underlying previous works [11, 20].

Lemma 4 *If f is an increasing transformation of a linear function, then for all $x \in \mathbb{R}^n$ the function $\alpha_f(x, \cdot)$ does not depend on x and is positively homogeneous with degree 1. In other words, for $x \in \mathbb{R}^n$, $\sigma > 0$ and $u = \left(u^1, \ldots, u^\lambda\right) \in \mathbb{R}^{n\lambda}$, $\alpha_f\left(x, \sigma u\right) = \sigma\alpha_f\left(0, u\right)$.*

Proof. By linearity $f(x + \sigma u^{1:\lambda}) \leq \cdots \leq f(x + \sigma u^{\lambda:\lambda})$ if and only if $f(u^{1:\lambda}) \leq \cdots \leq f(u^{\lambda:\lambda})$. Therefore $\alpha_f\left(x, \sigma u\right) = \sigma\left(u^{1:\lambda}, \ldots, u^{\mu:\lambda}\right) = \sigma\alpha_f\left(0, u\right)$. □

Let $l^\star$ be the linear function defined for all $x \in \mathbb{R}^n$ as $l^\star(x) = x_1$ and $U_1 = \left(U_1^1, \ldots, U_1^\lambda\right)$ where $U_1^1, \ldots, U_1^\lambda$ are i.i.d. with law $\mathcal{N}_n$. Define the step-size change $\Gamma^\star_{\text{linear}}$ as

$$\Gamma^\star_{\text{linear}} = \Gamma\left(\alpha_{l^\star}(0, U_1)\right). \quad (15)$$

We prove in the next proposition that for all nontrivial linear functions, the step-size multiplicative factor of the algorithm (7) and (8) has at all iterations the distribution of $\Gamma^\star_{\text{linear}}$. This result derives from the rotation invariance of the function Γ (see Assumption A2) and of the probability density function $p_{\mathcal{N}_{n\mu}} : u \mapsto \frac{1}{(2\pi)^{n\mu/2}} \exp\left(-\|u\|^2/2\right)$. The details of the proof are in Appendix A.

Proposition 1 *(Invariance of the step-size multiplicative factor on linear functions) Let f be an increasing transformation of a nontrivial linear function, i.e. satisfy F2. Assume that the sequence $\{U_{k+1}\,; k \in \mathbb{N}\}$ satisfies Assumption A5 and that Γ satisfies Assumption A2, i.e. Γ is invariant under rotation. Then for all $z \in \mathbb{R}^n$ and all natural integer k, the step-size multiplicative factor $\Gamma\left(\alpha_f(z, U_{k+1})\right)$ has the law of the step-size change $\Gamma^\star_{\text{linear}}$ defined in (15).*

The proposition shows that on any (nontrivial) linear function the step-size change factor is independent of X_k, Z_k and even σ_k. We can now state the result which is at the origin of the methodology used in this paper, namely that on scaling-invariant functions, $\{Z_k = (X_k - x^\star)/\sigma_k \,;k \in \mathbb{N}\}$ is a homogeneous Markov chain. For this, we introduce the following function

$$F_w(z, v) = \frac{z + \sum_{i=1}^{\mu} w_i v_i}{\varGamma(v)} \text{ for all } (z, v) \in \mathbb{R}^n \times \mathbb{R}^{n\mu}, \tag{16}$$

which allows to write Z_{k+1} as a deterministic function of Z_k and U_{k+1}. The following proposition establishes conditions under which $\{Z_k; k \in \mathbb{N}\}$ is a homogeneous Markov chain that is defined with (16), independently of $\{(X_k, \sigma_k) \,; k \in \mathbb{N}\}$. We refer to $\{Z_k; k \in \mathbb{N}\}$ as the σ-normalized chain. This is a particular case from a more abstract algorithm framework [14, Proposition 4.1].

Proposition 2 *Let f be a scaling invariant function with respect to $x^\star$. Define the sequence $\{(X_k, \sigma_k); k \in \mathbb{N}\}$ as in (5) and (6). Then $\{Z_k = (X_k - x^\star)/\sigma_k \,; k \in \mathbb{N}\}$ is a homogeneous Markov chain and for all natural integer k, the following equation holds*

$$Z_{k+1} = F_w \left(Z_k, \alpha_f \left(x^\star + Z_k,\, U_{k+1}\right)\right), \tag{17}$$

where α_f is defined in (3), F_w is defined in (16) and $\{U_{k+1} \,; k \in \mathbb{N}\}$ is the sequence of random inputs used to sample the candidate solutions in (1) corresponding to the random input in (7) and (8).

Proof. The definition of the selection function α_f allows to write (5) and (6) as (7) and (8). We have $Z_{k+1} = \frac{X_{k+1} - x^\star}{\sigma_{k+1}} = \frac{X_k - x^\star + \sum_{i=1}^{\mu} w_i \left[\alpha_f(X_k, \sigma_k U_{k+1})\right]_i}{\sigma_k \, \varGamma\left(\frac{\alpha_f(X_k, \sigma_k U_{k+1})}{\sigma_k}\right)} = \frac{Z_k + \sum_{i=1}^{\mu} w_i \frac{\left[\alpha_f(X_k, \sigma_k U_{k+1})\right]_i}{\sigma_k}}{\varGamma\left(\frac{\alpha_f(X_k, \sigma_k U_{k+1})}{\sigma_k}\right)}$. By Lemma 3, $\frac{\alpha_f(X_k, \sigma_k U_{k+1})}{\sigma_k} = \frac{\alpha_f(x^\star + X_k - x^\star, \sigma_k U_{k+1})}{\sigma_k} = \alpha_f(x^\star + \frac{X_k - x^\star}{\sigma_k}, U_{k+1})$. Then $Z_{k+1} = F_w \left(Z_k, \alpha_f \left(x^\star + Z_k, U_{k+1}\right)\right)$ and $\{Z_k; k \in \mathbb{N}\}$ is a homogeneous Markov chain. □

Three invariances are key to obtain that $\{Z_k = (X_k - x^\star)/\sigma_k \,; k \in \mathbb{N}\}$ is a homogeneous Markov chain: invariance to strictly increasing transformations (stemming from the comparison-based property of ESs), translation invariance, and scale invariance [14, Proposition 4.1]. The last two invariances are satisfied with the update we assume for mean and step-size.

3 Methodology and overview of the rest of the analysis

We present in this section the main idea behind the proof methodology used in this paper, namely how the stability study of an underlying Markov chain leads to convergence (or divergence) of the original algorithm. From there, we sketch the different steps of the analysis and present an overview of the structure of the rest of the mathematical analysis.

We aim at proving linear convergence that can be visualized by looking at the distance to the optimum: after an adaptation phase, we observe that the log

distance to the optimum diverges to minus infinity with a graph that resembles a straight line with random perturbations. The step-size converges to zero at the same linear rate (see Figure 2), the so-called convergence rate of the algorithm. Formally, in case of convergence, there exists $\mathrm{r} > 0$ such that

$$\lim_{k\to\infty} \frac{1}{k} \log \frac{\|X_k - x^\star\|}{\|X_0 - x^\star\|} = \lim_{k\to\infty} \frac{1}{k} \log \frac{\sigma_k}{\sigma_0} = -\mathrm{r} \tag{18}$$

where $x^\star$ is the optimum of the function.

We consider a scaling invariant function with respect to x^*. From Proposition 2, we know that $\{Z_k = (X_k - x^\star)/\sigma_k \,; k \in \mathbb{N}\}$ is a homogeneous Markov chain where $\{(X_k, \sigma_k) \,; k \in \mathbb{N}\}$ is the sequence of states of the step-size adaptive $(\mu/\mu_w, \lambda)$-ES defined in (5) and (8) (see Proposition 2). We use this Markov chain to write the log progress in the following way:

$$\begin{aligned} \log \frac{\|X_{k+1} - x^\star\|}{\|X_k - x^\star\|} &= \log \frac{\|Z_{k+1}\|}{\|Z_k\|} + \log \frac{\sigma_{k+1}}{\sigma_k} \\ &= \log \frac{\|Z_{k+1}\|}{\|Z_k\|} + \log \left(\Gamma\left(\alpha_f\left(x^\star + Z_k, U_{k+1}\right)\right)\right) \end{aligned} \tag{19}$$

where Γ and α_f are defined in (6) and in (3). This equation can now be used to express the term whose limit we need to investigate:

$$\frac{1}{k} \log \frac{\|X_k - x^\star\|}{\|X_0 - x^\star\|} = \frac{1}{k} \sum_{t=0}^{k-1} \log \frac{\|X_{t+1} - x^\star\|}{\|X_t - x^\star\|} \tag{20}$$

$$= \frac{1}{k} \sum_{t=0}^{k-1} \log \frac{\|Z_{t+1}\|}{\|Z_t\|} + \frac{1}{k} \sum_{t=0}^{k-1} \log(\Gamma(\alpha_f(x^\star + Z_t, U_{t+1}))) \ . \tag{21}$$

This latter equation suggests that if we can apply a law of large numbers to $\{Z_k \,; k \in \mathbb{N}\}$ and $\{(Z_k, U_{k+1}) \,; k \in \mathbb{N}\}$, the right-hand side of (21) converges when k goes to infinity to $\int \mathbb{E}_{U_1 \sim \mathcal{N}_{n\lambda}} \left[\log \left(\Gamma\left(\alpha_f(x^\star + z, U_1)\right)\right)\right] \pi(dz) = \mathbb{E}_\pi(\mathcal{R}_f)$ where $\mathcal{R}_f$ is defined as the expected change of the logarithm of the step-size for any state $z \in \mathbb{R}^n$ of the σ-normalized chain as

$$\mathcal{R}_f(z) = \mathbb{E}_{U_1 \sim \mathcal{N}_{n\lambda}} \left[\log \left(\Gamma\left(\alpha_f(x^\star + z, U_1)\right)\right)\right] \ , \tag{22}$$

and π is the invariant measure of $\{Z_k \,; k \in \mathbb{N}\}$. From there, we obtain the almost sure convergence of $\frac{1}{k} \log \frac{\|X_k - x^\star\|}{\|X_0 - x^\star\|}$ towards $\mathbb{E}_\pi(\mathcal{R}_f)$ expressed in (34) characterizing the asymptotic linear behavior of the algorithm. A similar equation can be established to prove the convergence of $1/k \log(\sigma_k/\sigma_0)$. Convergence of the expected log-progress can also be deduced from stability properties of $\{Z_k; k \geq 0\}$.

The idea to apply a Law of Large Numbers (LLN) to the chain $\{Z_k; k \geq 0\}$ to prove the asymptotic linear behavior of the underlying algorithm is the key behind the asymptotic almost sure linear behavior proof we provide. This seminal idea was introduced for self-adaptive ES on the sphere function [15] and exploited to prove their linear behavior [11] and generalized to a wider class of algorithms and functions [14].

Hence, in order to obtain a proof of the linear behavior of the studied algorithm following the idea sketched above, we need to investigate now the stability of the

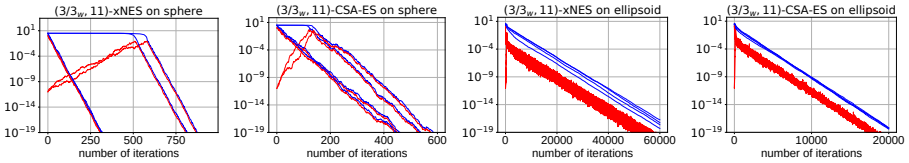

Fig. 2: Four independent runs of $(\mu/\mu_w, \lambda)$-xNES and $(\mu/\mu_w, \lambda)$-CSA1-ES (without cumulation) as presented in Section 2.1 on the functions $x \mapsto \|x\|^2$ (first two figures) and $x \mapsto \sum_{i=1}^{n} 10^{3\frac{i-1}{n-1}} x_i^2$ (last two figures). Illustration of $\|X_k\|$ in blue and σ_k in red where k is the number of iterations, $\mu = 3$, $\lambda = 11$ and $w_i = 1/\mu$. Initializations: σ_0 equals to 10^{-11} in two runs and 1 in the two other runs, X_0 is the all-ones vector in dimension 10.

chain $\{Z_k\,; k \in \mathbb{N}\}$ (and in turn $\{(Z_k, U_{k+1})\,; k \in \mathbb{N}\}$). In particular, we need to prove that it satisfies the mathematical properties referred to informally as stability properties (following a terminology by Meyn and Tweedie [50]) such that an LLN can be applied. It is not a trivial task and it will occupy a large part of the rest of the paper. While establishing stability properties to obtain an LLN we will prove stronger properties that will allow to state convergence of the expected log progress and a Central Limit Theorem. The outline of the remaining mathematical analysis and the proof structure is as follows:

- In Section 4, we introduce different notions related to Markov chains, notably the stability properties that we will prove like φ-irreducibility, aperiodicity, positivity, Harris-recurrence and geometric ergodicity. We also introduce the different practical tools to prove that a Markov chain satisfies those properties.
- In Section 5, we establish those stability properties for the Markov chain $\{Z_k; k \geq 0\}$ associated to a step-size adaptive $(\mu/\mu_w, \lambda)$-ES under the appropriate conditions on the objective functions and the step-size adaptation mechanism.
- In Section 6, we use those properties to prove the linear behavior of the studied algorithms. In addition to the asymptotic almost sure linear behavior stemming from the LLN, we establish convergence in terms of expected log progress and a Central Limit Theorem. Our conditions for linear convergence are expressed for an abstract step-size update. We investigate how those conditions translate to the case of the CSA and xNES step-size updates.

4 Reminders on Markov chains and various tools

We consider a Markov chain $\{Z_k\,; k \in \mathbb{N}\}$ on a measure space $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}), P)$ where $\mathcal{Z}$ in an open subset of $\mathbb{R}^n$, for all $k \in \mathbb{N}$ its k-step transition kernel as $P^k(z, A) = P\left(Z_k \in A | Z_0 = z\right)$ for $z \in \mathcal{Z}$, $A \in \mathcal{B}(\mathcal{Z})$. We also denote $P(z, A)$ and $P_z(A)$ as $P^1(z, A)$. We remind different stability notions investigated later on to prove in particular that $\{Z_k\,; k \in \mathbb{N}\}$ satisfies an LLN, a central limit theorem and that for some $z \in \mathcal{Z}$, $P^k(z, \cdot)$ converges to a stationary distribution. We additionally present different tools to be able to verify that a Markov chain satisfies those various properties.

4.1 Stability properties and practical drift conditions

If there exists a nontrivial measure φ on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ such that for all $A \in \mathcal{B}(\mathcal{Z})$, $\varphi(A) > 0$ implies $\sum_{k=1}^{\infty} P^k(z, A) > 0$ for all $z \in \mathcal{Z}$, then the chain is called φ-irreducible. A φ-irreducible Markov chain is Harris recurrent if for all $A \in \mathcal{B}(\mathcal{Z})$ with $\varphi(A) > 0$ and for all $z \in \mathcal{Z}$, $P_z(\eta_A = \infty) = 1$, where $\eta_A = \sum_{k=1}^{\infty} \mathbb{1}_{Z_k \in A}$ is the *occupation time* of A.

A σ-finite measure π on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ is an invariant measure for $\{Z_k; k \in \mathbb{N}\}$ if for all $A \in \mathcal{B}(\mathcal{Z})$, $\pi(A) = \int_{\mathcal{Z}} \pi(dz) P(z, A)$. A Harris recurrent chain admits a unique (up to constant multiples) invariant measure π (see [50, Theorem 10.0.1]). A φ-irreducible Markov chain admitting an invariant probability measure π is said positive. A positive Harris-recurrent chain satisfies an LLN as reminded below.

Theorem 1 *[50, Theorem 17.0.1] If $\{Z_k; k \in \mathbb{N}\}$ is a positive and Harris recurrent chain with invariant probability measure π, then the LLN holds for any π-integrable function g, i.e. for any g with $\mathbb{E}_\pi(|g|) < \infty$, $\lim_{k\to\infty} \frac{1}{k} \sum_{t=0}^{k-1} g(Z_t) = \mathbb{E}_\pi(g)$.*

We will need the notion of aperiodicity. Assume that d is a positive integer and $\{Z_k\,; k \in \mathbb{N}\}$ is a φ-irreducible Markov chain defined on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$. Let $(D_i)_{i=1,\ldots,d} \in \mathcal{B}(\mathcal{Z})^d$ be a sequence of disjoint sets. Then $(D_i)_{i=1,\ldots,d}$ is called a d-cycle if

(i) $P(z, D_{i+1}) = 1$ for all $z \in D_i$ and $i = 0, \ldots, d-1 \pmod d$,
(ii) $\Lambda\left(\left(\bigcup_{i=1}^{d} D_i\right)^c\right) = 0$ for all irreducibility measure Λ of $\{Z_k\,; k \in \mathbb{N}\}$.

If $\{Z_k; k \in \mathbb{N}\}$ is φ-irreducible, there exists a d-cycle where d is a positive integer [50, Theorem 5.4.4]. The largest d for which there exists a d-cycle is called the period of $\{Z_k\,; k \in \mathbb{N}\}$. We then say that a φ-irreducible Markov chain $\{Z_k\,; k \in \mathbb{N}\}$ on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ is aperiodic if it has a period of 1.

A set $C \in \mathcal{B}(\mathcal{Z})$ is called *small* if there exists a positive integer k and a nontrivial measure ν_k on $\mathcal{B}(\mathcal{Z})$ such that $P^k(z, A) \geq \nu_k(A)$ for all $z \in C$, $A \in \mathcal{B}(\mathcal{Z})$. We then say that C is a ν_k-small set [50].

Given an extended-valued, non-negative and measurable function $V : \mathcal{Z} \to \mathbb{R}_+ \cup \{\infty\}$ (called potential function), the drift operator is defined for all $z \in \mathcal{Z}$ as $\Delta V(z) = \mathbb{E}[V(Z_1)|Z_0 = z] - V(z) = \int_{\mathcal{Z}} V(y) P(z, \mathrm{d}y) - V(z)$. A φ-irreducible, aperiodic Markov chain $\{Z_k\,; k \in \mathbb{N}\}$ defined on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ satisfies a geometric drift condition if there exist $0 < \gamma < 1$, $b \in \mathbb{R}$, a small set C and a potential function V greater than 1, finite at some $z_0 \in \mathcal{Z}$ such that for all $z \in \mathcal{Z} : \Delta V(z) \leq (\gamma - 1)V(z) + b\mathbb{1}_C(z)$, or equivalently if $\mathbb{E}[V(Z_1)|Z_0 = z] \leq \gamma V(z) + b\mathbb{1}_C(z)$. The function V is called a geometric drift function and if $\{y \in \mathcal{Z}\,; V(y) < \infty\} = \mathcal{Z}$, we say that $\{Z_k\,; k \in \mathbb{N}\}$ is V-geometrically ergodic.

If a φ-irreducible and aperiodic Markov chain is V-geometrically ergodic, then it is positive and Harris recurrent [50, Theorem 13.0.1 and Theorem 9.1.8]. We prove a geometric drift condition in Section 5.3, this in turn implies positivity and Harris-recurrence.

From a geometric drift condition follows a stronger result than an LLN, namely a central limit theorem.

Theorem 2 *[50, Theorem 17.0.1 and Theorem 16.0.1] Let $\{Z_k\,; k \in \mathbb{N}\}$ be a φ-irreducible aperiodic Markov chain on $(\mathcal{Z}, \mathcal{B}(\mathcal{Z}))$ that is V-geometrically ergodic with*

invariant probability measure π. *For any function* g *on* $\mathcal{Z}$ *that satisfies* $g^2 \leq V$, *the central limit theorem holds for* $\{Z_k\,;k\in\mathbb{N}\}$ *in the following sense. Define* $\bar{g} = g - \mathbb{E}_\pi(g)$ *and for all positive integer* t, *define* $S_t(\bar{g}) = \sum_{k=0}^{t-1} \bar{g}(Z_k)$. *Then the constant* $\gamma^2 = \mathbb{E}_\pi[(\bar{g}(Z_0))^2] + 2\sum_{k=1}^{\infty} \mathbb{E}_\pi[\bar{g}(Z_0)\bar{g}(Z_k)]$ *is well defined, non-negative, finite and* $\lim_{t\to\infty} \frac{1}{t}\mathbb{E}_\pi[(S_t(\bar{g}))^2] = \gamma^2$. *Moreover if* $\gamma^2 > 0$ *then* $\frac{1}{\sqrt{t\gamma^2}} S_t(\bar{g})$ *converges in distribution to* $\mathcal{N}(0,1)$ *when* t *goes to* ∞, *else if* $\gamma^2 = 0$ *then* $\frac{1}{\sqrt{t}} S_t(\bar{g}) = 0$ *a.s.*

For a measurable function $h \geq 1$ on $\mathcal{Z}$, a φ-irreducible aperiodic Markov chain $\{Z_k\,;k\in\mathbb{N}\}$ defined on $(\mathcal{Z},\mathcal{B}(\mathcal{Z}))$ is positive Harris recurrent with invariant probability measure π such that h is π-integrable if and only if there exist $b\in\mathbb{R}$, a small set C and an extended-valued non-negative function $V \neq \infty$ such that

$$\Delta V(z) \leq -h(z) + b\mathbb{1}_C(z) \tag{23}$$

for all $z\in\mathcal{Z}$ [50, Theorem 14.0.1]. Recall that for a measurable function $h\geq 1$, we say that a general Markov chain $\{Z_k\,;k\in\mathbb{N}\}$ is h-ergodic if there exists a probability measure π such that $\lim_{k\to\infty} \|P^k(z,\cdot) - \pi\|_h = 0$ for any initial condition z. The probability measure π is then called the invariant probability measure of $\{Z_k\,;k\in\mathbb{N}\}$. If $h=1$, we say that $\{Z_k\,;k\in\mathbb{N}\}$ is ergodic.

