Markov chain Analysis of Evolution Strategies

Alexandre Chotard

UNIVERSITÉ PARIS SUD

Inria

UNIVERSITÉ PARIS-SUD

ECOLE DOCTORALE D'INFORMATIQUE
LABORATOIRE INRIA SACLAY

DISCIPLINE : INFORMATIQUE

THÈSE DE DOCTORAT

Soutenue le 24 Septembre 2015 par

Alexandre Chotard

Titre :
Analyse Markovienne des Stratégies d'Évolution

Directeur de thèse :	Nikolaus Hansen	Directeur de recherche (INRIA Saclay)
Co-directeur de thèse :	Anne Auger	Chargée de recherche (INRIA Saclay)
Composition du jury :		
Rapporteurs :	Dirk Arnold	Professor (Dalhousie University)
	Tobias Glasmachers	Junior Professor (Ruhr-Universität Bochum)
Examinateurs :	Gersende Fort	Directrice de Recherche (CNRS)
	François Yvon	Professeur (Université Paris-Sud)

Abstract

In this dissertation an analysis of Evolution Strategies (ESs) using the theory of Markov chains is conducted. We first develop sufficient conditions for a Markov chain to have some basic properties. We then analyse different ESs through underlying Markov chains. From the stability of these underlying Markov chains we deduce the log-linear divergence or convergence of these ESs on a linear function, with and without a linear constraint, which are problems that can be related to the log-linear convergence of ESs on a wide class of functions. More specifically, we first analyse an ES with cumulative step-size adaptation on a linear function and prove the log-linear divergence of the step-size; we also study the variation of the logarithm of the step-size, from which we establish a necessary condition for the stability of the algorithm with respect to the dimension of the search space. Then we study an ES with constant step-size and with cumulative step-size adaptation on a linear function with a linear constraint, using resampling to handle unfeasible solutions. We prove that with constant step-size the algorithm diverges, while with cumulative step-size adaptation, depending on parameters of the problem and of the ES, the algorithm converges or diverges log-linearly. We then investigate the dependence of the convergence or divergence rate of the algorithm with parameters of the problem and of the ES. Finally we study an ES with a sampling distribution that can be non-Gaussian and with constant step-size on a linear function with a linear constraint. We give sufficient conditions on the sampling distribution for the algorithm to diverge. We also show that different covariance matrices for the sampling distribution correspond to a change of norm of the search space, and that this implies that adapting the covariance matrix of the sampling distribution may allow an ES with cumulative step-size adaptation to successfully diverge on a linear function with any linear constraint.