A φ-irreducible aperiodic Markov chain on $\mathcal{Z}$ that satisfies (23) is h-ergodic if in addition $\{y\in\mathcal{Z}\,;V(y)<\infty\} = \mathcal{Z}$ [50, Theorem 14.0.1].

Prior to establishing a drift condition, we need to identify small sets. Using the notion of T-chain defined below, compact sets are small sets because for a φ-irreducible aperiodic T-chain, every compact set is a small set [50, Theorem 5.5.7 and Theorem 6.2.5].

The T-chain property calls for the notion of kernel: a kernel K is a function on $(\mathcal{Z},\mathcal{B}(\mathcal{Z}))$ such that for all $A\in\mathcal{B}(\mathcal{Z})$, $K(.,A)$ is a measurable function and for all $z\in\mathcal{Z}$, $K(z,.)$ is a signed measure. A non-negative kernel K satisfying $K(z,\mathcal{Z})\leq 1$ for all $z\in\mathcal{Z}$ is called substochastic. A substochastic kernel K satisfying $K(z,\mathcal{Z}) = 1$ for all $z\in\mathcal{Z}$ is a transition probability kernel. Let b be a probability distribution on $\mathbb{N}$ and denote by K_b the probability transition kernel defined as $K_b(z,A) = \sum_{k=0}^{\infty} b(k)P^k(z,A)$ for all $z\in\mathcal{Z}$, $A\in\mathcal{B}(\mathcal{Z})$. If T is a substochastic transition kernel such that $T(.,A)$ is lower semi-continuous for all $A\in\mathcal{B}(\mathcal{Z})$ and $K_b(z,A)\geq T(z,A)$ for all $z\in\mathcal{Z}$, $A\in\mathcal{B}(\mathcal{Z})$, then T is called a continuous component of K_b. If a Markov chain $\{Z_k\,;k\in\mathbb{N}\}$ admits a probability distribution b on $\mathbb{N}$ such that K_b has a continuous component T that satisfies $T(z,\mathcal{Z})>0$ for all $z\in\mathcal{Z}$, then $\{Z_k\,;k\in\mathbb{N}\}$ is called a T-chain.

4.2 Generalized law of large numbers

To apply an LLN for the convergence of the term $\frac{1}{k}\sum_{t=0}^{k-1}\log(\Gamma(\alpha_f(x^\star + Z_t, U_{t+1})))$ in (21), we proceed in two steps. First we prove that if $\{Z_k\,;k\in\mathbb{N}\}$ is defined as $Z_{k+1} = G(Z_k, U_{k+1})$ where $G:\mathcal{Z}\times\mathbb{R}^m\to\mathcal{Z}$ is a measurable function and $\{U_{k+1}\,;k\in\mathbb{N}\}$ is a sequence of i.i.d. random vectors, then the ergodic properties of $\{Z_k\,;k\in\mathbb{N}\}$ are transferred to $\{W_k = (Z_k, U_{k+2})\,;k\in\mathbb{N}\}$. Afterwards we apply a generalized LLN recalled in the following theorem.

Theorem 3 ([45, Theorem 1]) *Assume that $\{Z_k\,;k\in\mathbb{N}\}$ is a homogeneous Markov chain on an abstract measurable space $(\mathbf{E},\mathcal{E})$ that is ergodic with invariant probability measure π. For all measurable function $g:\mathbf{E}^\infty\to\mathbb{R}$ such that for all $s\in\mathbb{N}$, $\mathbb{E}_\pi(|g(Z_s,Z_{s+1},\dots)|)<\infty$ and for any initial distribution Λ, the generalized LLN holds as follows $\lim_{k\to\infty}\frac{1}{k}\sum_{t=0}^{k-1}g\,(Z_t,Z_{t+1},\dots)=\mathbb{E}_\pi(g(Z_s,Z_{s+1},\dots))$ P_Λ a.s. where P_Λ is the distribution of the process $\{Z_k\,;k\in\mathbb{N}\}$ on $(\mathbf{E}^\infty,\mathcal{E}^\infty)$.*

Theorem 3 generalizes the case where the initial state is distributed under the invariant measure [63, Theorems 3.5.7 and 3.5.8] to an arbitrary initial distribution.

If we have the generalized LLN for a chain $\{(Z_k,U_{k+2})\,;k\in\mathbb{N}\}$ on $\mathbb{R}^n\times\mathbb{R}^m$, then an LLN for the chain $\{(Z_k,U_{k+1})\,;k\in\mathbb{N}\}$ is directly implied. We formalize this statement in the next corollary.

Corollary 1 *Assume that $\{W_k=(Z_k,U_{k+2})\,;k\in\mathbb{N}\}$ is a homogeneous Markov chain on $\mathbb{R}^n\times\mathbb{R}^m$ that is ergodic with invariant probability measure π. Then the LLN holds for $\{(Z_k,U_{k+1})\,;k\in\mathbb{N}\}$ in the following sense. Define the function $T:(\mathbb{R}^n\times\mathbb{R}^m)^2\to\mathbb{R}^n\times\mathbb{R}^m$ as $T((z_1,u_3),(z_2,u_4))=(z_2,u_3)$. If $g:\mathbb{R}^n\times\mathbb{R}^m\to\mathbb{R}$ is such that for all $s\in\mathbb{N}$, $\mathbb{E}_\pi(|g\circ T|\,(W_s,W_{s+1}))<\infty$, then $\lim_{k\to\infty}\frac{1}{k}\sum_{t=0}^{k-1}g(Z_t,U_{t+1})=\mathbb{E}_\pi\left[(g\circ T)\,(W_s,W_{s+1})\right]$.*

Proof. We have $\lim_{k\to\infty}\frac{1}{k}\sum_{t=0}^{k-1}(g\circ T)\,(W_t,W_{t+1})=\mathbb{E}_\pi\left[(g\circ T)\,(W_s,W_{s+1})\right]$ thanks to Theorem 3. For $t\in\mathbb{N}$, $(g\circ T)\,(W_t,W_{t+1})=g\,(Z_{t+1},U_{t+2})$. Therefore

$$\mathbb{E}_\pi\left[(g\circ T)\,(W_s,W_{s+1})\right]=\lim_{k\to\infty}\frac{1}{k}\sum_{t=0}^{k-1}g\,(Z_{t+1},U_{t+2})=\lim_{k\to\infty}\frac{1}{k}\sum_{t=0}^{k-1}g(Z_t,U_{t+1}).\qquad\square$$

We formulate now that for a Markov chain following a non-linear state space model of the form $Z_{k+1}=G(Z_k,U_{k+1})$ with G measurable and $\{U_{k+1}\,;k\in\mathbb{N}\}$ i.i.d., then φ-irreducibility, aperiodicity and V-geometric ergodicity of Z_k are transferred to $\{W_k=(Z_k,U_{k+2})\,;k\in\mathbb{N}\}$. We provide a proof of this result in Appendix B for the sake of completeness.

Proposition 3 *Let $\{Z_k\,;k\in\mathbb{N}\}$ be a Markov chain on $(\mathcal{Z},\mathcal{B}(\mathcal{Z}))$ defined as $Z_{k+1}=G(Z_k,U_{k+1})$ where $G:\mathcal{Z}\times\mathbb{R}^m\to\mathcal{Z}$ is a measurable function and $\{U_{k+1}\,;k\in\mathbb{N}\}$ is a sequence of i.i.d. random vectors with probability measure Ψ. Consider $\{W_k=(Z_k,U_{k+2})\,;k\geq 0\}$, then it is a Markov chain on $\mathcal{B}(\mathcal{Z})\otimes\mathcal{B}(\mathbb{R}^m)$ which inherits properties of $\{Z_k\,;k\in\mathbb{N}\}$ in the following sense:*

- *If φ (resp. π) is an irreducibility (resp. invariant) measure of $\{Z_k\,;k\in\mathbb{N}\}$, then $\varphi\times\Psi$ (resp. $\pi\times\Psi$) is an irreducibility (resp. invariant) measure of $\{W_k\,;k\in\mathbb{N}\}$.*
- *The set of integers d such that there exists a d-cycle for $\{Z_k\,;k\in\mathbb{N}\}$ is equal to the set of integers d such that there exists a d–cycle for $\{W_k\,;k\in\mathbb{N}\}$. In particular $\{Z_k\,;k\in\mathbb{N}\}$ and $\{W_k\,;k\in\mathbb{N}\}$ have the same period. Therefore $\{Z_k\,;k\in\mathbb{N}\}$ is aperiodic if and only if $\{W_k\,;k\in\mathbb{N}\}$ is aperiodic.*
- *If C is a small set for $\{Z_k\,;k\in\mathbb{N}\}$, then $C\times\mathbb{R}^m$ is a small set for $\{W_k\,;k\in\mathbb{N}\}$.*
- *If $\{Z_k\,;k\in\mathbb{N}\}$ satisfies a drift condition*

$$\Delta V(z)\leq-\beta h(z)+b\mathbb{1}_C(z)\quad\text{for all } z\in\mathcal{Z},\tag{24}$$

where V is a potential function, $0<\beta<1$, $h\geq 0$ is a measurable function and $C\subset\mathcal{Z}$ is a measurable set, then $\{W_k\,;k\in\mathbb{N}\}$ satisfies the following drift condition for all $(z,u)\in\mathcal{Z}\times\mathbb{R}^m$: $\Delta\widetilde{V}(z,u)\leq-\beta\widetilde{h}(z,u)+b\mathbb{1}_{C\times\mathbb{R}^m}(z,u)$, where $\widetilde{V}:(z,u)\mapsto V(z)$ and $\widetilde{h}:(z,u)\mapsto h(z)$.

Remark that the drift condition in (24) includes the geometric drift condition by taking $h = V$, the drift condition for h-ergodicity by dividing the equation by β and assuming that $h \geq 1$, for positivity and Harris recurrence by taking $h = 1/\beta$, and for Harris recurrence by taking $h = 0$. This is obtained assuming that V and C satisfy the proper assumptions for the drift to hold.

4.3 φ-irreducibility, aperiodicity and T-chain property via deterministic control models

For the Markov chain considered, it is difficult to establish φ-irreducible, aperiodicity and the T-chain property "by hand". We thus resort to tools connecting those properties to stability properties of the underlying control model [50, Chapter 13] [19]. Assume that $\mathcal{Z}$ is an open subset of $\mathbb{R}^n$. We consider a Markov chain that takes the following form

$$Z_{k+1} = F\left(Z_k, \alpha\left(Z_k, U_{k+1}\right)\right), \tag{25}$$

where $Z_0 \in \mathcal{Z}$ and for all natural integer k, $F : \mathcal{Z} \times \mathbb{R}^{n\mu} \to \mathcal{Z}$ and $\alpha : \mathcal{Z} \times \mathbb{R}^{n\lambda} \to \mathbb{R}^{n\mu}$ are measurable functions, $U = \left\{U_{k+1} \in \mathbb{R}^{n\lambda}\,;\, k \in \mathbb{N}\right\}$ is a sequence of i.i.d. random vectors. We consider the following assumptions on the model:

B1. (Z_0, U) are random variables on a probability space $(\Omega, \mathcal{F}, P_{Z_0})$.
B2. Z_0 is independent of U.
B3. U is an independent and identically distributed process.
B4. For all $z \in \mathcal{Z}$, the random variable $\alpha(z, U_1)$ admits a probability density function denoted by p_z, such that the function $(z, u) \mapsto p_z(u)$ is lower semi-continuous.
B5. The function $F : \mathcal{Z} \times \mathbb{R}^{n\mu} \to \mathcal{Z}$ is C^1.

We recall the deterministic control model related to (25) denoted by CM(F) [19]. It is based on the notion of extended transition map function [49], defined recursively for all $z \in \mathcal{Z}$ as $S_z^0 = z$, and for all $k \in \mathbb{N}\backslash\{0\}$, $S_z^k : \mathbb{R}^{n\mu k} \to \mathcal{Z}$ such that for all $\mathbf{w} = (w_1, \ldots, w_k) \in \mathbb{R}^{n\mu k}$, $S_z^k(\mathbf{w}) = F\left(S_z^{k-1}\left(w_1, \ldots, w_{k-1}\right), w_k\right)$. Assume in the following that Assumptions B1−B4 are satisfied and that F is continuous.

Let us define the process W for all $k \in \mathbb{N}\backslash\{0\}$ and $z \in \mathcal{Z}$ as $W_1 = \alpha\left(z, U_1\right)$ and $W_k = \alpha\left(S_z^{k-1}(W_1, \ldots, W_{k-1}), U_k\right)$. Then the probability density function of $(W_1, W_2, \ldots, W_k)$ denoted by p_z^k is what is called the extended probability function. It is defined inductively for all $k \in \mathbb{N}\backslash\{0\}$, $\mathbf{w} = (w_1, \ldots, w_k) \in \mathbb{R}^{n\mu k}$ by $p_z^1(w_1) = p_z(w_1)$ and $p_z^k(\mathbf{w}) = p_z^{k-1}\left(w_1, \ldots, w_{k-1}\right) p_{S_z^{k-1}(w_1,\ldots,w_{k-1})}(w_k)$. For all $k \in \mathbb{N}\backslash\{0\}$ and for all $z \in \mathcal{Z}$, the control sets are finally defined as $\mathcal{O}_z^k = \left\{\mathbf{w} \in \mathbb{R}^{n\mu k}\ ;\ p_z^k(\mathbf{w}) > 0\right\}$. The control sets are open sets since F is continuous and the functions $(z, \mathbf{w}) \mapsto p_z^k(\mathbf{w})$ are lower semi-continuous (see [19] for more details).

The deterministic control model CM(F) is defined recursively for all $k \in \mathbb{N}$, $z \in \mathcal{Z}$ and $(w_1, \ldots, w_{k+1}) \in \mathcal{O}_z^{k+1}$ as $S_z^{k+1}\left(w_1, \ldots, w_{k+1}\right) = F\left(S_z^k\left(w_1, \ldots, w_k\right), w_{k+1}\right)$.

For $z \in \mathcal{Z}$, $A \in \mathcal{B}(\mathcal{Z})$ and $k \in \mathbb{N}\backslash\{0\}$, we say that $\mathbf{w} \in \mathbb{R}^{n\mu k}$ is a k-steps path from z to A if $\mathbf{w} \in \mathcal{O}_z^k$ and $S_z^k(\mathbf{w}) \in A$. We introduce for $z \in \mathcal{Z}$ and $k \in \mathbb{N}$ the set

of all states reachable from z in k steps by CM(F), denoted by $A_+^k(z)$ and defined as $A_+^0(z) = \{z\}$ and $A_+^k(z) = \left\{ S_z^k(\mathbf{w})\,;\, \mathbf{w} \in \mathcal{O}_z^k \right\}$.

A point $z \in \mathcal{Z}$ is a steadily attracting state if for all $y \in \mathcal{Z}$, there exists a sequence $\left\{ y_k \in A_+^k(y) | \, k \in \mathbb{N} \setminus \{0\} \right\}$ that converges to z.

The controllability matrix is defined for $k \in \mathbb{N}\setminus\{0\}$, $z \in \mathcal{Z}$ and $\mathbf{w} \in \mathbb{R}^{n\mu k}$ as the Jacobian matrix of $(w_1, \ldots, w_k) \mapsto S_z^k(w_1, \ldots, w_k)$ and denoted by $C_z^k(\mathbf{w})$. Namely, $C_z^k(\mathbf{w}) = \left[\frac{\partial S_z^k}{\partial w_1}(\mathbf{w}) | \ldots | \frac{\partial S_z^k}{\partial w_k}(\mathbf{w}) \right]$.

If F is C^1, the existence of a steadily attracting state z and a full-rank condition on a controllability matrix of z imply that a Markov chain following (25) is a φ-irreducible aperiodic T-chain, as reminded in the next theorem.

Theorem 4 *[19, Theorem 4.4: Practical condition to be a φ-irreducible aperiodic T-chain.] Consider a Markov chain $\{Z_k\,; k \in \mathbb{N}\}$ following the model (25) for which the conditions B1−B5 are satisfied. If there exist a steadily attracting state $z \in \mathcal{Z}$, $k \in \mathbb{N}\setminus\{0\}$ and $\mathbf{w} \in \mathcal{O}_z^k$ such that $\operatorname{rank}\left(C_z^k(\mathbf{w})\right) = n$, then $\{Z_k\,; k \in \mathbb{N}\}$ is a φ-irreducible aperiodic T-chain, and every compact set is a small set.*

The next lemma allows to loosen the full-rank condition stated above if the control set $\mathcal{O}_z^k$ is dense in $\mathbb{R}^{n\mu k}$.

Lemma 5 *Consider a Markov chain $\{Z_k\,; k \in \mathbb{N}\}$ following the model (25) for which the conditions B1−B5 are satisfied. Assume that there exist a positive integer k and $z \in \mathcal{Z}$ such that the control set $\mathcal{O}_z^k$ is dense in $\mathbb{R}^{n\mu k}$. If there exists $\widetilde{\mathbf{w}} \in \mathbb{R}^{n\mu k}$ such that $\operatorname{rank}(C_z^k(\widetilde{\mathbf{w}})) = n$, then the rank condition in Theorem 4 is satisfied, i.e. there exists $\mathbf{w} \in \mathcal{O}_z^k$ such that $\operatorname{rank}(C_z^k(\mathbf{w})) = n$.*

Proof. The function $w \mapsto S_z^k(w)$ is C^1 [19, Lemma 6.1]. Since the set of full rank matrices is open, there exists an open neighborhood $\mathcal{V}_{\widetilde{\mathbf{w}}}$ of $\widetilde{\mathbf{w}}$ such that for all $w \in \mathcal{V}_{\widetilde{\mathbf{w}}}$, $\operatorname{rank}(C_z^k(w)) = n$. By density of $\mathcal{O}_z^k$, the non-empty set $\mathcal{V}_{\widetilde{\mathbf{w}}} \cap \mathcal{O}_z^k$ contains an element $\mathbf{w}$. □

If z is steadily attracting, there exists under mild assumptions an open set outside of a ball centered at z, with positive measure with respect to the invariant probability measure of a chain following the model (25) as stated next.

Lemma 6 *Consider a Markov chain $\{Z_k\,; k \in \mathbb{N}\}$ on $\mathbb{R}^n$ following the model (25) for which the conditions B1−B5 are satisfied. Assume that there exist a steadily attracting state $z \in \mathbb{R}^n$ such that $\mathcal{O}_z^1$ is dense in $\mathbb{R}^n$ and $w \in \mathcal{O}_z^1$ with $\operatorname{rank}\left(C_z^1(w)\right) = n$. Assume also that $\{Z_k\,; k \in \mathbb{N}\}$ is a positive Harris recurrent chain with invariant probability measure π. Then there exists $0 < \epsilon < 1$ such that $\pi(\mathbb{R}^n \setminus \overline{\mathbf{B}(z, \epsilon)}) > 0$.*

Proof. A φ-irreducible Markov chain admits a maximal irreducibility measure ψ dominating any other irreducibility measure [50, Theorem 4.0.1]. In other words, for a measurable set A, $\psi(A) = 0$ induces that $\varphi(A) = 0$ for any irreducibility measure φ. The measure π is equivalent to the maximal irreducibility measure ψ [50, Theorem 10.4.9]. Since z is steadily attracting, $\operatorname{supp}\,\psi = \overline{A_+(z)} = \overline{\bigcup_{k\in\mathbb{N}} \left\{ S_z^k(\mathbf{w})\,;\mathbf{w} \in \mathcal{O}_z^k \right\}}$ [19, Propositions 3.3 and 4.2]. We have $\operatorname{rank}\left(C_z^1(w)\right) = n$, therefore the function $F(z, \cdot)$ is not constant. Along with the density of $\mathcal{O}_z^1$, we

obtain that there exists $\epsilon > 0$ and a vector $v \in \operatorname{supp} \psi$ such that $\|z - v\| = 2\,\epsilon$. By definition of the support, it follows that every open neighborhood of v has a positive measure with respect to π. Since $\mathbb{R}^n \setminus \overline{\mathbf{B}\left(z, \epsilon\right)}$ is an open neighborhood of v, the result of the lemma follows. □

5 Stability of the σ-normalized Markov chain $\{\boldsymbol{Z_k}\,;\boldsymbol{k} \in \mathbb{N}\}$

Assuming that f is a strictly increasing transformation of either a C^1 scaling-invariant function with a unique global argmin or a nontrivial linear function, we prove that if Assumptions A1−A5 are satisfied and the expected logarithm of the step-size increases on nontrivial linear functions, then the σ-normalized Markov chain is a φ-irreducible aperiodic T-chain that is geometrically ergodic. In particular, it is positive and Harris recurrent.