Contents

Notations

We denote $\mathbb{R}$ the set of real numbers, $\mathbb{R}_+$ the set of non-negative numbers, $\mathbb{R}_-$ the set of non-positive numbers, $\mathbb{N}$ the set of non-negative integers. For $n \in \mathbb{N}\backslash\{0\}$, $\mathbb{R}^n$ denotes the set of n-dimensional real vectors. For A a subset of $\mathbb{R}^n$, A^* denotes $A\backslash\{\mathbf{0}\}$, A^c denotes the complementary of A, $\mathbf{1}_A$ the indicator function of A, and $\Lambda_n(A)$ the Lebesgue measure on R^n of A. For A a finite set, we denote $\#A$ its cardinal. For A a set, 2^A denotes the power set of A. For F a family of subsets of $\mathbb{R}^n$, we denote $\sigma(F)$ the σ-algebra generated by F. Let f be a function defined on an open set of $\mathbb{R}^n$ and valued in $\mathbb{R}^m$, and take $p \in \mathbb{N}$, we say that f is a C^p function if it is continuous, and p-times continuously differentiable; if f is differentiable, we denote $D_{\boldsymbol{x}} f$ the differential of f with respect to $\boldsymbol{x} \in \mathbb{R}^n$; if $m = 1$ and f is differentiable, we denote $\nabla_{\boldsymbol{x}} f$ its gradient at $\boldsymbol{x} \in \mathbb{R}^n$. For $(a, b) \in \mathbb{N}^2$, $[a..b]$ denotes the set $\{i \in \mathbb{N} | a \leq i \leq b\}$. For $\boldsymbol{x} \in \mathbb{R}^n$, $\boldsymbol{x}^T$ denotes $\boldsymbol{x}$ transposed. For $n \in \mathbb{N}^*$, $\boldsymbol{Id}_n$ is the n-dimensional identity matrix. We denote $\mathcal{N}(0, 1)$ the standard normal law, and for $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{C}$ a covariance matrix of order n, $\mathcal{N}(\boldsymbol{x}, \boldsymbol{C})$ denotes the multivariate normal law of mean $\boldsymbol{x}$ and covariance matrix $\boldsymbol{C}$. For $\boldsymbol{X}$ a random vector, $\mathbf{E}(\boldsymbol{X})$ denotes the expected value of $\boldsymbol{X}$, and for π a distribution, $\boldsymbol{X} \sim \pi$ means that $\boldsymbol{X}$ has distribution π. For $(a, b) \in \mathbb{N} \times \mathbb{N}^*$, $a \mod b$ denotes a modulo b. For f and g two real-valued functions defined on $\mathbb{N}$, we write that $f \sim g$ when f is equal to g asymptotically, that $f = O(g)$ if there exists $C \in \mathbb{R}_+^*$ and $n_0 \in \mathbb{N}$ such that $|f(n)| \leq C|g(n)|$ for all $n \geq n_0$, and that $f = \Theta(g)$ if $f = O(g)$ and $g = O(f)$. For $\boldsymbol{x} \in \mathbb{R}^n$, $\|\boldsymbol{x}\|$ denotes the euclidean norm of $\boldsymbol{x}$, and for $r \in \mathbb{R}_+^*$, $B(\boldsymbol{x}, r)$ denotes the open ball for the euclidean norm centred in $\boldsymbol{x}$ of radius r, and for $i \in [1..n]$, $[\boldsymbol{x}]_i$ denotes the i^{th} coordinate of $\boldsymbol{x}$ in the canonical basis. We use the acronym i.i.d. for independent and identically distributed. For $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ a sequence of random vectors and $\boldsymbol{Y}$ a random vectors, we denote $\boldsymbol{X}_t \xrightarrow[t\to+\infty]{a.s.} \boldsymbol{Y}$ when the sequence $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ converges almost surely to $\boldsymbol{Y}$, and $\boldsymbol{X}_t \xrightarrow[t\to+\infty]{P} \boldsymbol{Y}$ when the sequence $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ converges in probability to $\boldsymbol{Y}$.

Chapter 1

Preamble

Optimization problems are frequently encountered in both science and industry. They consist in finding the optimum of a real-valued function f called the objective function and defined on a search space X. Depending on this search space, they can be broadly categorized into discrete or continuous optimization problems. Evolution Strategies (ESs) are stochastic continuous optimization algorithms that have been successfully applied to a wide range of real-world problem. These algorithms adapt a sampling distribution of the form $\boldsymbol{x} + \sigma H$ where H is a distribution with mean $\mathbf{0}$, generally taken as $\mathcal{N}(\mathbf{0}, \boldsymbol{C})$ a multivariate Gaussian distribution with covariance matrix $\boldsymbol{C}$; $\boldsymbol{x} \in X$ is the mean of the sampling distribution and $\sigma \in \mathbb{R}_+^*$ is called the step-size, and controls the standard deviation of the sampling distribution. ESs proceed to sample a population of points, that they rank according to their f-value, and use these points and their rankings to update the sampling distribution.