5.1 Irreducibility, aperiodicity and T-chain property of the σ-normalized Markov chain

Prior to establishing Harris recurrence and positivity of the chain $\{Z_k\,; k \in \mathbb{N}\}$, we need to establish the φ-irreducibility and aperiodicity as well as identify some small sets such that drift conditions can be used. Since the step-size change is a deterministic function of the random input used to update the mean, we use the tools reminded in Section 4.3 to establish these properties. The chain investigated satisfies $Z_{k+1} = F_w\left(Z_k, \alpha_f\left(x^\star + Z_k, U_{k+1}\right)\right)$ and therefore fits the model (25). We prove next that the necessary assumptions needed to use the tools are satisfied if f satisfies F1 or F2 because if f is a continuous scaling-invariant function with Lebesgue negligible level sets, then for all $z \in \mathbb{R}^n$, the random variable $\alpha_f(x^\star + z, U_1)$ admits a probability density function p_z^f such that $(z, u) \mapsto p_z^f(u)$ is lower semi-continuous [19, Proposition 5.2], i.e. B4 is satisfied.

Proposition 4 *Let f be scaling-invariant with respect to $x^\star$ defined as $\varphi \circ g$ where φ is strictly increasing and g is a continuous scaling-invariant function with Lebesgue negligible level sets. Let $\{Z_k\,; k \in \mathbb{N}\}$ be a σ-normalized Markov chain associated to the step-size adaptive $(\mu/\mu_w, \lambda)$-ES defined as in Proposition 2 satisfying $Z_{k+1} = F_w\left(Z_k, \alpha_f\left(x^\star + Z_k, U_{k+1}\right)\right)$. Then model (25) follows. In addition, if Assumption A1 is satisfied, then F_w is C^1 and thus B5 is satisfied. If Assumption A5 is satisfied, then Assumptions B1−B4 are satisfied and the probability density function of the random variable $\alpha_f(x^\star + z, U_{k+1})$ denoted by p_z^f and defined in (14) satisfies $(z, u) \mapsto p_z^f(u)$ is lower semi-continuous.*

In particular, if f satisfies F1 or F2, the assumption above on f holds such that the conclusions above are valid.

Proof. It follows from (17) that $\{Z_k\,; k \in \mathbb{N}\}$ is a homogeneous Markov chain following model (25). By (16), F_w is of class C^1 (B5 is satisfied) if A1 is satisfied ($\Gamma : \mathbb{R}^{n\mu} \to \mathbb{R}_+ \setminus \{0\}$ is C^1). If A5 is satisfied, then B1−B3 are also satisfied.

For all $z \in \mathbb{R}^n$, $\alpha_g(x^\star + z, U_{k+1})$ has a probability density function p_z^g such that $(z, u) \mapsto p_z^g(u)$ is lower semi-continuous [19, Proposition 5.2], and defined for all $z \in \mathbb{R}^n$ and $u \in \mathbb{R}^{n\mu}$ as in (14). With Lemma 1, $\alpha_f = \alpha_g$ and then B4 holds.

A nontrivial linear function is a continuous scaling-invariant function with Lebesgue negligible level sets. Also f still has Lebesgue negligible level sets in the case where it is a C^1 scaling-invariant function with a unique global argmin [66, Proposition 4.2]. □

We show in the following lemma the density of the control set in $\mathbb{R}^{n\mu}$ when the objective functions are strictly increasing transformations of continuous scaling-invariant functions with Lebesgue negligible level sets, especially for functions f that satisfy F1 or F2. This is useful for Proposition 5 and for the application of Lemma 5.

Lemma 7 *Let f be a scaling-invariant function defined as $\varphi \circ g$ where φ is strictly increasing and g is a continuous scaling-invariant function with Lebesgue negligible level sets. Assume that $\{Z_k\,; k \in \mathbb{N}\}$ is the σ-normalized Markov chain associated to a step-size adaptive $(\mu/\mu_w, \lambda)$-ES as defined in Proposition 2 such that A5 is satisfied. Then for all $z \in \mathbb{R}^n$, the control set $\mathcal{O}_z^1 = \left\{v \in \mathbb{R}^{n\mu}\,;\, p_z^f(v) > 0\right\}$ is dense in $\mathbb{R}^{n\mu}$.*

In particular, if f satisfies F1 or F2, the assumption above on f holds and thus the conclusions above are valid.

Proof. By Proposition 4, we obtain that for all $z \in \mathbb{R}^n$, p_z^f is defined as in (14). In addition, $\alpha_f = \alpha_g$ (see Lemma 1). Therefore $p_z^f = p_z^g > 0$ almost everywhere. Hence we have that $\mathcal{O}_z^1$ is dense in $\mathbb{R}^{n\mu}$. □

Thanks to Theorem 4, to ensure that $\{Z_k\,;\, k \in \mathbb{N}\}$ is a φ-irreducible aperiodic T-chain, we prove that 0 is a steadily attracting state and that there exists $w \in \mathcal{O}_0^1$ such that $\operatorname{rank}\left(C_0^1(w)\right) = n$. We start with the steady attractivity in the next proposition.

Proposition 5 *Let f be a scaling-invariant function defined as $\varphi \circ g$ where φ is strictly increasing and g is a continuous scaling-invariant function with Lebesgue negligible level sets. Assume that $\{Z_k\,; k \in \mathbb{N}\}$ is the σ-normalized Markov chain associated to a step-size adaptive $(\mu/\mu_w, \lambda)$-ES as defined in Proposition 2 such that Assumptions A1 and A5 are satisfied. Then 0 is a steadily attracting state of CM(F_w). Especially, if f satisfies F1 or F2, the assumption above on f holds and thus the conclusions above are valid.*

Proof. We fix $z \in \mathbb{R}^n$ and prove that there exists a sequence $\left\{z_k \in A_+^k(z)\,; k \in \mathbb{N}\right\}$ that converges to 0. We construct the sequence recursively as follows.

We define $z_0 = z$ and fix a natural integer k. We define z_{k+1} iteratively as follows. We set $\tilde{v}_k = -\frac{1}{\|w\|^2}\left(w_1 z_k, \ldots, w_\mu z_k\right)$, then $z_k + w^\top \tilde{v}_k = z_k - \frac{1}{\|w\|^2}\sum_{i=1}^{\mu} w_i^2 z_k = 0$. By continuity of F_w and density of $\mathcal{O}_{z_k}^1$ thanks to Lemma 7, there exists $v_k \in \mathcal{O}_{z_k}^1$ such that $\|F_w(z_k, v_k)\| = \|F_w(z_k, v_k) - F_w(z_k, \tilde{v}_k)\| \leq \frac{1}{2^{k+1}}$. Define $z_{k+1} = F_w(z_k, v_k)$. Then the sequence $(z_k)_{k\in\mathbb{N}}$ converges to 0. Now let us show that $z_k \in A_+^k(z)$ for all $k \in \mathbb{N}$. Since $A_+^0(z) = \{z\}$, then $z_0 = z \in A_+^0(z)$. We fix again a natural integer k and assume that $z_k \in A_+^k(z)$. It is then enough to prove that $z_{k+1} \in A_+^{k+1}(z)$. Recall that for all $\mathbf{u} \in \mathbb{R}^{n\mu(k+1)}$, $A_+^{k+1}(z) = \left\{S_z^{k+1}(\mathbf{u})\,;\mathbf{u} \in \mathcal{O}_z^{k+1}\right\}$, $S_z^{k+1}(\mathbf{u}) = F_w\left(S_z^k\left(u_1, \ldots, u_k\right), u_{k+1}\right)$, $p_z^{f,k+1}(\mathbf{u}) = p_z^{f,k}\left(u_1, \ldots, u_k\right) p^f_{S_z^k(u_1,\ldots,u_k)}(u_{k+1})$, $\mathcal{O}_z^{k+1} = \left\{\mathbf{u} \in \mathbb{R}^{n\mu(k+1)}\ ;\ p_z^{f,k+1}(\mathbf{u}) > 0\right\}$. Therefore by construction, $p_z^{f,k+1}(v_0, \ldots, v_k) =$

$p_z^{f,k}(v_0,\dots,v_{k-1})p_{z_k}^f(v_k) > 0$, hence $(v_0,\dots,v_k) \in \mathcal{O}_z^{k+1}$. Finally, $z_{k+1} = F_w(z_k, v_k) = S_z^{k+1}(v_0,\dots,v_k) \in A_+^{k+1}(z)$.

□

The next proposition ensures that the steadily attracting state 0 satisfies also the adequate full-rank condition on a controllability matrix of 0.

Proposition 6 *Let f be a scaling-invariant function defined as $\varphi \circ g$ where φ is strictly increasing and g is a continuous scaling-invariant function with Lebesgue negligible level sets. Assume that $\{Z_k\,; k \in \mathbb{N}\}$ is the σ-normalized Markov chain associated to a step-size adaptive $(\mu/\mu_w, \lambda)$-ES as defined in Proposition 2 such that Assumptions A1 and A5 are satisfied. Then there exists $w \in \mathcal{O}_0^1$ such that* $\operatorname{rank}\left(C_0^1(w)\right) = n$.

In particular, if f satisfies F1 or F2, the assumption above on f holds and thus the conclusions above are valid.

Proof. Lemma 5 along with the density of the control set $\mathcal{O}_0^1$ in Lemma 7 ensure that it is enough to prove the existence of $v \in \mathbb{R}^{n\mu}$ such that $\operatorname{rank}\left(C_0^1(v)\right) = n$. Let us show that the matrix $C_0^1(0) = \frac{\partial S_0^1}{\partial v_1}(0)$ has a full rank, with $S_0^1 : v \in \mathbb{R}^{n\mu} \mapsto F_w(0, v) \in \mathbb{R}^n$. This is equivalent to showing that the differential $DS_0^1(0) : \mathbb{R}^{n\mu} \to \mathbb{R}^n$ of S_0^1 at 0 is surjective. Denote by l the linear function $h \in \mathbb{R}^{n\mu} \mapsto \sum_{i=1}^{\mu} w_i h_i \in \mathbb{R}^n$. Then $S_0^1 = l/\Gamma$ and then $DS_0^1(h) = Dl(h)\frac{1}{\Gamma(h)} + l(h)D(\frac{1}{\Gamma})(h)$. Since $l(0) = 0$, it follows that $DS_0^1(0) = \frac{l}{\Gamma(0)}$ and finally we obtain that $DS_0^1(0)$ is surjective. □

By applying Propositions 4, 5 and 6 along with Theorem 4, we directly deduce that the σ-normalized Markov chain associated to a step-size adaptive $(\mu/\mu_w, \lambda)$-ES is a φ-irreducible aperiodic T-chain. More formally, the next proposition holds.

Proposition 7 *Let f be a scaling-invariant function defined as $\varphi \circ g$ where φ is strictly increasing and g is a continuous scaling-invariant function with Lebesgue negligible level sets. Assume that $\{Z_k\,; k \in \mathbb{N}\}$ is the σ-normalized Markov chain associated to a step-size adaptive $(\mu/\mu_w, \lambda)$-ES as defined in Proposition 2 such that Assumptions A1 and A5 are satisfied. Then $\{Z_k\,; k \in \mathbb{N}\}$ is a φ-irreducible aperiodic T-chain, and every compact set is a small set.*

In particular, if f satisfies F1 or F2, the assumption above on f holds and thus the conclusions above are valid.

5.2 Convergence in distribution of the step-size multiplicative factor

In order to prove that $\{Z_k\,; k \in \mathbb{N}\}$ satisfies a geometric drift condition, we investigate the distribution of $\{Z_k\,; k \in \mathbb{N}\}$ outside of a compact set (small set). Intuitively, when Z_k is very large, i.e. $X_k - x^\star$ large compared to the step-size σ_k, the algorithm sees the function f in a small neighborhood from $X_k - x^\star$ where f resembles a linear function (this holds under regularity conditions on the level sets of f). Formally we prove that for all $k \in \mathbb{N}$, the step-size multiplicative factor

$\Gamma\left(\alpha_f(x^\star + z, U_{k+1})\right)$ converges in distribution[3] towards the step-size change on nontrivial linear functions $\Gamma^\star_{\text{linear}}$ defined in (15), when $\|z\|$ goes to ∞.

To do so we derive in Proposition 8 an intermediate result that requires to introduce a specific nontrivial linear function l^f_z defined as follows.

We consider a scaling-invariant function f with respect to its unique global argmin $x^\star$. Then the function $\tilde{f} : x \mapsto f(x^\star + x) - f(x^\star)$ is C^1 scaling-invariant with respect to 0 which is the unique global argmin. There exists a vector in the closed unit ball $z^f_0 \in \overline{\mathbf{B}(0,1)}$ whose $\tilde{f}$-level set is included in the closed unit ball, that is $\mathcal{L}_{\tilde{f},z^f_0} \subset \overline{\mathbf{B}(0,1)}$ and such that for all $z \in \mathcal{L}_{\tilde{f},z^f_0}$, the scalar product between z and the gradient of f at $x^\star + z$ satisfies $z^\top \nabla f(x^\star + z) > 0$ [66, Corollary 4.1 and Proposition 4.10]. In addition, any half-line of origin 0 intersects the level set $\mathcal{L}_{\tilde{f},z^f_0}$ at a unique point. We denote for all $z \neq 0$ by t^f_z the unique scalar of $(0,1]$ such that $t^f_z \frac{z}{\|z\|}$ belongs to the level set $\mathcal{L}_{\tilde{f},z^f_0} \subset \overline{\mathbf{B}(0,1)}$. We finally define for all $z \neq 0$, the nontrivial linear function l^f_z for all $w \in \mathbb{R}^n$ as

$$l^f_z(w) = w^\top \nabla f\left(x^\star + t^f_z \frac{z}{\|z\|}\right). \tag{26}$$

We state below the intermediate result that when $\|z\|$ goes to ∞, the selection random vector $\alpha_f(x^\star + z, U_1)$ has asymptotically the distribution of the selection random vector on the linear function l^f_z. According to Lemma 4, the latter does not depend on the current location and is equal to the distribution of $\alpha_{l^f_z}(0, U_1)$.

Proposition 8 *Let f be a C^1 scaling-invariant function with a unique global argmin. For all $\varphi : \mathbb{R}^{n\mu} \to \mathbb{R}$ continuous and bounded, $\lim_{\|z\|\to\infty} \int \varphi(u)\left(p^f_z(u) - p^{l^f_z}_z(u)\right) \mathrm{d}u = 0$ where l^f_z is defined as in (26). In other words, the selection random vectors $\alpha_f(x^\star + z, U_1)$ and $\alpha_{l^f_z}(0, U_1)$ have asymptotically the same distribution when $\|z\|$ goes to ∞.*

Proof idea. We sketch the proof idea and refer to Appendix C for the full proof. Note beforehand that $\alpha_f(x^\star + z, U_1) = \alpha_{\tilde{f}}(z, U_1)$ so that we assume without loss of generality that $x^\star = 0$ and $f(0) = 0$. If f is a C^1 scaling-invariant function with a unique global argmin, we can construct a positive number δ_f such that for all element z of the compact set $\mathcal{L}_{f,z^f_0} + \overline{\mathbb{B}(0, 2\delta_f)}$, $z^\top \nabla f(z) > 0$ [66, Proposition 4.11]. In particular, this result produces a compact neighborhood of the level set $\mathcal{L}_{f,z^f_0}$ where ∇f does not vanish. This helps to establish the limit of $\mathbb{E}\left[\varphi(\alpha_f(z, U_1))\right]$ when $\|z\|$ goes to ∞. We prove it by exploiting the uniform continuity of a function that we obtain thanks to its continuity on the compact set $\left(\mathcal{L}_{f,z^f_0} + \overline{\mathbb{B}(0,\delta_f)}\right) \times [0, \delta_f]$ [13]. □

Thanks to Proposition 8 and Proposition 1, we can finally state in the next theorem the convergence in distribution of the step-size multiplicative factor for f satisfying F1 towards $\Gamma^\star_{\text{linear}}$ defined in (15).

[3] Recall that a sequence of real-valued random variables $\{Y_k\}_{k\in\mathbb{N}}$ converges in distribution to a random variable Y if $\lim_{k\to\infty} F_{Y_k}(x) = F_Y(x)$ for all continuity point x of F_Y, where F_{Y_k} and F_Y are respectively the cumulative distribution functions of Y_k and Y.
The Portmanteau lemma [16] ensures that $\{Y_k\}_{k\in\mathbb{N}}$ converges in distribution to Y if and only if for all bounded and continuous function φ, $\lim_{k\to\infty} \mathbb{E}\left[\varphi(Y_k)\right] = \mathbb{E}\left[\varphi(Y)\right]$.

Theorem 5 *Let f be a scaling-invariant function satisfying F1. Assume that $\{U_{k+1}\,;k\in\mathbb{N}\}$ satisfies Assumption A5, Γ is continuous and satisfies Assumption A2, i.e. Γ is invariant under rotation. Then for all natural integer k, $\Gamma\left(\alpha_f(x^\star+z,U_{k+1})\right)$ converges in distribution to $\Gamma^\star_{\text{linear}}$ defined in (15), when $\|z\|\to\infty$.*

Proof. Let $\varphi:\Gamma(\mathbb{R}^{n\mu})\to\mathbb{R}$ be continuous and bounded. It is enough to prove that $\lim_{\|z\|\to\infty}\mathbb{E}_{U_1\sim\mathcal{N}_{n\lambda}}\left[\varphi\left(\Gamma\left(\alpha_f(x^\star+z,U_1)\right)\right)\right]=\mathbb{E}\left[\varphi\left(\Gamma^\star_{\text{linear}}\right)\right]$ and apply the Portmanteau lemma. By Proposition 8, $\lim_{\|z\|\to\infty}\int\varphi\left(\Gamma(u)\right)\left(p^f_z(u)-p^{l^f_z}_z(u)\right)\mathrm{d}u=0$. Then $\lim_{\|z\|\to\infty}\mathbb{E}_{U_1\sim\mathcal{N}_{n\lambda}}\left[\varphi\left(\Gamma\left(\alpha_f(x^\star+z,U_1)\right)\right)\right]-\mathbb{E}_{U_1\sim\mathcal{N}_{n\lambda}}\left[\varphi\left(\Gamma\left(\alpha_{l^f_z}(x^\star+z,U_1)\right)\right)\right]=0$. With Proposition 1, $\mathbb{E}_{U_1\sim\mathcal{N}_{n\lambda}}\left[\varphi\left(\Gamma\left(\alpha_{l^f_z}(x^\star+z,U_1)\right)\right)\right]=\mathbb{E}\left[\varphi\left(\Gamma^\star_{\text{linear}}\right)\right]$. □

5.3 Geometric ergodicity of the σ-normalized Markov chain

The convergence in distribution of the step-size multiplicative factor while optimizing a function f that satisfies F1 proven in Theorem 5 allows us to control the behavior of the σ-normalized chain when its norm goes to ∞. More specifically, we use it to show the geometric ergodicity of $\{Z_k\,;k\in\mathbb{N}\}$ defined as in Proposition 2 for f satisfying F1 or F2. Beforehand, let us show the following proposition, which is a first step towards the construction of a geometric drift function.

Proposition 9 *Let f be a scaling-invariant function that satisfies F1 or F2 and $\{Z_k\,;k\in\mathbb{N}\}$ be the σ-normalized Markov chain associated to a step-size adaptive $(\mu/\mu_w,\lambda)$-ES defined as in Proposition 2. We assume that Γ is continuous and Assumptions A2, A3 and A5 are satisfied. Then for all $\alpha>0$, $\lim_{\|z\|\to\infty}\frac{\mathbb{E}[\|Z_1\|^\alpha|Z_0=z]}{\|z\|^\alpha}=\mathbb{E}\left[\frac{1}{[\Gamma^\star_{\text{linear}}]^\alpha}\right]$ where $\Gamma^\star_{\text{linear}}$ is the random variable defined in (15) that represents the step-size change on any nontrivial linear function.*

Proof. Let $z\neq 0$. Since $Z_1=F_w\left(Z_0,\alpha_f\left(x^\star+Z_0,U_1\right)\right)=\frac{Z_0+w^\top\alpha_f(x^\star+Z_0,U_1)}{\Gamma\left(\alpha_f(x^\star+Z_0,U_1)\right)}$, then $\mathbb{E}\left[\|Z_1\|^\alpha|Z_0=z\right]/\|z\|^\alpha-\mathbb{E}\left[1/\Gamma\left(\alpha_f(x^\star+z,U_1)\right)^\alpha\right]=\mathbb{E}\left[\frac{\left\|\frac{z}{\|z\|}+\frac{w^\top\alpha_f(x^\star+z,U_1)}{\|z\|}\right\|^\alpha-1}{\Gamma\left(\alpha_f(x^\star+z,U_1)\right)^\alpha}\right]$. The function Γ is lower bounded by $m_\Gamma>0$ thanks to Assumption A3. In addition, $\|w^\top\alpha_f\left(x^\star+z,U_1\right)\|\leq\|w\|\,\|U_1\|$. Then the term

$$\left|\left\|\frac{z}{\|z\|}+\frac{1}{\|z\|}w^\top\alpha_f(x^\star+z,U_1)\right\|^\alpha-1\right|/\Gamma\left(\alpha_f(x^\star+z,U_1)\right)^\alpha \tag{27}$$

converges almost surely towards 0 when $\|z\|$ goes to ∞, and is bounded (when $\|z\|\geq 1$) by the integrable random variable $\frac{1+(1+\|w\|\,\|U_1\|)^\alpha}{m^\alpha_\Gamma}$. Then it follows by the dominated convergence theorem that

$$\lim_{\|z\|\to\infty}\mathbb{E}\left[\|Z_1\|^\alpha|Z_0=z\right]/\|z\|^\alpha-\mathbb{E}\left[1/\Gamma\left(\alpha_f(x^\star+z,U_1)\right)^\alpha\right]=0. \tag{28}$$

Since $x\mapsto 1/x^\alpha$ is continuous and bounded on $\Gamma\left(\mathbb{R}^{n\mu}\right)\subset[m_\Gamma,\infty)$, then for f satisfying F1, Theorem 5 implies that $\lim_{\|z\|\to\infty}\mathbb{E}\left[1/\Gamma\left(\alpha_f(x^\star+z,U_1)\right)^\alpha\right]$ exists and

is equal to $\mathbb{E}\left[1/[\Gamma^\star_{\text{linear}}]^\alpha\right]$. Starting from (28) and using Proposition 1 to replace $\mathbb{E}\left[\frac{1}{\Gamma\left(\alpha_f(x^\star+z,U_1)\right)^\alpha}\right]$ by $\mathbb{E}\left[\frac{1}{[\Gamma^\star_{\text{linear}}]^\alpha}\right]$, the same conclusion holds for f satisfying F2. Thereby $\lim_{\|z\|\to\infty}\mathbb{E}\left[\|Z_1\|^\alpha|Z_0=z\right]/\|z\|^\alpha = \mathbb{E}\left[\frac{1}{[\Gamma^\star_{\text{linear}}]^\alpha}\right]$. □

We introduce the next two lemmas, that allow to go from Proposition 9 to a formulation with the multiplicative log-step-size factor.

Lemma 8 *Let f be a continuous scaling-invariant function with respect to $x^\star$ with Lebesgue negligible level sets, let $z\in\mathbb{R}^n$. Assume that Γ satisfies Assumption A4. Then $u\mapsto \log\left(\Gamma\left(\alpha_f(x^\star+z,u)\right)\right)$ is $\mathcal{N}_{n\lambda}$-integrable with*

$$\mathbb{E}_{U_1\sim\mathcal{N}_{n\lambda}}\left[\left|\log\left(\Gamma\left(\alpha_f(x^\star+z,U_1)\right)\right)\right|\right] \le \frac{\lambda!\,\mathbb{E}_{W_1\sim\mathcal{N}_{n\mu}}\left[|\log\circ\Gamma|\left(W_1\right)\right]}{(\lambda-\mu)!}. \tag{29}$$

Proof. With (14), we have $\frac{(\lambda-\mu)!}{\lambda!}\mathbb{E}\left[\left|\log\left(\Gamma\left(\alpha_f(x^\star+z,U_1)\right)\right)\right|\right] \le \int_{\mathbb{R}^n}|\log\circ\Gamma|\left(v\right)\prod_{i=1}^{\mu}p_{\mathcal{N}_n}(v^i)\mathrm{d}v = \mathbb{E}\left[|\log\circ\Gamma|\left(\mathcal{N}_{n\mu}\right)\right]$, and A4 says that $\log\circ\Gamma$ is $\mathcal{N}_{n\mu}$-integrable. □

The next lemma states that if the expected logarithm of the step-size change is positive, then we can find $\alpha>0$ such that the limit in Proposition 9 is strictly smaller than 1. This is the key lemma to have the condition in the main results expressed as $\mathbb{E}\left[\log\left(\Gamma^\star_{\text{linear}}\right)\right]>0$, instead of $\mathbb{E}\left[1/[\Gamma^\star_{\text{linear}}]^\alpha\right]<1$ for a positive α [13].

Lemma 9 *Assume that Γ satisfies Assumptions A3 and A4. If $\mathbb{E}\left[\log\left(\Gamma^\star_{\text{linear}}\right)\right]>0$, then there exists $0<\alpha<1$ such that $\mathbb{E}\left[\frac{1}{[\Gamma^\star_{\text{linear}}]^\alpha}\right]<1$, where $\Gamma^\star_{\text{linear}}$ is defined in (15).*

Proof. Lemma 8 ensures that $\log\left(\Gamma^\star_{\text{linear}}\right)$ is integrable. For $\alpha>0$, $\frac{1}{[\Gamma^\star_{\text{linear}}]^\alpha} = \exp\left[-\alpha\log\left(\Gamma^\star_{\text{linear}}\right)\right] = 1-\alpha\log\left(\Gamma^\star_{\text{linear}}\right)+o(\alpha)$. Then the random variable $A(\alpha) = \left(\frac{1}{[\Gamma^\star_{\text{linear}}]^\alpha}-1+\alpha\log\left(\Gamma^\star_{\text{linear}}\right)\right)/\alpha$ depending on the parameter α converges almost surely towards 0 when α goes to 0.

Let $u\in\mathbb{R}^{n\mu}$ and $\alpha\in(0,1)$. Define $\varphi_u : c\mapsto \frac{1}{\Gamma(u)^c} = \exp(-c\log(\Gamma(u)))$ on $[0,\alpha]$. By the mean value theorem, there exists $c_{u,\alpha}\in(0,\alpha)$ such that $\left(\frac{1}{\Gamma(u)^\alpha}-1\right)/\alpha = \varphi_u'(c_{u,\alpha}) = -\log(\Gamma(u))\frac{1}{\Gamma(u)^{c_{u,\alpha}}}$. In addition, $\frac{1}{\Gamma(u)^{c_{u,\alpha}}}\le\frac{1}{m_\Gamma^{c_{u,\alpha}}}$ thanks to Assumption A3, and $\frac{1}{m_\Gamma^{c_{u,\alpha}}} = \exp\left(-c_{u,\alpha}\log(m_\Gamma)\right)\le\exp\left(|\log(m_\Gamma)|\right)$. Therefore $|A(\alpha)|\le\left(1+\exp\left(|\log(m_\Gamma)|\right)\right)|\log\left(\Gamma^\star_{\text{linear}}\right)|$. The latter is integrable thanks to Assumption A4, and does not depend on α. Then by the dominated convergence theorem, $\mathbb{E}\left[A(\alpha)\right]$ converges to 0 when α goes to 0 or equivalently $\mathbb{E}\left[\frac{1}{[\Gamma^\star_{\text{linear}}]^\alpha}\right] = 1-\alpha\mathbb{E}\left[\log\left(\Gamma^\star_{\text{linear}}\right)\right]+o(\alpha)$. Hence there exists $0<\alpha<1$ small enough such that $\mathbb{E}\left[\frac{1}{[\Gamma^\star_{\text{linear}}]^\alpha}\right]<1$. □

We now have enough material to state and prove the desired geometric ergodicity of the σ-normalized Markov chain in the following theorem.

Theorem 6 *(Geometric ergodicity) Let f be a scaling-invariant function that satisfies F1 or F2. Let $\{Z_k\,;k\in\mathbb{N}\}$ be the σ-normalized Markov chain associated to a step-size adaptive $(\mu/\mu_w,\lambda)$-ES defined as in Proposition 2 such that Assumptions A1−A5 are satisfied. Assume that $\mathbb{E}\left[\log\left(\Gamma^\star_{\text{linear}}\right)\right]>0$ where $\Gamma^\star_{\text{linear}}$ is defined in (15).*

Then there exists $0<\alpha<1$ such that the function $V:z\mapsto 1+\|z\|^\alpha$ is a geometric drift function for the Markov chain $\{Z_k\,;k\in\mathbb{N}\}$. Therefore $\{Z_k\,;k\in\mathbb{N}\}$ is V-geometrically ergodic, admits an invariant probability measure π and is Harris recurrent.

Proof. Proposition 7 shows that $\{Z_k\,;k\in\mathbb{N}\}$ is a φ-irreducible aperiodic T-chain. With [50, Theorem 5.5.7 and Theorem 6.2.5], every compact set is a small set. Since $\mathbb{E}\left[\log\left(\Gamma^\star_{\text{linear}}\right)\right]>0$, by Lemma 9 there exists $0<\alpha<1$ such that $\mathbb{E}\left[\frac{1}{[\Gamma^\star_{\text{linear}}]^\alpha}\right]<1$. Define $V:z\mapsto 1+\|z\|^\alpha$. By Proposition 9, $\lim\limits_{\|z\|\to\infty}\mathbb{E}\left[\|Z_1\|^\alpha|Z_0=z\right]/\|z\|^\alpha = \mathbb{E}\left[1/[\Gamma^\star_{\text{linear}}]^\alpha\right]$. Since $\mathbb{E}\left[V(Z_1)|Z_0=z\right]/V(z)=\left(1+\mathbb{E}\left[\|Z_1\|^\alpha|Z_0=z\right]\right)/\left(1+\|z\|^\alpha\right)$, $\lim_{\|z\|\to\infty}\mathbb{E}\left[V(Z_1)|Z_0=z\right]/V(z)=\mathbb{E}\left[1/[\Gamma^\star_{\text{linear}}]^\alpha\right]$. Let $\gamma=\frac{1}{2}\left(1+\mathbb{E}\left[\frac{1}{[\Gamma^\star_{\text{linear}}]^\alpha}\right]\right)<1$. There exists $r>0$ such that for all $\|z\|>r$

$$\mathbb{E}\left[V(Z_1)|Z_0=z\right]/V(z)<\gamma. \tag{30}$$

In addition, since $\|z+w^\top\alpha_f\left(x^\star+z,U_1\right)\|\le\|z\|+\|w\|\|U_1\|$ then $\mathbb{E}\left[V(Z_1)|Z_0=z\right]\le 1+\mathbb{E}\left[\left(\|z\|+\|w\|\|U_1\|\right)^\alpha\right]/m_\Gamma^\alpha$. Since $z\mapsto 1+\mathbb{E}\left[\left(\|z\|+\|w\|\|U_1\|\right)^\alpha\right]/m_\Gamma^\alpha-\gamma V(z)$ is continuous on the compact $\overline{\mathbf{B}\left(0,r\right)}$, it is bounded on that compact. Denote by $b\in\mathbb{R}_+$ an upper bound. We have proven that for all $z\in\overline{\mathbf{B}\left(0,r\right)}$, $\mathbb{E}\left[V(Z_1)|Z_0=z\right]\le\gamma V(z)+b$. This result, along with (30), show that for all $z\in\mathbb{R}^n$, $\mathbb{E}\left[V(Z_1)|Z_0=z\right]\le\gamma V(z)+b\mathbb{1}_{\overline{\mathbf{B}(0,r)}}(z)$. Therefore $\{Z_k\,;k\in\mathbb{N}\}$ is V-geometrically ergodic. Then thanks to [50, Theorem 15.0.1], $\{Z_k;k\in\mathbb{N}\}$ is positive and Harris recurrent with invariant probability measure π. □

6 Main results: linear behavior as a consequence of the stability and integrability

We are now almost ready to establish the main results of the paper. Yet, we first prove in the next section the integrability of $z\mapsto\log\|z\|$ and $\mathcal{R}_f$ defined in (22), with respect to the invariant probability measure of the Markov chain $\{Z_k\,;k\in\mathbb{N}\}$ whose existence is proven in Theorem 6. We state and prove in Section 6.2 the linear behavior of the studied class of algorithms for an abstract step-size update satisfying A1-A4 on scaling invariant functions. We provide in Section 6.3 a Central Limit Theorem for approximating the convergence rate. We investigate in Section 6.4 how the CSA-ES and xNES satisfy the required conditions for a linear behavior providing sufficient conditions expressed in terms of parameters of the algorithms.

6.1 Integrabilities with respect to the invariant probability measure

For a scaling-invariant function f that satisfies F1 or F2, the limit in Theorem 7 is expressed as $\mathbb{E}_\pi(\mathcal{R}_f)$ where the function $\mathcal{R}_f$ is defined as in (22) and π is a probability measure. Therefore the π-integrability of the function $z\mapsto\mathcal{R}_f(z)$ is

necessary to obtain Theorem 7. In the following, we present a result stronger than its π-integrability, that is the boundedness of $\mathcal{R}_f$ under some assumptions.

Proposition 10 *Let f be a continuous scaling-invariant function with Lebesgue negligible level sets. Let $\{(X_k, \sigma_k)\,; k \in \mathbb{N}\}$ be the sequence defined in (5) and (6) such that Assumptions A4 and A5 are satisfied. Then the function $z \mapsto \left|\mathcal{R}_f\right|(z)$ is bounded by $\frac{\lambda!}{(\lambda-\mu)!}\mathbb{E}_{W\sim\mathcal{N}_{n\mu}}\left[\left|\log \circ\, \Gamma\right|(W)\right]$, where the function $z \mapsto \mathcal{R}_f(z)$ is defined as in (22).*

If in addition the following holds: (i) f satisfies F1 or F2, (ii) Assumptions A1−A3 are satisfied and (iii) the expected log step-size change satisfies $\mathbb{E}\left[\log\left(\Gamma^\star_{\text{linear}}\right)\right] > 0$ where $\Gamma^\star_{\text{linear}}$ is defined in (15), then $\mathbb{E}_\pi(\left|\mathcal{R}_f\right|) = \int \left|\mathcal{R}_f(z)\right| \pi(dz) < \infty$ that is $z \mapsto \mathcal{R}_f(z)$ is π-integrable where π is the invariant probability measure of $\{Z_k\,; k \in \mathbb{N}\}$ defined as in Proposition 2.

Proof. Lemma 8 shows that for all $z \in \mathbb{R}^n$, $z \mapsto \log\left(\Gamma\left(\alpha_f(x^\star + z, u)\right)\right)$ is $\mathcal{N}_{n\lambda}$-integrable with $\mathbb{E}\left[\left|\log\left(\Gamma\left(\alpha_f(x^\star + z, U_1)\right)\right)\right|\right] \leq \dfrac{\lambda!}{(\lambda-\mu)!}\mathbb{E}_{W\sim\mathcal{N}_{n\mu}}\left[\left|\log \circ\, \Gamma\right|(W)\right]$. Then $\left|\mathcal{R}_f\right|$ is bounded since $\left|\mathcal{R}_f(z)\right| \leq \frac{\lambda!}{(\lambda-\mu)!}\mathbb{E}_{W\sim\mathcal{N}_{n\mu}}\left[\left|\log \circ\, \Gamma\right|(W)\right]$ for all $z \in \mathbb{R}^n$. If in addition Assumptions A1−A3 are satisfied and $\mathbb{E}\left[\log\left(\Gamma^\star_{\text{linear}}\right)\right] > 0$, Theorem 6 ensures that $\{Z_k\,; k \in \mathbb{N}\}$ is a positive Harris recurrent chain with invariant probability measure π. Hence the integrability with respect to π. □

We prove in the following the π-integrability of $z \mapsto \log\|z\|$, where π is the invariant probability measure of the σ-normalized chain, under some assumptions.

Proposition 11 *Let f satisfy F1 or F2 and $\{Z_k\,; k \in \mathbb{N}\}$ be the Markov chain defined as in Proposition 2 such that Assumptions A1−A5 are satisfied. Assume that $\mathbb{E}\left[\log\left(\Gamma^\star_{\text{linear}}\right)\right] > 0$ where $\Gamma^\star_{\text{linear}}$ is defined in (15). Then $\{Z_k\,; k \in \mathbb{N}\}$ has an invariant probability measure π and $z \mapsto \log\|z\|$ is π-integrable.*

Proof. Theorem 6 ensures that $\{Z_k\,; k \in \mathbb{N}\}$ is V-geometrically ergodic with invariant probability measure π, where $V : z \in \mathbb{R}^n \mapsto 1 + \|z\|^\alpha \in \mathbb{R}_+$. We define for all $z \in \mathbb{R}^n$, $g(z) = \frac{(\lambda-\mu)!}{2\lambda!}\left|\log\|z\|\right|$. The π-integrability of g is obtained if there exist a set A with $\pi(A) > 0$ such that $\displaystyle\int_A g(z)\pi(\mathrm{d}z) < \infty$, and a measurable function h with $h\mathbb{1}_{A^c} \geq g\mathbb{1}_{A^c}$ such that (i) $\displaystyle\int_{A^c} P(z, \mathrm{d}y)h(y) < h(z) - g(z)\,, \forall z \in A^c$ and (ii) $\displaystyle\sup_{z\in A}\int_{A^c} P(z, \mathrm{d}y)h(y) < \infty$ [67, Theorem 1]. For $z \in \mathbb{R}^n$ and $v \in \mathbb{R}^{n\mu}$, denote $\varphi(z, v)$ as $\varphi(z, v) = p_{\mathcal{N}_{n\mu}}\left(v - \frac{1}{\|w\|^2}\left(w_1 z, \ldots, w_\mu z\right)\right)\mathbb{1}_{\|w^\top v\|\leq 1}$. We prove in a first time that $\displaystyle\lim_{\|z\|\to 0}\int \left|\log\|w^\top v\|\right|\varphi(z, v)\mathrm{d}v < \infty$. We have $(2\pi)^{n\mu/2}\varphi(z, v)$ which is equal to $\exp\left(\frac{1}{2}\left(-\|v\|^2 - \frac{\|w\|^2\|z\|^2}{\|w\|^4} + \frac{2(w^\top v)^\top z}{\|w\|^2}\right)\right)\mathbb{1}_{\|w^\top v\|\leq 1}$ which is smaller than $\exp\left(\frac{1}{2}\left(-\|v\|^2 - \frac{\|w\|^2\|z\|^2}{\|w\|^4} + \frac{2\|w^\top v\|\|z\|}{\|w\|^2}\right)\right)\mathbb{1}_{\|w^\top v\|\leq 1}$. Then for $z \in \mathbb{R}^n$ and $v \in \mathbb{R}^{n\mu}$,

$$(2\pi)^{n\mu/2}\varphi(z, v) \leq \exp\left(\frac{1}{2}\left(-\|v\|^2 - \frac{\|w\|^2\|z\|^2}{\|w\|^4} + \frac{2\|z\|}{\|w\|^2}\right)\right)\mathbb{1}_{\|w^\top v\|\leq 1}. \tag{31}$$

Therefore for $z \in \overline{\mathbf{B}\left(0, 1\right)}$ and $v \in \mathbb{R}^{n\mu}$, $\varphi(z, v) \leq \exp\left(\frac{1}{\|w\|^2}\right)\varphi(0, v)$.