ESs are known in practice to achieve log-linear convergence (i.e. the distance to the optimum decreases exponentially fast, see Section 2.1) on a wide class of functions. To achieve a better understanding of ESs, it is important to know the convergence rate and its dependence to the search problem (e.g. the dimension of the search space) or in different update rules or parameters of ESs. Log-linear convergence has been shown for different ESs on the sphere function $f_{\text{sphere}} : \boldsymbol{x} \in \mathbb{R}^n \mapsto \|\boldsymbol{x}\|^2$ using tools from the theory of Markov chains (see [18, 24, 33]) by proving the positivity, Harris recurrence or geometric ergodicity of an underlying Markov chain (these concepts are defined in Section 1.2). A methodology on a wide class of functions called scaling-invariant (see (2.33) for a definition of scaling invariant functions) for proving the geometric ergodicity of underlying Markov chains from which the log-linear convergence of the algorithm can be deduced is proposed in [25], and has been used to prove the log-linear convergence of a specific ES [24] on positively homogeneous functions (see (2.34) for a definition of positively homogeneous functions). In [2] the local convergence of a continuous-time ES is shown on C^2 functions using ordinary differential equations. In both [24] and [2] a shared assumption is that the standard deviation σ_t of the sampling distribution diverges log-linearly on the linear function, making the study of ESs on the linear function a key to the convergence of ESs on a wide range of functions.

The ergodicity (or more precisely, f-ergodicity as defined in 1.2.9) of Markov chains underlying ESs is a crucial property regarding Monte-Carlo simulations, as it implies that a law of large

numbers applies, and so shows that Monte-Carlo simulations provide a consistent estimator of $\mathbf{E}_\pi(f(\Phi_0))$, where $(\Phi_t)_{t\in\mathbb{N}}$ is a f-ergodic Markov chain and π is its invariant measure, as defined in 1.2.8. This allows the use of Monte-Carlo simulations to estimate the convergence rate of the algorithm, and evaluate the influence of different parameters on this convergence rate.
The work presented in this thesis can be divided in two parts: the contributions in Chapter 3 improve techniques from Markov chain theory so that they can be applied to problems met in continuous optimization and allow us to analyse easily a broader class of algorithms, and the contributions in Chapter 4 analyse ESs on different linear problems.

1.1 Overview of Contributions

1.1.1 Sufficient conditions for φ-irreducibility, aperiodicity and T-chain property

In order to show the ergodicity of a Markov chain $\mathbf{\Phi} = (\Phi_t)_{t\in\mathbb{N}}$ valued in an open space $X\subset\mathbb{R}^n$, we use some basic Markov chain properties (namely φ-irreducibility, aperiodicity, and that compact sets are small sets for the chain, which are concepts defined in 1.2). For some Markov chains arising from algorithms that we want to analyse, showing these basic properties turned out to be unexpectingly difficult as the techniques used with success in other scenarios failed, as outlined in Section 4.1. In [98, Chapter 7] powerful tools can be found and be used to show the basic properties we require. However, [98, Chapter 7] assumes that the Markov chain of interest follows a certain model, namely that there exists an open set $O\subset\mathbb{R}^p$, a C^∞ function $F : X\times O\to X$ and $(\boldsymbol{U}_t)_{t\in\mathbb{N}^*}$ a sequence of i.i.d. random vectors valued in O and admitting a lower semi-continuous density, such that $\Phi_{t+1} = F(\Phi_t, \boldsymbol{U}_{t+1})$ for all $t\in\mathbb{N}$. For some of the Markov chains that we analyse we cannot find an equivalent model: the corresponding function F is not even continuous, or the random vectors $(\boldsymbol{U}_t)_{t\in\mathbb{N}^*}$ are not i.i.d.. However, in Chapter 3 which contains the article [42] soon to be submitted to the journal Bernoulli we show that we can adapt the results of [98, Chapter 7] to a more general model $\Phi_{t+1} = F(\Phi_t, \boldsymbol{W}_{t+1})$, with F typically a C^1 function, $\boldsymbol{W}_{t+1} = \alpha(\Phi_t, \boldsymbol{U}_{t+1})$ and $(\boldsymbol{U}_t)_{t\in\mathbb{N}^*}$ are i.i.d. such that $\alpha(\boldsymbol{x}, \boldsymbol{U}_1)$ admits a lower semi-continuous density. The function α is in our cases typically not continuous, and the sequence $(\boldsymbol{W}_t)_{t\in\mathbb{N}^*}$ is typically not i.i.d.. We then use these results to solve cases that we could not solve before.