Since $v \mapsto \left|\log \|w^\top v\|\right| \varphi(0,v)$ is Lebesgue integrable, it follows by the dominated convergence theorem that $z \mapsto \int \left|\log \|w^\top v\|\right| \varphi(z,v)\mathrm{d}v$ is continuous on $\overline{\mathbf{B}(0,1)}$ and $\lim_{\|z\|\to 0} \int \left|\log \|w^\top v\|\right| \varphi(z,v)\mathrm{d}v < \infty$. In addition, $\lim_{\|z\|\to 0} g(z) = \infty$. Then there exists $\epsilon_1 \in (0,1)$ such that for $z \in \overline{\mathbf{B}(0,\epsilon_1)}$:

$$\int \left|\log \|w^\top v\|\right| \varphi(z,v)\mathrm{d}v + 2\,\mathbb{E}_{W\sim\mathcal{N}_{n\mu}}\left[|\log\circ\Gamma|(W)\right] \leq g(z). \tag{32}$$

We define ϵ_2 from Lemma 6 and denote $\epsilon = \min(\epsilon_1, \epsilon_2)$. Define $A = \mathbb{R}^n \setminus \overline{\mathbf{B}(0,\epsilon)}$. Then from Lemma 6 it follows that $\pi(A) > 0$. Note also that $A^c = \overline{\mathbf{B}(0,\epsilon)}$. In addition, g is dominated by the π-integrable function V around ∞, then $\int_A g(z)\pi(\mathrm{d}z) < \infty$. We define now the function h for all $z \in \mathbb{R}^n$ as $h(z) = 2g(z)\mathbb{1}_{A^c}(z)$. Then $h\mathbb{1}_{A^c} \geq g\mathbb{1}_{A^c}$. It remains to verify the items (i) and (ii) from above to obtain the π-integrability of g. We give in the following an upper bound of $K(z) = \int_{A^c} P(z,\mathrm{d}y)h(y) = -\frac{(\lambda-\mu)!}{\lambda!}\mathbb{E}_z\left[\mathbb{1}_{\overline{\mathbf{B}(0,\epsilon)}}(Z_1)\log\|Z_1\|\right]$. We have $K(z) \leq -\frac{(\lambda-\mu)!}{\lambda!}\int_{\|z+w^\top v\|\leq\Gamma(v)} \log\frac{\|z+w^\top v\|}{\Gamma(v)} p_z^f(v)\mathrm{d}v$. With (14), $\frac{(\lambda-\mu)!}{\lambda!}p_z^f \leq p_{\mathcal{N}_{n\mu}}$. Then $K(z) \leq \int |\log(\Gamma(v))|\, p_{\mathcal{N}_{n\mu}}(v)\mathrm{d}v + \int_{\|z+w^\top v\|\leq\Gamma(v)} \left|\log\|z+w^\top v\|\right| p_{\mathcal{N}_{n\mu}}(v)\mathrm{d}v$. We split the latter integral between the events $\left\{\|z+w^\top v\| \leq \min(1,\Gamma(v))\right\}$ and the events $\left\{1 < \|z+w^\top v\| \leq \Gamma(v)\right\}$. Then $K(z) \leq \int_{\|z+w^\top v\|\leq\min(1,\Gamma(v))} \left|\log\|z+w^\top v\|\right| p_{\mathcal{N}_{n\mu}}(v)\mathrm{d}v + \int_{\Gamma(v)\geq 1} \log(\Gamma(v)) p_{\mathcal{N}_{n\mu}}(v)\mathrm{d}v + \int |\log(\Gamma(v))|\, p_{\mathcal{N}_{n\mu}}(v)\mathrm{d}v$. Hence $K(z) \leq 2\,\mathbb{E}_{W\sim\mathcal{N}_{n\mu}}\left[|\log\circ\Gamma|(W)\right] - \int_{\|z+w^\top v\|\leq 1} \log\|z+w^\top v\| p_{\mathcal{N}_{n\mu}}(v)\mathrm{d}v$. With a translation $v \to v - \frac{1}{\|w\|^2}(w_1 z, \ldots, w_\mu z)$ within the last integrand, we obtain:

$$K(z) \leq 2\,\mathbb{E}_{W\sim\mathcal{N}_{n\mu}}\left[|\log\circ\Gamma|(W)\right] + \int \left|\log\|w^\top v\|\right| \varphi(z,v)\mathrm{d}v. \tag{33}$$

Equations (32) and (33) show that for $z \in A^c = \overline{\mathbf{B}(0,\epsilon)}$, $\int_{A^c} P(z,\mathrm{d}y)h(y) \leq g(z) = h(z) - g(z)$. Therefore the item (i) follows. With (31), it follows that there exist $c_1 > 0$ and $c_2 > 0$ such that for $\|z\| \geq c_1$ and $v \in \mathbb{R}^n$, $\varphi(z,v) \leq c_2\varphi(0,v)$. Thanks to the dominated convergence theorem, $\lim_{\|z\|\mapsto\infty} \int \left|\log\|w^\top v\|\right| \varphi(z,v)\mathrm{d}v = 0$. Therefore that integral is bounded outside of a compact. In addition, $z \mapsto \int \left|\log\|w^\top v\|\right| \varphi(z,v)\mathrm{d}v$ is continuous and is bounded on any compact included in $\overline{A}$. Then along with (33) it follows that $\sup_{z\in A}\int_{A^c} P(z,\mathrm{d}y)h(y) < \infty$. Hence the item (ii) is also satisfied, which ends the integrability proof of $z \mapsto \log\|z\|$. □

6.2 Linear behavior for an abstract step-size update

We are now ready to establish the linear behavior of the $(\mu/\mu_w, \lambda)$-ES. Our condition for the linear behavior stemming from the drift condition for geometric ergodicity established in Theorem 6 is that the expected logarithm of the step-size change function Γ on a nontrivial linear function is positive. By Proposition 1, when f satisfies F2, the expected change of the logarithm of the step-size is constant and for all z, $\mathcal{R}_f(z) = \mathcal{R}_f(-x^\star) = \mathbb{E}\left[\log\left(\Gamma^\star_{\text{linear}}\right)\right]$ where $\Gamma^\star_{\text{linear}}$ is defined in (15). Our main result states that if the expected logarithm of the step-size increases on nontrivial linear functions, i.e. if $\mathbb{E}\left[\log\left(\Gamma^\star_{\text{linear}}\right)\right] > 0$, then almost sure linear behavior holds on functions satisfying F1 or F2. If f satisfies F2, then almost sure linear divergence holds with a divergence rate of $\mathbb{E}\left[\log\left(\Gamma^\star_{\text{linear}}\right)\right]$. More precisely the following results hold.

Theorem 7 *Let f be a scaling-invariant function with respect to $x^\star$. Assume that f satisfies F1 (in which case $x^\star$ is the global optimum) or F2. Let $\{(X_k, \sigma_k)\,; k \in \mathbb{N}\}$ be the sequence defined in (5) and (6) such that Assumptions A1−A5 are satisfied. Let $\{Z_k = (X_k - x^\star)/\sigma_k\,; k \in \mathbb{N}\}$ be the σ-normalized Markov chain (Proposition 2). If the expected logarithm of the step-size increases on nontrivial linear functions, i.e. if $\mathbb{E}\left[\log\left(\Gamma^\star_{\text{linear}}\right)\right] > 0$ where $\Gamma^\star_{\text{linear}}$ is defined in (15), then $\{Z_k\,; k \in \mathbb{N}\}$ admits an invariant probability measure π such that $\mathcal{R}_f$ defined in (22) is π-integrable. And for all $(X_0, \sigma_0) \in \left(\mathbb{R}^n \setminus \{x^\star\}\right) \times (0, \infty)$, linear behavior of X_k and σ_k as in (18) holds almost surely with*

$$\lim_{k\to\infty} \frac{1}{k} \log \frac{\|X_k - x^\star\|}{\|X_0 - x^\star\|} = \lim_{k\to\infty} \frac{1}{k} \log \frac{\sigma_k}{\sigma_0} = \mathbb{E}_\pi(\mathcal{R}_f) \ . \tag{34}$$

In addition, for all initial conditions $(X_0, \sigma_0) = (x, \sigma) \in \mathbb{R}^n \times (0, \infty)$, the expected log-progress behaves linearly with

$$\lim_{k\to\infty} \mathbb{E}_{\frac{x-x^\star}{\sigma}}\left[\log \frac{\|X_{k+1} - x^\star\|}{\|X_k - x^\star\|}\right] = \lim_{k\to\infty} \mathbb{E}_{\frac{x-x^\star}{\sigma}}\left[\log \frac{\sigma_{k+1}}{\sigma_k}\right] = \mathbb{E}_\pi(\mathcal{R}_f) \ . \tag{35}$$

If f satisfies F2, then $\mathcal{R}_f$ is constant equal to $\mathbb{E}_\pi(\mathcal{R}_f) = \mathbb{E}\left[\log\left(\Gamma^\star_{\text{linear}}\right)\right] > 0$, and then both X_k and σ_k diverge to infinity with a divergence rate of $\mathbb{E}\left[\log\left(\Gamma^\star_{\text{linear}}\right)\right]$.

If $\mathbb{E}_\pi(\mathcal{R}_f) < 0$, then X_k converges linearly to the global optimum $x^\star$ with a convergence rate of $-\mathbb{E}_\pi(\mathcal{R}_f)$ and the step-size converges linearly to zero.

Proof. Theorem 6 ensures that $\{Z_k\,; k \in \mathbb{N}\}$ is a positive Harris recurrent chain with invariant probability measure π. We start from (21). Since $z \mapsto \log \|z\|$ is π-integrable, Theorem 1 ensures that the LLN holds with $\lim_{k\to\infty} \frac{1}{k} \sum_{t=0}^{k-1} \log \frac{\|Z_{t+1}\|}{\|Z_t\|} = \int \log\left(\|z\|\right) \pi(\mathrm{d}z) - \int \log\left(\|z\|\right) \pi(\mathrm{d}z) = 0$.

Let us consider the chain $\{W_k = (Z_k, U_{k+2})\,; k \in \mathbb{N}\}$. Then thanks to Proposition 3, $\{W_k = (Z_k, U_{k+2})\,; k \in \mathbb{N}\}$ is geometrically ergodic with invariant probability measure $\pi \times \mathcal{N}_{n\lambda}$. Define the function g for $((z_1, u_3), (z_2, u_4)) \in \left(\mathbb{R}^n \times \mathbb{R}^{n\lambda}\right)^2$ as $g\left((z_1, u_3), (z_2, u_4)\right) = \log(\Gamma(\alpha_f(x^\star + z_2, u_3)))$. We have by Proposition 10 that for all natural integer t, $\mathbb{E}_{\pi\times\mathcal{N}_{n\lambda}}(|g(W_t, W_{t+1})|) \leq \frac{\lambda!}{(\lambda-\mu)!} \mathbb{E}_{Y\sim\mathcal{N}_{n\mu}}\left[|\log \circ \Gamma|\,(Y)\right] < \infty$. By Theorem 3 or Corollary 1, for any initial distribution, $\frac{1}{k} \sum_{t=0}^{k-1} \log(\Gamma(\alpha_f(x^\star + Z_t, U_{t+1})))$ converges almost surely towards $\mathbb{E}_{\pi\times\mathcal{N}_{n\lambda}}(g(W_1, W_2)) = \mathbb{E}_\pi(\mathcal{R}_f)$.

Let us prove now (35). Equation (19) implies that for all $z \in \mathbb{R}^n$

$$\begin{aligned}\mathbb{E}_z\left[\log\frac{\|X_{k+1}-x^\star\|}{\|X_k-x^\star\|}\right] &= \mathbb{E}_z\left[\log\frac{\|Z_{k+1}\|}{\|Z_k\|}\right] + \mathbb{E}_z\left[\log\frac{\sigma_{k+1}}{\sigma_k}\right] \\ &= \int P^{k+1}(z,\mathrm{d}y)\log(\|y\|) - \int P^k(z,\mathrm{d}y)\log(\|y\|) \\ &\quad + \int P^k(z,\mathrm{d}y)\mathcal{R}_f(y).\end{aligned}$$

Define h on $\mathbb{R}^n$ as $h(z) = 1 + |\log\|z\||$ for all $z \in \mathbb{R}^n$ which is π-integrable thanks to Proposition 11. Then $z \mapsto \log\|z\|$ is π-integrable, and for $z \in \{y \in \mathbb{R}^n\,; V(y) < \infty\} = \mathbb{R}^n$, $\lim_{k\to\infty} \|P^k(z,\cdot) - \pi\|_h = 0$ [50, Theorem 14.0.1]. Then $\lim_{k\to\infty} \int P^{k+1}(z,\mathrm{d}y)\log(\|y\|) = \lim_{k\to\infty} \int P^k(z,\mathrm{d}y)\log(\|y\|) = \int \log(\|y\|)\pi(\mathrm{d}y)$. In addition, $\left|\mathcal{R}_f\right|/h$ is bounded, then $\lim_{k\to\infty} \int P^k(z,\mathrm{d}y)\mathcal{R}_f(y) = \int \mathcal{R}_f(y)\pi(\mathrm{d}y) = \mathbb{E}_\pi(\mathcal{R}_f)$, and finally (35) follows. We also note that if f satisfies F2, then thanks to Proposition 1, for all $z \in \mathbb{R}^n$, $\mathcal{R}_f(z) = \mathbb{E}\left[\log\left(\Gamma^\star_{\text{linear}}\right)\right]$, hence $\mathcal{R}_f$ is constant. Then $\mathbb{E}_\pi\left(\mathcal{R}_f\right) = \int \mathcal{R}_f(z)\pi(\mathrm{d}z) = \mathbb{E}\left[\log\left(\Gamma^\star_{\text{linear}}\right)\right]$. If in addition $\mathbb{E}\left[\log\left(\Gamma^\star_{\text{linear}}\right)\right] > 0$, we obtain that $\|X_k\|$ and σ_k both diverge to ∞ when k goes to ∞. □

The result that both the step-size and log distance converge (resp. diverge) to the optimum (resp. to ∞) at the same rate is noteworthy and directly follows from our theory. In addition, we provide the exact expression of the rate. Yet it is expressed using the stationary distribution of the Markov chain $\{Z_k\,; k \in \mathbb{N}\}$ for which we know little information. From a practical perspective, while we never know the optimum of a function on a real problem, (34) suggests that we can track the evolution of the step-size to define a termination criterion based on the tolerance of the x-values.

6.3 Central Limit Theorem

The rate of convergence (or divergence) of a step-size adaptive $(\mu/\mu_w, \lambda)$-ES given in (34) is expressed as $\left|\mathbb{E}_\pi(\mathcal{R}_f)\right|$ where π is the invariant probability measure of the σ-normalized Markov chain and $\mathcal{R}_f$ is defined in (22). Yet we do not have an explicit expression for π and thus of $\mathbb{E}_\pi(\mathcal{R}_f)$. However, we can approximate $\mathbb{E}_\pi(\mathcal{R}_f)$ with Monte Carlo simulations. We present a central limit theorem for the approximation of $\mathbb{E}_\pi(\mathcal{R}_f)$ as $\frac{1}{t}\sum_{k=0}^{t-1}\mathcal{R}_f(Z_k)$ where $\{Z_k; k \in \mathbb{N}\}$ is the homogeneous Markov chain defined in Proposition 2.

Theorem 8 *(Central limit theorem for the expected logarithm of the step-size) Let f be a scaling-invariant function with respect to $x^\star$ that satisfies F1 or F2. Let $\{(X_k, \sigma_k)\,; k \in \mathbb{N}\}$ be the sequence defined in (5) and (6) such that Assumptions A1–A5 are satisfied. If the expected logarithm of the step-size increases on nontrivial linear functions, i.e. if $\mathbb{E}\left[\log\left(\Gamma^\star_{\text{linear}}\right)\right] > 0$ where $\Gamma^\star_{\text{linear}}$ is defined in (15), then $\{Z_k = (X_k - x^\star)/\sigma_k\,; k \in \mathbb{N}\}$ is a Markov chain admitting an invariant probability measure π. Define $\mathcal{R}_f$ as in (22) and for all positive integer t, define $S_t(\mathcal{R}_f) = \sum_{k=0}^{t-1}\mathcal{R}_f(Z_k)$. Then the constant γ^2*

defined as
$\mathbb{E}_\pi\left[\left(\mathcal{R}_f(Z_0)-\mathbb{E}_\pi(\mathcal{R}_f)\right)^2\right]+2\sum_{k=1}^{\infty}\mathbb{E}_\pi\left[\left(\mathcal{R}_f(Z_0)-\mathbb{E}_\pi(\mathcal{R}_f)\right)\left(\mathcal{R}_f(Z_k)-\mathbb{E}_\pi(\mathcal{R}_f)\right)\right]$ *is well defined, non-negative, finite and* $\lim_{t\to\infty}\frac{1}{t}\mathbb{E}_\pi\left[\left(S_t(\mathcal{R}_f)-t\,\mathbb{E}_\pi(\mathcal{R}_f)\right)^2\right]=\gamma^2$.

If $\gamma^2>0$, *then the central limit theorem holds in the sense that for any initial condition* z_0, $\sqrt{\frac{t}{\gamma^2}}\left(\frac{1}{t}S_t(\mathcal{R}_f)-\mathbb{E}_\pi(\mathcal{R}_f)\right)$ *converges in distribution to* $\mathcal{N}(0,1)$. *If* $\gamma^2=0$, *then* $\lim_{t\to\infty}\frac{S_t(\mathcal{R}_f)-t\mathbb{E}_\pi(\mathcal{R}_f)}{\sqrt{t}}=0$ *a.s.*

Proof. Thanks to Proposition 10, $\left|\mathcal{R}_f\right|$ is bounded. And then there exists a positive constant K large enough such that $\mathcal{R}_f^2\leq KV$ where V is the geometric drift function of $\{Z_k\,;k\in\mathbb{N}\}$ given by Theorem 6. Then KV remains a geometric drift function. Thanks to Theorem 2, the constant γ defined as $\mathbb{E}_\pi\left[\left(\mathcal{R}_f(Z_0)-\mathbb{E}_\pi(\mathcal{R}_f)\right)^2\right]+2\sum_{k=1}^{\infty}\mathbb{E}_\pi\left[\left(\mathcal{R}_f(Z_0)-\mathbb{E}_\pi(\mathcal{R}_f)\right)\left(\mathcal{R}_f(Z_k)-\mathbb{E}_\pi(\mathcal{R}_f)\right)\right]$ is well defined, non-negative, finite and $\lim_{t\to\infty}\frac{1}{t}\mathbb{E}_\pi\left[\left(S_t(\mathcal{R}_f)-t\mathbb{E}_\pi(\mathcal{R}_f)\right)^2\right]=\gamma^2$. Moreover if $\gamma^2>0$, then the CLT holds for any z_0 as follows $\lim_{t\to\infty}P_{z_0}\left((t\gamma^2)^{-\frac{1}{2}}\left(S_t(\mathcal{R}_f)-t\mathbb{E}_\pi(\mathcal{R}_f)\right)\leq z\right)=\int_{-\infty}^{z}\frac{1}{\sqrt{2\pi}}e^{-u^2/2}\mathrm{d}u$. Which can be rephrased as $\frac{1}{\sqrt{t\gamma^2}}\left(S_t(\mathcal{R}_f)-t\mathbb{E}_\pi(\mathcal{R}_f)\right)$ converges in distribution to $\mathcal{N}(0,1)$ when $t\to\infty$. And if $\gamma=0$, then $\lim_{t\to\infty}(S_t(\mathcal{R}_f)-t\,\mathbb{E}_\pi(\mathcal{R}_f))/\sqrt{t}=0$ *a.s.* □

6.4 Sufficient conditions for the linear behavior of the $(\mu/\mu_w,\lambda)$-CSA1-ES and the $(\mu/\mu_w,\lambda)$-xNES

Theorems 7 and 8 hold for an abstract step-size update function Γ that satisfies Assumptions A1−A4. For the step-size update functions of the $(\mu/\mu_w,\lambda)$-CSA1-ES and the $(\mu/\mu_w,\lambda)$-xNES defined in (10) and (11), sufficient and necessary conditions to obtain a step-size increase on linear functions are presented in the next proposition. They are expressed using the weights and the μ best order statistics $\mathcal{N}^{1:\lambda},\ldots,\mathcal{N}^{\mu:\lambda}$ of a sample of λ standard normal distributions $\mathcal{N}^1,\ldots,\mathcal{N}^\lambda$ defined such as $\mathcal{N}^{1:\lambda}\leq\mathcal{N}^{2:\lambda}\leq\cdots\leq\mathcal{N}^{\lambda:\lambda}$.

Proposition 12 (Necessary and sufficient condition for step-size increase on nontrivial linear functions) *For the* $(\mu/\mu_w,\lambda)$*-CSA-ES algorithm without cumulation,* $\mathbb{E}\left[\log\left((\Gamma_{\mathrm{CSA1}})^\star_{\mathrm{linear}}\right)\right]=\frac{1}{2d_\sigma n}\left(\mathbb{E}\left[\left(\sum_{i=1}^{\mu}\frac{w_i}{\|w\|}\mathcal{N}^{i:\lambda}\right)^2\right]-1\right)$. *Therefore, the expected logarithm of the step-size increases on nontrivial linear functions if and only if* $\mathbb{E}\left[\left(\sum_{i=1}^{\mu}\frac{w_i}{\|w\|}\mathcal{N}^{i:\lambda}\right)^2\right]>1$.

For the $(\mu/\mu_w,\lambda)$*-xNES without covariance matrix adaptation, if* $w_i\geq 0$ *for all* $i=1,\ldots,\mu$, $\mathbb{E}\left[\log\left((\Gamma_{\mathrm{xNES}})^\star_{\mathrm{linear}}\right)\right]=\frac{1}{2d_\sigma n}\left(\sum_{i=1}^{\mu}\frac{w_i}{\sum_{j=1}^{\mu}w_j}\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]-1\right)$. *Therefore, the expected logarithm of the step-size increases on nontrivial linear functions if and*

only if $\sum_{i=1}^{\mu} \frac{w_i}{\sum_{j=1}^{\mu} w_j} \mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right] > 1$. *In addition, this latter equation is satisfied if* λ, μ *and* w *are set such that* $\lambda \geq 3$, $\mu < \frac{\lambda}{2}$ *and* $w_1 \geq w_2 \geq \cdots \geq w_\mu \geq 0$.

Proof. We first prove the statement related to the $(\mu/\mu_w, \lambda)$-CSA1-ES. Then we show the condition regarding the $(\mu/\mu_w, \lambda)$-xNES. Finally we prove the general practical condition that allows to obtain the condition regarding the xNES algorithm.

If m is a positive integer and $u = \left(u^1, \ldots, u^m\right) \in \mathbb{R}^{nm}$, we denote $u_1 = \left(u_1^1, \ldots, u_1^m\right)$ and $u_{-1} = \left(u_{-1}^1, \ldots, u_{-1}^m\right)$ where $u_{-1}^i = \left(u_2^i, \ldots, u_n^i\right)$ for $i = 1, \ldots, m$. Define the nontrivial linear function $l^\star$ such that $l^\star(x) = x_1$ for $x \in \mathbb{R}^n$, and denote by e_1 the unit vector $(1, \ldots, 0)$.

Part 1. We prove that $\mathbb{E}_{U_1 \sim \mathcal{N}_{n\lambda}}\left[\log\left(\Gamma_{\text{CSA1}}\left(\alpha_{l^\star}(e_1, U_1)\right)\right)\right]$ has the same sign than $\mathbb{E}\left[\left(\sum_{i=1}^{\mu} \frac{w_i}{\|w\|} \mathcal{N}^{i:\lambda}\right)^2\right] - 1$, and apply Theorem 7. We have
$2d_\sigma \|w\|^2 n \mathbb{E}_{U_1 \sim \mathcal{N}_{n\lambda}}\left[\log\left(\Gamma_{\text{CSA1}}\left(\alpha_{l^\star}(e_1, U_1)\right)\right)\right] =$
$\left(\mathbb{E}_{U_1 \sim \mathcal{N}_{n\lambda}}\left[\| \sum_{i=1}^{\mu} w_i \left[\alpha_{l^\star}(e_1, U_1)\right]_i \|^2\right] - \|w\|^2 n\right)$. Therefore it is enough to show that $\mathbb{E}\left[\left\|\sum_{i=1}^{\mu} \frac{w_i}{\|w\|} \left[\alpha_{l^\star}(e_1, U_1)\right]_i\right\|^2\right] - n = \mathbb{E}\left[\left(\sum_{i=1}^{\mu} \frac{w_i}{\|w\|} \mathcal{N}^{i:\lambda}\right)^2\right] - 1$. Recall that the probability density function of $\alpha_{l^\star}(e_1, U_1)$ is $p_{e_1}^{l^\star}$ defined for all $u \in \mathbb{R}^{n\mu}$ as $p_{e_1}^{l^\star}(u) = \frac{\lambda!}{(\lambda-\mu)!}(1 - Q_{e_1}^{l^\star}(u^\mu))^{\lambda-\mu} \prod_{i=1}^{\mu-1} \mathbb{1}_{\{l^\star(u^i) < l^\star(u^{i+1})\}} \prod_{i=1}^{\mu} p_{\mathcal{N}_n}(u^i)$.

Denote $A = \mathbb{E}_{U_1 \sim \mathcal{N}_{n\lambda}}\left[\|\sum_{i=1}^{\mu} \frac{w_i}{\|w\|} \left[\alpha_{l^\star}(e_1, U_1)\right]_i\|^2\right]$. It follows that

$$A = \frac{\lambda!}{(\lambda-\mu)!} \int \left\|\sum_{i=1}^{\mu} \frac{w_i}{\|w\|} u^i\right\|^2 (1 - Q_{e_1}^{l^\star}(u^\mu))^{\lambda-\mu} \prod_{j=1}^{\mu-1} \mathbb{1}_{\{l^\star(u^j) < l^\star(u^{j+1})\}} \prod_{j=1}^{\mu} p_{\mathcal{N}_n}(u^j) \mathrm{d}u$$

$$= \int \left(\left\|\sum_{i=1}^{\mu} \frac{w_i}{\|w\|} u_1^i\right\|^2 + \left\|\sum_{i=1}^{\mu} \frac{w_i}{\|w\|} u_{-1}^i\right\|^2\right) P\left(\mathcal{N} > u_1^\mu\right)^{\lambda-\mu} \prod_{j=1}^{\mu-1} \mathbb{1}_{\{u_1^j < u_1^{j+1}\}} \prod_{j=1}^{\mu} p_{\mathcal{N}}(u_1^j)$$

$\prod_{j=1}^{\mu} p_{\mathcal{N}_{n-1}}(u_{-1}^j)\, \mathrm{d}u$. We expand the integrand, the first term is $\mathbb{E}\left[\left(\sum_{i=1}^{\mu} \frac{w_i}{\|w\|} \mathcal{N}^{i:\lambda}\right)^2\right]$.

Denote $B = \mathbb{E}\left[\left(\sum_{i=1}^{\mu} \frac{w_i}{\|w\|} \mathcal{N}^{i:\lambda}\right)^2\right]$ and $C = B - A$. Then $\frac{(\lambda-\mu)!}{\lambda!} C$ equals

$$\int \left\|\sum_{i=1}^{\mu} \frac{w_i}{\|w\|} u_{-1}^i\right\|^2 P\left(\mathcal{N} > u_1^\mu\right)^{\lambda-\mu} \prod_{j=1}^{\mu-1} \mathbb{1}_{\{u_1^j < u_1^{j+1}\}} \prod_{j=1}^{\mu} p_{\mathcal{N}}(u_1^j) p_{\mathcal{N}_{n-1}}(u_{-1}^j) \mathrm{d}u.$$

Then $C = \int_{\mathbb{R}^\mu} \frac{\lambda!}{(\lambda-\mu)!} P\left(\mathcal{N} > u_1^\mu\right)^{\lambda-\mu} \prod_{j=1}^{\mu-1} \mathbb{1}_{\{u_1^j < u_1^{j+1}\}} \prod_{j=1}^{\mu} p_{\mathcal{N}}(u_1^j) \mathrm{d}u_1$
$\int_{\mathbb{R}^{(n-1)\mu}} \left\|\sum_{i=1}^{\mu} \frac{w_i}{\|w\|} u_{-1}^i\right\|^2 \prod_{j=1}^{\mu} p_{\mathcal{N}_{n-1}}(u_{-1}^j) \mathrm{d}u_{-1}$. The first integral equals 1 as it is the integral of a probability density function. The second integral is equal to $\mathbb{E}\left[\|\sum_{i=1}^{\mu} \frac{w_i}{\|w\|} W_i\|^2\right]$ where $W_1, \ldots, W_\mu$ are i.i.d. random variables of law $\mathcal{N}_{n-1}$. Then the law of $\sum_{i=1}^{\mu} \frac{w_i}{\|w\|} W_i$ is $\mathcal{N}_{n-1}$. Then $\mathbb{E}\left[\left\|\sum_{i=1}^{\mu} \frac{w_i}{\|w\|} W_i\right\|^2\right] = n - 1$. Hence
$\mathbb{E}_{U_1 \sim \mathcal{N}_{n\lambda}}\left[\left\|\sum_{i=1}^{\mu} \frac{w_i}{\|w\|} \left[\alpha_{l^\star}(e_1, U_1)\right]_i\right\|^2\right] - \mathbb{E}\left[\left(\sum_{i=1}^{\mu} \frac{w_i}{\|w\|} \mathcal{N}^{i:\lambda}\right)^2\right] = n - 1$, which ends this part.

Part 2. For the second item, we show that $\mathbb{E}_{U_1\sim\mathcal{N}_{n\lambda}}\left[\log\left(\Gamma_{\text{xNES}}\left(\alpha_{l^\star}(e_1,U_1)\right)\right)\right]$ has the same sign than $\sum_{i=1}^{\mu}\frac{w_i}{\sum_{j=1}^{\mu}w_j}\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]-1$, and apply Theorem 7. We have $\mathbb{E}_{U_1\sim\mathcal{N}_{n\lambda}}\left[\log\left(\Gamma_{\text{xNES}}\left(\alpha_{l^\star}(e_1,U_1)\right)\right)\right]=$ $\frac{1}{2d_\sigma n\sum_{i=1}^{\mu}w_i}\sum_{i=1}^{\mu}w_i\left(\mathbb{E}_{U_1\sim\mathcal{N}_{n\lambda}}\left[\|\left[\alpha_{l^\star}(e_1,U_1)\right]_i\|^2\right]-n\right)$. Then it is enough to show: $\sum_{i=1}^{\mu}w_i\left(\mathbb{E}_{U_1\sim\mathcal{N}_{n\lambda}}\left[\|\left[\alpha_{l^\star}(e_1,U_1)\right]_i\|^2\right]-n\right)=\sum_{i=1}^{\mu}w_i\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]-\sum_{i=1}^{\mu}w_i$. Denote $A=\sum_{i=1}^{\mu}w_i\mathbb{E}_{U_1\sim\mathcal{N}_{n\lambda}}\left[\|\left[\alpha_{l^\star}(e_1,U_1)\right]_i\|^2\right]$. It follows $A=\frac{\lambda!}{(\lambda-\mu)!}\int\sum_{i=1}^{\mu}w_i\|u^i\|^2(1-Q_{e_1}^{l^\star}(u^\mu))^{\lambda-\mu}\prod_{j=1}^{\mu-1}\mathbb{1}_{\{l^\star(u^j)<l^\star(u^{j+1})\}}\prod_{j=1}^{\mu}p_{\mathcal{N}_n}(u^j)\mathrm{d}u$ which is equal to $\frac{\lambda!}{(\lambda-\mu)!}\int\left(\sum_{i=1}^{\mu}w_i\|u_1^i\|^2+\sum_{i=1}^{\mu}w_i\|u_{-1}^i\|^2\right)P\Big(\mathcal{N}>u_1^\mu\Big)^{\lambda-\mu}\prod_{j=1}^{\mu-1}\mathbb{1}_{\{u_1^j<u_1^{j+1}\}}$ $\prod_{j=1}^{\mu}p_{\mathcal{N}}(u_1^j)\prod_{j=1}^{\mu}p_{\mathcal{N}_{n-1}}(u_{-1}^j)\,\mathrm{d}u$. Then after expansion, the integral of the first term of the integrand equals $\frac{(\lambda-\mu)!}{\lambda!}\sum_{i=1}^{\mu}w_i\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]$. Denote $B=\sum_{i=1}^{\mu}w_i\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]$ and $C=A-B$. Then $\frac{(\lambda-\mu)!}{\lambda!}C=\int\sum_{i=1}^{\mu}w_i\|u_{-1}^i\|^2\left(P\left(\mathcal{N}>u_1^\mu\right)\right)^{\lambda-\mu}$ $\prod_{j=1}^{\mu-1}\mathbb{1}_{\{u_1^j<u_1^{j+1}\}}\prod_{j=1}^{\mu}p_{\mathcal{N}}(u_1^j)p_{\mathcal{N}_{n-1}}(u_{-1}^j)\mathrm{d}u$. Then $C=\int_{\mathbb{R}^\mu}\frac{\lambda!}{(\lambda-\mu)!}\prod_{j=1}^{\mu-1}\mathbb{1}_{\{u_1^j<u_1^{j+1}\}}$ $P\left(\mathcal{N}>u_1^\mu\right)^{\lambda-\mu}\prod_{j=1}^{\mu}p_{\mathcal{N}}(u_1^j)\mathrm{d}u_1\int_{\mathbb{R}^{(n-1)\mu}}\sum_{i=1}^{\mu}w_i\|u_{-1}^i\|^2\prod_{j=1}^{\mu}p_{\mathcal{N}_{n-1}}(u_{-1}^j)\mathrm{d}u_{-1}$. The first integral equals 1 as it is the integral of a probability density function. The second one equals $\sum_{i=1}^{\mu}w_i\mathbb{E}\left[\|\mathcal{N}_{n-1}\|^2\right]=(n-1)\sum_{i=1}^{\mu}w_i$. We finally have that $\sum_{i=1}^{\mu}w_i\mathbb{E}_{U_1\sim\mathcal{N}_{n\lambda}}\left[\|\left[\alpha_{l^\star}(e_1,U_1)\right]_i\|^2\right]-\sum_{i=1}^{\mu}w_i\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]=(n-1)\sum_{i=1}^{\mu}w_i$.

Part 3. If $(X_1,\ldots,X_\lambda)$ is distributed according to $\left(\mathcal{N}^{1:\lambda},\ldots,\mathcal{N}^{\lambda:\lambda}\right)$, then $X_1\leq\cdots\leq X_\lambda$ and then $-X_\lambda\leq\cdots\leq-X_1$. Therefore $(-X_\lambda,\ldots,-X_1)$ is also distributed according to $\left(\mathcal{N}^{1:\lambda},\ldots,\mathcal{N}^{\lambda:\lambda}\right)$. Assume that $\lambda\geq3$ and $\mu>\frac{\lambda}{2}$. We show the results in two parts.

Part 3.1. First we assume that $w_1=\cdots=w_\mu=\frac{1}{\mu}$. In this case, we have to prove that: $1<\sum_{i=1}^{\mu}\frac{w_i}{\sum_{j=1}^{\mu}w_j}\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]=\frac{1}{\mu}\sum_{i=1}^{\mu}\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]$. Since $\mathcal{N}^{1:\lambda}\leq\cdots\leq\mathcal{N}^{\lambda:\lambda}$ is equivalent to $-\mathcal{N}^{\lambda:\lambda}\leq\cdots\leq-\mathcal{N}^{1:\lambda}$, then $\left(\mathcal{N}^{1:\lambda},\ldots,\mathcal{N}^{\lambda:\lambda}\right)$ has the distribution of $\left(-\mathcal{N}^{\lambda:\lambda},\ldots,-\mathcal{N}^{1:\lambda}\right)$. And then for $i=1,\ldots,\lambda$, $\left(\mathcal{N}^{i:\lambda}\right)^2$ has the distribution of $\left(\mathcal{N}^{\lambda-i+1:\lambda}\right)^2$. It follows that $\sum_{i=1}^{\lambda}\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]=2\sum_{i=1}^{\mu}\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]+\sum_{i=\mu+1}^{\lambda-\mu}\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]$. Moreover, $\sum_{i=1}^{\lambda}\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]=\sum_{i=1}^{\lambda}\mathbb{E}\left[\left(\mathcal{N}^{i}\right)^2\right]=\lambda$, meaning that we lose the selection effect of the order statistics when we do the above

summation. Both equations above ensure that

$$2\sum_{i=1}^{\mu}\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]+\sum_{i=\mu+1}^{\lambda-\mu}\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]=\lambda. \tag{36}$$

For any $j\in\{\mu+1,\dots,\lambda-\mu\}$ and any $i\in\{1\dots,\mu\}$, $\mathcal{N}^{i:\lambda}\leq\mathcal{N}^{j:\lambda}\leq\mathcal{N}^{\lambda+1-i:\lambda}$. Therefore if $\mathcal{N}^{j:\lambda}\geq 0$, $\left(\mathcal{N}^{j:\lambda}\right)^2\leq\left(\mathcal{N}^{\lambda+1-i:\lambda}\right)^2$, and if $\mathcal{N}^{j:\lambda}\leq 0$, $\left(\mathcal{N}^{j:\lambda}\right)^2\leq\left(\mathcal{N}^{i:\lambda}\right)^2$. Since $\left(\mathcal{N}^{\lambda+1-i:\lambda}\right)^2$ has the distribution of $\left(\mathcal{N}^{i:\lambda}\right)^2$, it follows that for all $j\in\{\mu+1,\dots,\lambda-\mu\}$ and $i\in\{1\dots,\mu\}$: $\left(\mathcal{N}^{j:\lambda}\right)^2\leq\left(\mathcal{N}^{i:\lambda}\right)^2$, and it is straightforward to see that the we do not have almost sure equality. It then follows that for all $j\in\{\mu+1,\dots,\lambda-\mu\}$ [4] and $i\in\{1\dots,\mu\}:\mathbb{E}\left[\left(\mathcal{N}^{j:\lambda}\right)^2\right]<\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]$. Therefore for all $j\in\{\mu+1,\dots,\lambda-\mu\}$

$$\mathbb{E}\left[\left(\mathcal{N}^{j:\lambda}\right)^2\right]<\frac{1}{\mu}\sum_{i=1}^{\mu}\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]. \tag{37}$$

With (37) and (36), we have $\lambda=2\sum_{i=1}^{\mu}\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]+\sum_{i=\mu+1}^{\lambda-\mu}\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]<2\sum_{i=1}^{\mu}\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]+\frac{\lambda-2\mu}{\mu}\sum_{i=1}^{\mu}\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]=\frac{\lambda}{\mu}\sum_{i=1}^{\mu}\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]$. Finally it follows that

$$\frac{1}{\mu}\sum_{i=1}^{\mu}\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]>1. \tag{38}$$

Part 3.2. Now we fall back to the general assumption where $w_1\geq\cdots\geq w_\mu$. Let us prove beforehand that:

$$\mathbb{E}\left[\left(\mathcal{N}^{1:\lambda}\right)^2\right]\geq\mathbb{E}\left[\left(\mathcal{N}^{2:\lambda}\right)^2\right]\geq\cdots\geq\mathbb{E}\left[\left(\mathcal{N}^{\mu:\lambda}\right)^2\right]. \tag{39}$$

Let $i\in\{1,\dots,\mu-1\}$. We have that $\mathcal{N}^{i:\lambda}\leq\mathcal{N}^{i+1:\lambda}\leq\mathcal{N}^{\lambda+1-i:\lambda}$. Then if $\mathcal{N}^{i+1:\lambda}\geq 0$, $\left(\mathcal{N}^{i+1:\lambda}\right)^2\leq\left(\mathcal{N}^{\lambda+1-i:\lambda}\right)^2$ and if $\mathcal{N}^{i+1:\lambda}\leq 0$, $\left(\mathcal{N}^{i+1:\lambda}\right)^2\leq\left(\mathcal{N}^{i:\lambda}\right)^2$. Since $\left(\mathcal{N}^{\lambda+1-i:\lambda}\right)^2$ and $\left(\mathcal{N}^{i:\lambda}\right)^2$ have the same distribution, it follows that $\left(\mathcal{N}^{i+1:\lambda}\right)^2\leq\left(\mathcal{N}^{i:\lambda}\right)^2$. Therefore (39) holds.

To prove the general case, we use the Chebyshev's sum inequality which states that if $a_1\geq a_2\geq\cdots\geq a_\mu$ and $b_1\geq b_2\geq\cdots\geq b_\mu$, then $\frac{1}{\mu}\sum_{k=1}^{\mu}a_kb_k\geq\left(\frac{1}{\mu}\sum_{k=1}^{\mu}a_k\right)\left(\frac{1}{\mu}\sum_{k=1}^{\mu}b_k\right)$. By applying Chebyshev's sum inequality on $w_1\geq\cdots\geq w_\mu$ and $\mathbb{E}\left[\left(\mathcal{N}^{1:\lambda}\right)^2\right]\geq\mathbb{E}\left[\left(\mathcal{N}^{2:\lambda}\right)^2\right]\geq\cdots\geq\mathbb{E}\left[\left(\mathcal{N}^{\mu:\lambda}\right)^2\right]$, it follows that $\frac{1}{\mu}\sum_{i=1}^{\mu}w_i\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]\geq\left(\frac{1}{\mu}\sum_{j=1}^{\mu}w_j\right)\left(\frac{1}{\mu}\sum_{i=1}^{\mu}\mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]\right)$. Therefore,

[4] Note that the set $\{\mu+1,\dots,\lambda-\mu\}$ is not empty since $1\leq\mu<\frac{\lambda}{2}$.

$\sum_{i=1}^{\mu} \frac{w_i}{\sum_{j=1}^{\mu} w_j} \mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right] \geq \frac{1}{\mu} \sum_{i=1}^{\mu} \mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right]$. And the first case in (38) ensures that $\sum_{i=1}^{\mu} \frac{w_i}{\sum_{j=1}^{\mu} w_j} \mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right] \geq \frac{1}{\mu} \sum_{i=1}^{\mu} \mathbb{E}\left[\left(\mathcal{N}^{i:\lambda}\right)^2\right] > 1$.

□

The positivity of $\mathbb{E}\left[\log\left(\Gamma^{\star}_{\text{linear}}\right)\right]$ is the main assumption for our main results. In this context, Proposition 12 gives more practical and concrete ways to obtain the conclusion of Theorems 7 and 8 for the $(\mu/\mu_w, \lambda)$-CSA1-ES and $(\mu/\mu_w, \lambda)$-xNES. In the case where $\mu = 1$, the two conditions given in the previous proposition for CSA and xNES are equivalent and yield the equation $\mathbb{E}\left[\left(\mathcal{N}^{1:\lambda}\right)^2\right] > 1$. The latter is satisfied if $\lambda \geq 3$ and $\mu = 1$, which is the linear divergence condition on linear functions of the $(1, \lambda)$-CSA1-ES [20]. Conditions similar to the one given for CSA in the previous proposition had already been derived for the so-called mutative self-adaptation of the step-size [26].

7 Conclusion and discussion

We have proven the asymptotic linear behavior of step-size adaptive $(\mu/\mu_w, \lambda)$-ESs on composites of strictly increasing functions with continuously differentiable scaling-invariant functions. The step-size update has been modeled as an abstract function of the random input multiplied by the current step-size. Two well-known step-size adaptation mechanisms are included in this model, namely derived from the Exponential Natural Evolution Strategy (xNES) [25] and the Cumulative Step-size Adaptation (CSA) [29] without cumulation.

Our main condition for the linear behavior proven in Theorem 7 reads "the logarithm of the step-size increases on linear functions", formally, stated as $\mathbb{E}\left[\log\left(\Gamma^{\star}_{\text{linear}}\right)\right] > 0$ where $\Gamma^{\star}_{\text{linear}}$ is the step-size change on nontrivial linear functions. This condition is equivalent to the geometric divergence of the step-size on nontrivial linear functions, as shown by the next lemma.

Lemma 10 *Let f be an increasing transformation of a nontrivial linear function, i.e. satisfy F2. Let $\{(X_k, \sigma_k)\,; k \in \mathbb{N}\}$ be the sequence defined in (5) and (6). Assume that $\{U_{k+1}\,; k \in \mathbb{N}\}$ satisfies Assumption A5 and that Γ satisfies Assumptions A2 and A4, i.e. Γ is invariant under rotation and $\log \circ \Gamma$ is $\mathcal{N}_{n\mu}$-integrable. Then* $\lim_{k\to\infty} \frac{1}{k} \log \frac{\sigma_k}{\sigma_0} = \mathbb{E}\left[\log\left(\Gamma^{\star}_{\text{linear}}\right)\right]$.