1.1.2 Analysis of Evolution Strategies

In Chapter 4 we analyse ESs on different problems.
In Section 4.2 we present an analysis of the so-called $(1,\lambda)$-CSA-ES algorithm on a linear function. The results are presented in a technical report [43] containing [46] which was published at the conference Parallel Problem Solving from Nature in 2012 and including the full proofs of the propositions found in [46], and a proof of the log-linear divergence of the algorithm. We prove that the step-size of the algorithm diverges log-linearly, which is the desired behaviour on a linear function. The divergence rate is explicitly given, which allow us to see how it depends of the parameters of the problem or of the algorithm. Also, a study

of the variance of the logarithm of the step-size is conducted, and the scaling of the variance with the dimension gives elements as how to adapt some parameters of the algorithm with the dimension.

In Section 4.3 we present two analyses of a $(1,\lambda)$-ES on a linear function with a linear constraint, handling the constraint through resampling unfeasible points.

The first analysis in Section 4.3.1 is presented in [45] which was accepted for publication at the Evolutionary Computation Journal in 2015, and is an extension of [44] which was published at the conference Congress on Evolutionary Computation in 2014. It first shows that a $(1,\lambda)$-ES algorithm with constant step-size diverges almost surely. Then for the $(1,\lambda)$-ES with cumulative step-size adaptation (see 2.3.8) it shows the geometric ergodicity of the Markov chain composed of the distance from the mean of the sampling distribution to the constraint normalized by the step-size. This geometric ergodicity justifies the use of Monte-Carlo simulations of the convergence rate of the step-size, which shows that when the angle θ between the gradients of the constraint and of the objective function is close to 0, the step-size converges log-linearly, while for values close enough to $\pi/2$ the algorithm diverges log-linearly. Log-linear divergence being desired here, the algorithm fails to solve the problem for small values of θ, and otherwise succeeds. The paper then analyses how its parameters affect the convergence rate and the critical value of θ which triggers convergence or divergence of the algorithm.

The second analysis in Section 4.3.2 is presented in a technical report containing [47], published at the conference Parallel Problem Solving from Nature in 2014, and the full proofs of the propositions found in [47]. It analyses a $(1,\lambda)$-ES with constant step-size and general (i.e. not necessary Gaussian) sampling distribution. It establishes sufficient conditions on the sampling distribution for the positivity, Harris recurrence and geometric ergodicity of the Markov chain composed of the distance from the mean of the sampling distribution to the constraint. The positivity and Harris recurrence is then used to apply a law of large numbers and deduce the divergence of the algorithm. It is then shown that changing the covariance matrix of the sampling distribution is equivalent to a change of norm which imply a change of the angle between the gradients of the constraint and of the function. This relates to the results presented in 4.3.1, showing that on this problem and if the covariance matrix is correctly adapted then the cumulative step-size adaptation is successful. Finally, sufficient conditions on the marginals of the sampling distribution and the copula combining them are given to get the absolute continuity of the sampling distribution.

1.2 A short introduction to Markov Chain Theory

Markov chain theory offers useful tools to show the log-linear convergence of optimization algorithms, and justifying the use of Monte Carlo simulations to estimate convergence rates. Markov chains are key to the results of this thesis, and therefore we give in this section an introduction to the concepts that we will be using throughout the thesis.

1.2.1 A definition of Markov chains through transition kernels

Let X be an open set of $\mathbb{R}^n$ that we call the state space, equipped with its borel σ-algebra $\mathscr{B}(X)$. A function $P : X \times \mathscr{B}(X) \to \mathbb{R}$ is called a **kernel** if

- for all $\boldsymbol{x} \in X$, the function $A \in \mathscr{B}(X) \mapsto P(\boldsymbol{x}, A)$ is a measure,
- for all $A \in \mathscr{B}(X)$ the function $\boldsymbol{x} \in X \mapsto P(\boldsymbol{x}, A)$ is a measurable function.