Proof. We have $\frac{1}{k} \log \frac{\sigma_k}{\sigma_0} = \frac{1}{k} \sum_{t=0}^{k-1} \log \frac{\sigma_{t+1}}{\sigma_t}$. With (15) and Proposition 1, $\sigma_{t+1} = \sigma_t \, \Gamma\left(\alpha_{l^\star}(0, U_{t+1})\right)$ where $l^\star$ is the linear function defined as $l^\star(x) = x_1$ for $x \in \mathbb{R}^n$. Therefore $\frac{1}{k} \log \frac{\sigma_k}{\sigma_0} = \frac{1}{k} \sum_{t=0}^{k-1} \left(\log \circ \Gamma \circ \alpha_{l^\star}\right)(0, U_{t+1})$. Using Assumption A3 and Lemma 8, we have that the function $u \mapsto \left(\log \circ \Gamma \circ \alpha_{l^\star}\right)(0, u)$ is $\mathcal{N}_{n\lambda}$-integrable. Then by the LLN applied to the i.i.d. sequence $\{U_{k+1}\,; k \in \mathbb{N}\}$, $\frac{1}{k} \log \frac{\sigma_k}{\sigma_0}$ converges almost surely to $\mathbb{E}\left[\log\left(\Gamma^{\star}_{\text{linear}}\right)\right]$. □

Geometric divergence of the step-size on a linear function is also the main condition when analyzing the deterministic flow of the IGO algorithm [3]. For the (1+1)-ES and the $(1, \lambda)$ self-adaptive ES, a different condition than $\mathbb{E}\left[\log\left(\Gamma^{\star}_{\text{linear}}\right)\right] > 0$ has

been used to characterize the step-size increase on linear functions: there exists $\beta > 0$ such that $\mathbb{E}\left[\frac{1}{{\Gamma^\star_{\text{linear}}}^\beta}\right] < 1$ [11,13]. With the concavity of the logarithm and Jensen's inequality, we have that $\log\left(\mathbb{E}\left[\frac{1}{{\Gamma^\star_{\text{linear}}}^\beta}\right]\right) \geq \mathbb{E}\left[\log\left(\frac{1}{{\Gamma^\star_{\text{linear}}}^\beta}\right)\right] = -\beta\,\mathbb{E}\left[\log\left(\Gamma^\star_{\text{linear}}\right)\right]$. Therefore $\mathbb{E}\left[\frac{1}{{\Gamma^\star_{\text{linear}}}^\beta}\right] < 1$ implies $\mathbb{E}\left[\log\left(\Gamma^\star_{\text{linear}}\right)\right] > 0$ and our condition that "the logarithm of the step-size increases on linear functions" is tighter than the previously used.

Our main condition for the linear behavior of the $(\mu/\mu_w, \lambda)$-CSA-ES algorithm without cumulation is formulated based on λ, μ, the weights w and the order statistics of the standard normal distribution as $\mathbb{E}\left[\left(\sum_{i=1}^{\mu} \frac{w_i}{\|w\|}\mathcal{N}^{i:\lambda}\right)^2\right] > 1$, see Proposition 12. For $\mu = 1$, this condition is satisfied when $\lambda \geq 3$.

The linear divergence of both the incumbent and the step-size was proven for a $(1, \lambda)$-ES without cumulation on linear functions whenever $\lambda \geq 3$ with a divergence rate equal to $\frac{\mathbb{E}\left[(\mathcal{N}^{1:\lambda})^2\right]-1}{2d_\sigma n}$ [20]. Proposition 12 extends this result to values of $\mu > 1$. Note that our results cover both, linear divergence on strictly increasing transformations of nontrivial linear functions and linear behavior on strictly increasing transformations of C^1 scaling-invariant functions with a unique global argmin.

Our methodology leans on investigating the stability of the σ-normalized homogeneous Markov chain to be able to apply an LLN and obtain the limit of the log-distance to the optimum divided by the iteration index. Then we obtain an exact expression of the rate of convergence or divergence as an expectation with respect to the stationary distribution of the σ-normalized chain. This is an elegant feature of our analysis. Other approaches [1,2,39–42] provide bounds on the convergence rate but not its exact expression with the advantage that the bounds are often expressed depending on dimension or population size and thus describe the scaling of the algorithm with respect to relevant parameters.

The class of scaling-invariant functions is, as far as we can see, the largest class to which our methodology can conceivably be applied—because on any wider class of functions, a selection function for the σ-normalized Markov chain can not anymore reflect the selection operation in the underlying chain. We require additionally that the objective function is a strictly increasing transformation of either a continuously differentiable function with a unique global argmin or a nontrivial linear function. Many non-convex functions with non-convex sublevel sets are included.

The implied requirement of smooth level sets is instrumental for our analysis. We believe that there exist unimodal functions with non-smooth level sets on which scale-invariant ESs can not converge to the global optimum with probability one independently of the initial conditions, for example $x \mapsto \sum_{i=1}^{n} \sqrt{|x_i|}$. However, smooth level sets are not a necessary condition for convergence—we consistently observe convergence on $x \mapsto \sum_{i=1}^{n} |x_i|$ for smaller values of n and understand the reason why ESs succeed on the one-norm but fail on the 1/2-norm function. Capturing this distinction in a rigorous analysis of the Markov chain remains an open challenge.

A broader function class has been analyzed by requiring a drift condition to hold on the whole state-space [2] while our methodology requires the drift condition to only hold outside of a small set (when the step-size is much smaller

than the distance to the optimum). Hence in our approach, it suffices to control the behavior in the limit when the step-size normalized by the distance to the optimum approaches zero.

A major limitation of our current analysis is the omission of cumulation that is used in the $(\mu/\mu_w, \lambda)$-CSA-ES to adapt the step-size (we have set the cumulation parameter to 1, see Section 2.2). In case of a parent population of size $\mu = 1$, Chotard et al. [20] obtain linear divergence of the step-size on linear functions also with cumulation. However, no proof of linear behavior exists, to our knowledge, on functions whose level sets are not affine subspaces. While we consider cumulation a crucial component in practice, proving the drift condition for the stability of the Markov chain is much harder when the state space is extended with the cumulative evolution path and this remains an open challenge.

Technically, our results rely on proving φ-irreducibility, positivity and Harris-recurrence of the σ-normalized Markov chain. The φ-irreducibility is difficult to prove directly for the class of algorithms studied in this paper while it is relatively easy to prove for the $(1, \lambda)$-ES with self-adaptation [11] or for the (1+1)-ES with one-fifth success rule [13]. We circumvented the problem by looking at the stability of an underlying deterministic control model and exploit its connection to the stability of Markov chains [19]. Positivity and Harris-recurrence are proven using Foster-Lyapunov drift conditions [50]. We prove a drift condition for geometric ergodicity that implies positivity and Harris-recurrence. The main ingredient for obtaining the drift condition is the convergence in distribution of the step-size change towards the step-size change on a linear function when $Z_k = z$ goes to infinity. It implies that the drift condition holds for $Z_k = z$ outside a compact set. We also prove in Lemma 6 the existence of non-negligible sets with respect to the invariant probability measure π, outside of a neighborhood of a steadily attracting state. This is used in Proposition 11 to obtain the π-integrability of the function $z \mapsto \log \|z\|$.

We have developed generic results to facilitate further studies of similar Markov chains. More specifically, applying an LLN to the σ-normalized chain is not enough to conclude linear convergence. We introduce the technique to apply the generalized LLN to an abstract chain $\{(Z_k, U_{k+2}) ; k \in \mathbb{N}\}$ and prove that stability properties from $\{Z_k ; k \geq 0\}$ are transferred to $\{(Z_k, U_{k+2}) ; k \in \mathbb{N}\}$.

Acknowledgements

Part of this research has been conducted in the context of a research collaboration between Storengy and Inria. We particularly thank F. Huguet and A. Lange from Storengy for their strong support.

References

1. Akimoto, Y., Auger, A., Glasmachers, T.: Drift theory in continuous search spaces: expected hitting time of the (1+1)-ES with 1/5 success rule. In: Proceedings of the Genetic and Evolutionary Computation Conference, pp. 801–808 (2018)
2. Akimoto, Y., Auger, A., Glasmachers, T., Morinaga, D.: Global linear convergence of evolution strategies on more than smooth strongly convex functions. SIAM Journal on Optimization (2022). Accepted

3. Akimoto, Y., Auger, A., Hansen, N.: Convergence of the continuous time trajectories of isotropic evolution strategies on monotonic $\mathcal{C}^2$-composite functions. In: International Conference on Parallel Problem Solving from Nature, pp. 42–51. Springer (2012)
4. Akimoto, Y., Auger, A., Hansen, N.: An ODE method to prove the geometric convergence of adaptive stochastic algorithms. Stochastic Processes and their Applications **145**, 269–307 (2022)
5. Akimoto, Y., Nagata, Y., Ono, I., Kobayashi, S.: Bidirectional relation between CMA evolution strategies and natural evolution strategies. In: International Conference on Parallel Problem Solving from Nature, pp. 154–163. Springer (2010)
6. Akimoto, Y., Nagata, Y., Ono, I., Kobayashi, S.: Theoretical analysis of evolutionary computation on continuously differentiable functions. In: Proceedings of the 12th annual conference on Genetic and evolutionary computation, pp. 1401–1408 (2010)
7. Akimoto, Y., Nagata, Y., Ono, I., Kobayashi, S.: Theoretical foundation for CMA-ES from information geometry perspective. Algorithmica **64**(4), 698–716 (2012)
8. Arnold, D.V.: Optimal weighted recombination. In: International Workshop on Foundations of Genetic Algorithms, pp. 215–237. Springer (2005)
9. Arnold, D.V.: Weighted multirecombination evolution strategies. Theoretical computer science **361**(1), 18–37 (2006)
10. Arnold, D.V., Beyer, H.G.: Random dynamics optimum tracking with evolution strategies. In: International Conference on Parallel Problem Solving from Nature, pp. 3–12. Springer (2002)
11. Auger, A.: Convergence results for the (1, λ)-SA-ES using the theory of ϕ-irreducible Markov chains. Theoretical Computer Science **334**(1-3), 35–69 (2005)
12. Auger, A., Hansen, N.: A restart CMA evolution strategy with increasing population size. In: 2005 IEEE congress on evolutionary computation, vol. 2, pp. 1769–1776. IEEE (2005)
13. Auger, A., Hansen, N.: 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. arXiv preprint arXiv:1310.8397 (2013)
14. Auger, A., Hansen, N.: Linear convergence of comparison-based step-size adaptive randomized search via stability of Markov chains. SIAM Journal on Optimization **26**(3), 1589–1624 (2016)
15. Bienvenüe, A., François, O.: Global convergence for evolution strategies in spherical problems: some simple proofs and difficulties. Theoretical Computer Science **306**(1-3), 269–289 (2003)
16. Billingsley, P.: Convergence of probability measures. Wiley (1999)
17. Bosman, P.A., Grahl, J., Thierens, D.: Benchmarking parameter-free AMaLGaM on functions with and without noise. Evolutionary Computation **21**(3), 445–469 (2013)
18. Bouter, A., Alderliesten, T., Bosman, P.A.: Achieving Highly Scalable Evolutionary Real-Valued Optimization by Exploiting Partial Evaluations. Evolutionary Computation **29**(1), 129–155 (2021)
19. Chotard, A., Auger, A.: Verifiable conditions for the irreducibility and aperiodicity of Markov chains by analyzing underlying deterministic models. Bernoulli **25**(1), 112–147 (2019)
20. Chotard, A., Auger, A., Hansen, N.: Cumulative step-size adaptation on linear functions. In: International Conference on Parallel Problem Solving from Nature, pp. 72–81. Springer (2012)
21. Diouane, Y., Gratton, S., Vicente, L.N.: Globally convergent evolution strategies. Mathematical Programming **152**(1), 467–490 (2015)
22. Feoktistov, V.: Differential evolution. Springer (2006)
23. García-Martínez, C., Gutiérrez, P.D., Molina, D., Lozano, M., Herrera, F.: Since CEC 2005 competition on real-parameter optimisation: a decade of research, progress and comparative analysis's weakness. Soft Computing **21**(19), 5573–5583 (2017)
24. Glasmachers, T., Krause, O.: Convergence analysis of the Hessian estimation evolution strategy. Evolutionary computation **30**(1), 27–50 (2022)
25. Glasmachers, T., Schaul, T., Yi, S., Wierstra, D., Schmidhuber, J.: Exponential natural evolution strategies. In: Proceedings of the 12th annual conference on Genetic and evolutionary computation, pp. 393–400 (2010)
26. Hansen, N.: An analysis of mutative σ-self-adaptation on linear fitness functions. Evolutionary computation **14**(3), 255–275 (2006)
27. Hansen, N.: The CMA evolution strategy: a comparing review. Towards a new evolutionary computation pp. 75–102 (2006)

28. Hansen, N.: Benchmarking a BI-Population CMA-ES on the BBOB-2009 function testbed. In: Proceedings of the 11th annual conference companion on genetic and evolutionary computation conference: late breaking papers, pp. 2389–2396 (2009)
29. Hansen, N.: The CMA evolution strategy: A tutorial. arXiv preprint arXiv:1604.00772 (2016)
30. Hansen, N., Arnold, D.V., Auger, A.: Evolution Strategies. In: Springer handbook of computational intelligence, pp. 871–898. Springer, Berlin (2015)
31. Hansen, N., Auger, A.: Principled design of continuous stochastic search: From theory to practice. In: Theory and principled methods for the design of metaheuristics, pp. 145–180. Springer (2014)
32. Hansen, N., Auger, A., Ros, R., Finck, S., Pošík, P.: Comparing results of 31 algorithms from the black-box optimization benchmarking BBOB-2009. In: Proceedings of the 12th annual conference companion on Genetic and evolutionary computation, pp. 1689–1696 (2010)
33. Hansen, N., Gemperle, F., Auger, A., Koumoutsakos, P.: When do heavy-tail distributions help? In: Parallel Problem Solving from Nature-PPSN IX, pp. 62–71. Springer (2006)
34. Hansen, N., Kern, S.: Evaluating the CMA evolution strategy on multimodal test functions. In: International Conference on Parallel Problem Solving from Nature, pp. 282–291. Springer (2004)
35. Hansen, N., Müller, S.D., Koumoutsakos, P.: Reducing the time complexity of the derandomized evolution strategy with covariance matrix adaptation (CMA-ES). Evolutionary computation **11**(1), 1–18 (2003)
36. Hansen, N., Ostermeier, A.: Completely derandomized self-adaptation in evolution strategies. Evolutionary computation **9**(2), 159–195 (2001)
37. Hansen, N., Ros, R., Mauny, N., Schoenauer, M., Auger, A.: Impacts of invariance in search: When CMA-ES and PSO face ill-conditioned and non-separable problems. Applied Soft Computing **11**(8), 5755–5769 (2011)
38. Igel, C., Suttorp, T., Hansen, N.: A computational efficient covariance matrix update and a (1+ 1)-CMA for evolution strategies. In: Proceedings of the 8th annual conference on Genetic and evolutionary computation, pp. 453–460 (2006)
39. Jägersküpper, J.: Analysis of a simple evolutionary algorithm for minimization in euclidean spaces. In: International Colloquium on Automata, Languages, and Programming, pp. 1068–1079. Springer (2003)
40. Jägersküpper, J.: Rigorous runtime analysis of the (1+1)-ES: 1/5-rule and ellipsoidal fitness landscapes. In: International Workshop on Foundations of Genetic Algorithms, pp. 260–281. Springer (2005)
41. Jägersküpper, J.: How the (1+1)-ES using isotropic mutations minimizes positive definite quadratic forms. Theoretical Computer Science **361**(1), 38–56 (2006)
42. Jägersküpper, J.: Algorithmic analysis of a basic evolutionary algorithm for continuous optimization. Theoretical Computer Science **379**(3), 329–347 (2007)
43. Jamieson, K.G., Nowak, R., Recht, B.: Query complexity of derivative-free optimization. Advances in Neural Information Processing Systems **25** (2012)
44. Jastrebski, G.A., Arnold, D.V.: Improving evolution strategies through active covariance matrix adaptation. In: 2006 IEEE international conference on evolutionary computation, pp. 2814–2821. IEEE (2006)
45. Jensen, S.T., Rahbek, A.: On the law of large numbers for (geometrically) ergodic Markov chains. Econometric Theory pp. 761–766 (2007)
46. Kappler, C.: Are evolutionary algorithms improved by large mutations? In: International Conference on Parallel Problem Solving from Nature–PPSN IV, pp. 346–355. Springer (1996)
47. Kennedy, J., Eberhart, R.: Particle swarm optimization. In: Proceedings of ICNN'95-international conference on neural networks, vol. 4, pp. 1942–1948. IEEE (1995)
48. Kern, S., Müller, S.D., Hansen, N., Büche, D., Ocenasek, J., Koumoutsakos, P.: Learning probability distributions in continuous evolutionary algorithms–a comparative review. Natural Computing **3**(1), 77–112 (2004)
49. Meyn, S.P., Caines, P.: Asymptotic behavior of stochastic systems possessing Markovian realizations. SIAM journal on control and optimization **29**(3), 535–561 (1991)
50. Meyn, S.P., Tweedie, R.L.: Markov chains and stochastic stability. Springer Science & Business Media (2012)

51. Morinaga, D., Akimoto, Y.: Generalized drift analysis in continuous domain: linear convergence of (1+1)-ES on strongly convex functions with lipschitz continuous gradients. In: Proceedings of the 15th ACM/SIGEVO Conference on Foundations of Genetic Algorithms, pp. 13–24 (2019)
52. Nesterov, Y.: Introductory lectures on convex optimization: A basic course, vol. 87. Springer Science & Business Media (2003)
53. Ollivier, Y., Arnold, L., Auger, A., Hansen, N.: Information-geometric optimization algorithms: A unifying picture via invariance principles. The Journal of Machine Learning Research **18**(1), 564–628 (2017)
54. Rechenberg, I.: Optimierung technischer Systeme nach Prinzipien der biologischen Evolution Dr.-Ing. Dissertation. Tech. rep., Verlag Frommann-Holzboog, Stuttgart-Bad Cannstatt (1973)
55. Rechenberg, I.: Evolutionsstrategie'94. frommann-holzboog (1994)
56. Rios, L.M., Sahinidis, N.V.: Derivative-free optimization: a review of algorithms and comparison of software implementations. Journal of Global Optimization **56**(3), 1247–1293 (2013)
57. Ros, R., Hansen, N.: A simple modification in CMA-ES achieving linear time and space complexity. In: International Conference on Parallel Problem Solving from Nature, pp. 296–305. Springer (2008)
58. Schaul, T.: Natural evolution strategies converge on sphere functions. In: Proceedings of the 14th annual conference on Genetic and evolutionary computation, pp. 329–336 (2012)
59. Schaul, T., Glasmachers, T., Schmidhuber, J.: High dimensions and heavy tails for natural evolution strategies. In: Proceedings of the 13th annual conference on Genetic and evolutionary computation, pp. 845–852 (2011)
60. Schwefel, H.P.: Numerische Optimierung von Computer-Modellen mittels der Evolutionsstrategie: mit einer vergleichenden Einführung in die Hill-Climbing-und Zufallsstrategie, vol. 1. Springer (1977)
61. Schwefel, H.P.: Evolution and Optimum Seeking. Sixth-Generation Computer Technology Series. John Wiley & Sons, Inc., New York (1995)
62. Storn, R., Price, K.: Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. Journal of global optimization **11**(4), 341–359 (1997)
63. Stout, W.F., Stout, W.F.: Almost sure convergence, vol. 24. Academic press (1974)
64. Suttorp, T., Hansen, N., Igel, C.: Efficient covariance matrix update for variable metric evolution strategies. Machine Learning **75**(2), 167–197 (2009)
65. Teytaud, O., Gelly, S.: General lower bounds for evolutionary algorithms. In: Parallel Problem Solving from Nature-PPSN IX, pp. 21–31. Springer (2006)
66. Touré, C., Gissler, A., Auger, A., Hansen, N.: Scaling-invariant functions versus positively homogeneous functions. Journal of Optimization Theory and Applications **191**(1), 363–383 (2021)
67. Tweedie, R.: The existence of moments for stationary Markov chains. Journal of Applied Probability **20**(1), 191–196 (1983)
68. Yao, X., Liu, Y.: Fast evolution strategies. In: International Conference on Evolutionary Programming, pp. 149–161. Springer (1997)

A Proof of Proposition 1

With Lemma 1, we assume without loss of generality that f is a nontrivial linear function. Let us remark beforehand that the random variable $\alpha_f(z, U_1)$ does not depend on z thanks to Lemma 4. Let $\varphi : \Gamma(\mathbb{R}^{n\mu}) \to \mathbb{R}$ be a continuous and bounded function, it is then enough to prove that $\mathbb{E}_{U_1 \sim \mathcal{N}_{n\lambda}}\left[\varphi\left(\Gamma(\alpha_f(z, U_1))\right)\right] = \mathbb{E}_{U_1 \sim \mathcal{N}_{n\lambda}}\left[\varphi\left(\Gamma(\alpha_{l^\star}(0, U_1))\right)\right]$. Denote by e_1 the unit vector $(1, 0, \ldots, 0)$, then for all $x \in \mathbb{R}^n$, $l^\star(x) = e_1^\top x$. Denote by $\tilde{e}_1$ the σ-normalized gradient of f at some point. Then there exists $K > 0$ such that for all $x \in \mathbb{R}^n$, $f(x) = K\tilde{e}_1^\top x$. And by the Gram-Schmidt process, there exist $(e_2, \ldots, e_n)$ and $(\tilde{e}_2, \ldots, \tilde{e}_n)$ such that $(e_1, e_2, \ldots, e_n)$ and $(\tilde{e}_1, \tilde{e}_2, \ldots, \tilde{e}_n)$ are orthonormal bases. Denote by T the linear function defined as $T(e_i) = \tilde{e}_i$ for $i = 1, \ldots, n$. Then T is an orthogonal matrix. For all $x \in \mathbb{R}^n$, $\tilde{e}_1^\top T(x) = e_1^\top x$, and $\|T(x)\| = \|x\|$. Denote $A = \mathbb{E}_{U_1 \sim \mathcal{N}_{n\lambda}}\left[\varphi\left(\Gamma(\alpha_f(z, U_1))\right)\right]$. We do a change of variable $u \mapsto \left(T(u^1), \ldots, T(u^\mu)\right)$. Then $\frac{(\lambda-\mu)!}{\lambda!} A = \int \varphi\left(\Gamma(u)\right) \mathbb{1}_{\tilde{e}_1^\top(u^2 - u^1) > 0, \ldots, \tilde{e}_1^\top(u^\mu - u^{\mu-1}) > 0}$
$P\left(\tilde{e}_1^\top \mathcal{N}_n > \tilde{e}_1^\top u^\mu\right)^{\lambda-\mu} p_{\mathcal{N}_n}(u^1) \ldots p_{\mathcal{N}_n}(u^\mu) \mathrm{d}u^1 \ldots \mathrm{d}u^\mu = \int \varphi\left(\Gamma\left(T(u^1), \ldots, T(u^\mu)\right)\right)$

$\mathbb{1}_{e_1^\top(u^2-u^1)>0,\dots,e_1^\top(u^\mu-u^{\mu-1})>0} P\left(e_1^\top \mathcal{N}_n > e_1^\top u^\mu\right)^{\lambda-\mu} p_{\mathcal{N}_n}(T(u^1))\dots p_{\mathcal{N}_n}(T(u^\mu))\mathrm{d}u^1\dots \mathrm{d}u^\mu$, thanks to the fact that $e_1^\top \mathcal{N}_n \sim \tilde{e}_1^\top \mathcal{N}_n \sim \mathcal{N}(0,1)$. Since Γ and $p_{\mathcal{N}_n}$ are invariant under rotation, $\mathbb{E}_{U_1\sim\mathcal{N}_{n\lambda}}\left[\varphi\left(\Gamma(\alpha_f(z,U_1))\right)\right] = \mathbb{E}_{U_1\sim\mathcal{N}_{n\lambda}}\left[\varphi\left(\Gamma(\alpha_{l^\star}(0,U_1))\right)\right]$.

B Proof of Proposition 3

We have $Z_{k+1} = G(Z_k, U_{k+1})$ and U_{k+3} is independent from $\{W_t\,; t \le k\}$, then $\{W_k\,; k\in\mathbb{N}\}$ is a Markov chain on $\mathcal{B}(\mathcal{Z})\otimes\mathcal{B}(\mathbb{R}^m)$. Let $(A,B)\in\mathcal{B}(\mathcal{Z})\times\mathcal{B}(\mathbb{R}^m)$ and $(z,u)\in\mathcal{Z}\times\mathbb{R}^m$. Then by independence $P\left((Z_{t+1},U_{t+3})\in A\times B|(Z_t,U_{t+2})=(z,u)\right) = P\left(Z_{t+1}\in A|Z_t=z\right)P\left(U_{t+3}\in B\right)$. For $(A,B)\in\mathcal{B}(\mathcal{Z})\times\mathcal{B}(\mathbb{R}^m)$, for $(z,u)\in\mathcal{Z}\times\mathbb{R}^m$, $\sum_{k=1}^\infty P^k((z,u),A\times B) = \Psi(B)\sum_{k=1}^\infty P^k(z,A)$. Therefore $\sum_{k=1}^\infty P^k((z,u),\cdot)$ is a product measure.

Let φ be an irreducible measure of $\{Z_k\,; k\in\mathbb{N}\}$ and let $E\in\mathcal{B}(\mathcal{Z})\otimes\mathcal{B}(\mathbb{R}^m)$. By definition of a product measure, $(\varphi\times\Psi)(E) = \int \varphi(E^v)\Psi(\mathrm{d}v)$ and thus $\sum_{k=1}^\infty P^k((z,u),E) = \int \sum_{k=1}^\infty P^k(z,E^v)\Psi(\mathrm{d}v)$ where $E^v = \{z\in\mathcal{Z}\,;(z,v)\in E\}$. If $\sum_{k=1}^\infty P^k((z,u),E) = 0$, then $0 = \sum_{k=1}^\infty P^k(z,E^v)$ for almost all v and then $\varphi(E^v)=0$ for almost all v. Then $(\varphi\times\Psi)(E) = \int \varphi(E^v)\Psi(\mathrm{d}v) = 0$, hence the $(\varphi\times\Psi)$-irreducibility of $\{W_k\,; k\in\mathbb{N}\}$.

Let us show that $\pi\times\Psi$ is an invariant probability measure of $\{W_k\,; k\in\mathbb{N}\}$ when π is an invariant measure of $\{Z_k\,; k\in\mathbb{N}\}$. Assume that $(A,B)\in\mathcal{B}(\mathcal{Z})\times\mathcal{B}(\mathbb{R}^m)$. Then $\int P\left((Z_1,U_3)\in A\times B|(Z_0,U_2)=(z,u)\right)(\pi\times\Psi)(\mathrm{d}(z,u)) = \int P_z\left(Z_1\in A\right)\Psi(B)\pi(\mathrm{d}z)\Psi(\mathrm{d}u) = \Psi(B)\pi(A) = (\pi\times\Psi)(A\times B)$. Hence $\pi\times\Psi$ is an invariant probability of $\{W_k\,; k\in\mathbb{N}\}$. Assume that $\{W_k\,; k\in\mathbb{N}\}$ has a d-cycle $(D_i)_{i=1,\dots,d}\in(\mathcal{B}(\mathcal{Z})\otimes\mathcal{B}(\mathbb{R}^m))^d$. Define for $i=1,\dots,d$, $\widetilde{D}_i = \{z\in\mathcal{Z}|\exists\, u\in\mathbb{R}^m\,;(z,u)\in D_i\}$ and let us prove that $\left(\widetilde{D}_i\right)_{i=1,\dots,d}$ is a d-cycle of $\{Z_k\,; k\in\mathbb{N}\}$.

Let $z\in\widetilde{D}_i$ and $i = 0,\dots,d-1 \pmod d$. There exists $u\in\mathbb{R}^m$ such that $(z,u)\in D_i$. Then $1 = P((z,u),D_{i+1}) = P\left((Z_1,U_3)\in D_{i+1}|Z_0=z\right) \le P\left(Z_1\in\widetilde{D}_{i+1}|Z_0=z\right)$. Thereforer $P\left(Z_1\in\widetilde{D}_{i+1}|Z_0=z\right)=1$.

If Λ is an irreducible measure of $\{Z_k\,; k\in\mathbb{N}\}$, then we have proven above that $\Lambda\times\Psi$ is an irreducible measure of $\{W_k\,; k\in\mathbb{N}\}$. Then $0 = (\Lambda\times\Psi)\left(\left(\bigcup_{i=1}^d D_i\right)^c\right)$. For $i=1,\dots,d$, $(\Lambda\times\Psi)(D_i) = \int \Lambda(D_i^v)\Psi(\mathrm{d}v) \le \int \Lambda(\widetilde{D}_i)\Psi(\mathrm{d}v) = \Lambda(\widetilde{D}_i)$. Then $\Lambda\left(\bigcup_{i=1}^d \widetilde{D}_i\right) = \sum_{i=1}^d \Lambda(\widetilde{D}_i) \ge (\Lambda\times\Psi)\left(\bigcup_{i=1}^d D_i\right)$. Hence $\Lambda\left(\left(\bigcup_{i=1}^d \widetilde{D}_i\right)^c\right) = 0$ and finally we have a d-cycle for $\{Z_k\,; k\in\mathbb{N}\}$. Similarly we can show that if $\{Z_k\,; k\in\mathbb{N}\}$ has a d-cycle, then $\{W_k\,; k\in\mathbb{N}\}$ also has a d-cycle. Now assume that C is a small set of $\{Z_k\,; k\in\mathbb{N}\}$. Then there exists a positive integer k and a nontrivial measure ν_k on $\mathcal{B}(\mathcal{Z})$ such that $P^k(z,A)\ge\nu_k(A)$ for all $z\in C$, $A\in\mathcal{B}(\mathcal{Z})$. If $(z,u)\in C\times\mathbb{R}^m$ and $E\in\mathcal{B}(\mathcal{Z})\otimes\mathcal{B}(\mathbb{R}^m)$, $P^k((z,u),E)\ge(\nu_k\times\Psi)(E)$ and therefore $C\times\mathbb{R}^m$ is a small set of $\{W_k\,; k\in\mathbb{N}\}$. The drift condition for $\{W_k\,; k\in\mathbb{N}\}$ follows directly from the drift condition for $\{Z_k\,; k\in\mathbb{N}\}$.

C Proof of Proposition 8

To prove the convergence in distribution of the step-size multiplicative factor for a function f that satisfies F1 or F2, we use the intermediate result given by Proposition 8, that asymptotically links $\Gamma\left(\alpha_f(x^\star+z,U_1)\right)$ to the random variable $\Gamma\left(\alpha_{l_z^f}(z,U_1)\right)$ where the nontrivial linear function l_z^f depends on z, ∇f, and is introduced in (26). Since $\alpha_f(x^\star+z,U_1) = \alpha_{\tilde{f}}(z,U_1)$, we assume without loss of generality that $x^\star=0$ and $f(0)=0$.

The next lemma is our fist step towards understanding the asymptotic behavior of $\alpha_f(z,U_1)$ for a C^1 scaling-invariant function f with a unique global argmin. For $\varphi:\mathbb{R}^{n\mu}\to\mathbb{R}$ continuous

and bounded, we approximate $\mathbb{E}\left[\varphi(\alpha_f(z,U_1))\right]$ by using the explicit definition of p_z^f in (14), and observing the integrals in the balls $\mathbf{B}\left(0,\sqrt{\|z\|}\right)$, such that the f-values we consider are relatively close to the f-values of $\frac{t_z^f}{\|z\|}z \in \mathcal{L}_{f,z_0}$.

Lemma 11 *Let f be a C^1 scaling-invariant function with a unique global argmin assumed to be in 0 such that $f(0)=0$. For $(z,w,v)\in(\mathbb{R}^n)^3$, define the function $h:(z,w,v)\mapsto \mathbb{1}_{\left\{f\left(\frac{t_z^f}{\|z\|}z+\frac{t_z^f}{\|z\|}w\right)>f\left(\frac{t_z^f}{\|z\|}z+\frac{t_z^f}{\|z\|}v\right)\right\}}$. Then for all $\varphi:\mathbb{R}^{n\mu}\to\mathbb{R}$ continuous and bounded:*

$\lim_{\|z\|\to\infty}\int_{\|u\|\leq\sqrt{\|z\|}}\left(\int_{\|w\|\leq\sqrt{\|z\|}}h(z,w,u^\mu)p_{\mathcal{N}_n}(w)\mathrm{d}w\right)^{\lambda-\mu}$
$\frac{\lambda!}{(\lambda-\mu)!}\varphi(u)\prod_{i=1}^{\mu-1}h(z,u^{i+1},u^i)\prod_{i=1}^{\mu}p_{\mathcal{N}_n}(u^i)\mathrm{d}u-\int\varphi(u)p_z^f(u)\mathrm{d}u=0.$

Proof. For $z\in\mathbb{R}^n$ and $u\in\mathbb{R}^{n\mu}$, define $A(z)=\left|\int\varphi(u)p_z^f(u)\mathrm{d}u-\int_{\|u\|\leq\sqrt{\|z\|}}\varphi(u)p_z^f(u)\mathrm{d}u\right|$. Then $A(z)\leq\frac{\lambda!}{(\lambda-\mu)!}\|\varphi\|_\infty\int_{\|u\|>\sqrt{\|z\|}}\prod_{i=1}^{\mu}p_{\mathcal{N}_n}(u^i)\mathrm{d}u=\frac{\lambda!}{(\lambda-\mu)!}\|\varphi\|_\infty\int_{\|u\|>\sqrt{\|z\|}}p_{\mathcal{N}_{n\mu}}(u)\mathrm{d}u=\frac{\lambda!}{(\lambda-\mu)!}\|\varphi\|_\infty\left(1-P\left(\|\mathcal{N}_{n\mu}\|\leq\sqrt{\|z\|}\right)\right)$.

Then by scaling-invariance with a multiplication by $t_z^f/\|z\|$, $\lim_{\|z\|\to\infty}\int_{\|u\|\leq\sqrt{\|z\|}}\varphi(u)$ $\left(\mathbb{E}\left[h(z,\mathcal{N}_n,u^\mu)\right]\right)^{\lambda-\mu}\prod_{i=1}^{\mu-1}h(z,u^{i+1},u^i)\prod_{i=1}^{\mu}p_{\mathcal{N}_n}(u^i)\mathrm{d}u-\frac{(\lambda-\mu)!}{\lambda!}\int\varphi(u)p_z^f(u)\mathrm{d}u=0$. Also, $\mathbb{E}\left[h(z,\mathcal{N}_n,u^\mu)\right]-\int_{\|w\|\leq\sqrt{\|z\|}}h(z,w,u^\mu)p_{\mathcal{N}_n}(w)\mathrm{d}w=\int_{\|w\|>\sqrt{\|z\|}}h(z,w,u^\mu)p_{\mathcal{N}_n}(w)\mathrm{d}w\leq 1-P\left(\|\mathcal{N}_n\|\leq\sqrt{\|z\|}\right)$. Hence along with the dominated convergence theorem, the lemma is proven. □

We are now ready to prove the proposition.
Let $\varphi:\mathbb{R}^{n\mu}\to\mathbb{R}$ be continuous and bounded. Using Lemma 11, it is enough to prove that $\lim_{\|z\|\to\infty}\int_{\|u\|\leq\sqrt{\|z\|}}\left(\int_{\|w\|\leq\sqrt{\|z\|}}h(z,w,u^\mu)p_{\mathcal{N}_n}(w)\mathrm{d}w\right)^{\lambda-\mu}\varphi(u)$ $\prod_{i=1}^{\mu-1}h(z,u^{i+1},u^i)\prod_{i=1}^{\mu}p_{\mathcal{N}_n}(u^i)\mathrm{d}u-\frac{(\lambda-\mu)!}{\lambda!}\int\varphi(u)p_z^{lf}(u)\mathrm{d}u=0$. We define the function g on the compact $\left(\mathcal{L}_{f,z_0^f}+\overline{\mathbb{B}(0,\delta_f)}\right)\times[0,\delta_f]$ as, for $(x,\rho)\in\left(\mathcal{L}_{f,z_0^f}+\overline{\mathbb{B}(0,\delta_f)}\right)\times(0,\delta_f]$, $g(x,\rho)=\int_{\|u\|\leq\frac{1}{\sqrt{\rho}}}\left(\int_{\|w\|\leq\frac{1}{\sqrt{\rho}}}1_{\theta(w,u^\mu,x)>0}\,p_{\mathcal{N}_n}(w)\mathrm{d}w\right)^{\lambda-\mu}\varphi(u)\prod_{i=1}^{\mu-1}\mathbb{1}_{\theta(u^{i+1},u^i,x)>0}p_{\mathcal{N}_{n\mu}(u)}\mathrm{d}u$, with $\theta(w,v,x)=(w-v)^\top\nabla f\left(x+t_x^f\rho(v+\tau_x^\rho(v,w)(w-v))\right)$ and $\tau_x^\rho(v^1,v^2)\in(0,1)$ defined thanks to the mean value theorem by $f(x+t_x^f\rho v^2)-f(x+t_x^f\rho v^1)=t_x^f\theta(v^2,v^1,x)$. For $x\in\mathcal{L}_{f,z_0^f}+\overline{\mathbb{B}(0,\delta_f)}$, $g(x,0)=\int\varphi(u)P\left((\mathcal{N}_n-u^\mu)^\top\nabla f(x)>0\right)^{\lambda-\mu}\prod_{i=1}^{\mu-1}\mathbb{1}_{(u^{i+1}-u^i)^\top\nabla f(x)>0}p_{\mathcal{N}_{n\mu}}(u)\mathrm{d}u$. Note that $g\left(t_z^f\frac{z}{\|z\|},0\right)=\frac{(\lambda-\mu)!}{\lambda!}\int\varphi(u)p_z^{lf}(u)\mathrm{d}u$. With Lemma 11, $\lim_{\|z\|\to\infty}g\left(t_z^f\frac{z}{\|z\|},\frac{1}{\|z\|}\right)-\frac{(\lambda-\mu)!}{\lambda!}\int\varphi(u)p_z^f(u)\mathrm{d}u=0$. Therefore it is enough to prove that g is uniformly continuous in order to obtain that $\frac{(\lambda-\mu)!}{\lambda!}\left(\lim_{\|z\|\to\infty}\int\varphi(u)p_z^f(u)\mathrm{d}u-\int\varphi(u)p_z^{lf}(u)\mathrm{d}u\right)$ is equal to $\lim_{\|z\|\to\infty}g\left(t_z^f\frac{z}{\|z\|},\frac{1}{\|z\|}\right)-g\left(t_z^f\frac{z}{\|z\|},0\right)$ which is equal to 0.

For $(x,\rho)\in\left(\mathcal{L}_{f,z_0^f}+\overline{\mathbb{B}(0,\delta_f)}\right)\times(0,\delta_f]$, for $u\in\overline{\mathbb{B}(0,1/\sqrt{\rho})}$, $w\in\overline{\mathbb{B}(0,1/\sqrt{\rho})}$, $\nabla f\Big(x+t_x^f\rho(u^\mu+\tau_x^\rho(u^\mu,w)(w-u^\mu))\Big)\neq 0$ since $x+t_x^f\rho(u^\mu+\tau_x^\rho(u^\mu,w)(w-u^\mu))\in\mathcal{L}_{f,z_0^f}+\overline{\mathbb{B}(0,2\delta_f)}$. Then the set $\{w\in\mathbb{R}^n;\theta(w,u^\mu,x)=0\}$ is Lebesgue negligible. In addition, the function $y\mapsto\mathbb{1}_{y>0}$ is continuous on $\mathbb{R}\setminus\{0\}$, therefore for almost all w, $(x,\rho,u^\mu)\mapsto\mathbb{1}_{\|w\|\leq\frac{1}{\sqrt{\rho}}}\mathbb{1}_{\theta(w,u^\mu,x)}p_{\mathcal{N}_n}(w)$ is continuous and bounded by the integrable function $p_{\mathcal{N}_n}$. Then by domination, for almost all u,

$(x,\rho) \mapsto \mathbb{1}_{\|u\| \leq \frac{1}{\sqrt{\rho}}} \left(\int_{\|w\| \leq \frac{1}{\sqrt{\rho}}} \mathbb{1}_{\theta(w,u^\mu,x)>0}\, p_{\mathcal{N}_n}(w)\mathrm{d}w \right)^{\lambda-\mu}$ is continuous. Similarly $(x,\rho) \mapsto \mathbb{1}_{\|u\| \leq \frac{1}{\sqrt{\rho}}} \prod_{i=1}^{\mu-1} \mathbb{1}_{\theta(u^{i+1},u^i,x)>0}$ is continuous for almost all u. Therefore g is continuous on $\left(\mathcal{L}_{f,z_0^f} + \overline{\mathbb{B}(0,\delta_f)} \right) \times (0,\delta_f]$, and for all $x \in \mathcal{L}_{f,z_0^f} + \overline{\mathbb{B}(0,\delta_f)}$, $\lim_{\rho\to 0} g(x,\rho)$ exists and equals

$$\int \lim_{\rho\to 0} \mathbb{1}_{\|u\| \leq \frac{1}{\sqrt{\rho}}} \varphi(u) \prod_{i=1}^{\mu-1} \mathbb{1}_{\theta(u^{i+1},u^i,x)>0} \left(\int_{\|w\| \leq \frac{1}{\sqrt{\rho}}} 1_{\theta(w,u^\mu,x)>0}\, p_{\mathcal{N}_n}(w)\mathrm{d}w \right)^{\lambda-\mu} p_{\mathcal{N}_{n\mu}(u)}\mathrm{d}u$$

which is equal to $g(x,0)$. Finally g is continuous on the compact $\left(\mathcal{L}_{f,z_0^f} + \overline{\mathbb{B}(0,\delta_f)} \right) \times [0,\delta_f]$; it is thereby uniformly continuous on that compact.