Furthermore, if for all $\boldsymbol{x} \in X$, $P(\boldsymbol{x}, X) \leq 1$ we call P a **substochastic transition kernel**, and if for all $\boldsymbol{x} \in X$, $P(\boldsymbol{x}, X) = 1$, we call P a **probability transition kernel**, or simply a transition kernel. Intuitively, for a specific sequence of random variables $(\Phi_t)_{t\in\mathbb{N}}$, the value $P(\boldsymbol{x}, A)$ represents the probability that $\Phi_{t+1} \in A$ knowing that $\Phi_t = \boldsymbol{x}$. Given a transition kernel P, we define P^1 as P, and inductively for $t \in \mathbb{N}^*$, P^{t+1} as

$$P^{t+1}(\boldsymbol{x}, A) = \int_X P^t(\boldsymbol{x}, d\boldsymbol{y}) P(\boldsymbol{y}, A) \ . \tag{1.1}$$

Let $(\Omega, \mathscr{B}(\Omega), P_0)$ be a probability space, and $\boldsymbol{\Phi} = (\Phi_t)_{t\in\mathbb{N}}$ be a sequence of random variables defined on Ω and valued in X, and let P be a probability transition kernel. Denote $(\mathscr{F}_t)_{t\in\mathbb{N}}$ the filtration such that $\mathscr{F}_t := \sigma(\Phi_k \mid k \leq t)$. Following [118, Definition 2.3], we say that $\boldsymbol{\Phi}$ is a **time-homogeneous Markov chain** with probability transition kernel P if for all $t \in \mathbb{N}^*$, $k \in [0..t-1]$ and any bounded real-valued function f defined on X,

$$\mathbf{E}_0\left(f(\Phi_t)\middle|\mathscr{F}_k\right) = \int_X f(\boldsymbol{y}) P^{t-k}(\cdot, d\boldsymbol{y}) \qquad P_0 - \text{a.s.} \ , \tag{1.2}$$

where $\mathbf{E}_0$ is the expectation operator with respect to P_0.
Less formally the expected value of $f(\Phi_t)$, knowing all the past information of $(\Phi_i)_{i\in[0..k]}$ and that Φ_k is distributed according to P_0, is equal to the expected value of $f(\Phi_{t-k})$ with Φ_0 distributed according to P_0. The value $P^t(\boldsymbol{x}, A)$ represents the probability of the Markov chain $\boldsymbol{\Phi}$ to be in A, t time steps after starting from $\boldsymbol{x}$.

1.2.2 φ-irreducibility

A Markov chain is said **φ-irreducible** if there exists a non trivial measure φ on $\mathscr{B}(X)$ such that for all $A \in \mathscr{B}(X)$

$$\varphi(A) > 0 \Rightarrow \forall \boldsymbol{x} \in X, \ \sum_{t\in\mathbb{N}^*} P^t(\boldsymbol{x}, A) > 0 \ . \tag{1.3}$$

Every point from the support of φ is **reachable** [98, Lemma 6.1.4], meaning any neighbourhood of a point in the support has a positive probability of being reached from anywhere in the state space. This ensures that the state space cannot be cut into disjoints sets that would never communicate through the Markov chain with each other.
A φ-irreducible Markov chain admits the existence of a **maximal irreducibility measure** ([98, Proposition 4.2.2]), that we denote ψ, which dominates any other irreducibility measure. This

allows us to define $\mathscr{B}^+(X)$, the set of sets with positive ψ-measure:

$$\mathscr{B}^+(X) = \{A \in \mathscr{B}(X) \mid \psi(A) > 0\} \ . \tag{1.4}$$

1.2.3 Small and petite sets

A set $C \in \mathscr{B}(X)$ is called a **small set** if there exists $m \in \mathbb{N}^*$ and ν_m a non-trivial measure on $\mathscr{B}(X)$ such that

$$P^m(\boldsymbol{x}, A) \geq \nu_m(A) \ \text{, for all } \boldsymbol{x} \in C \text{ and for all } A \in \mathscr{B}(X) \ . \tag{1.5}$$

A set $C \in \mathscr{B}(X)$ is called a **petite set** if there exists α a probability distribution on $\mathbb{N}^*$ and ν_α a non-trivial measure on $\mathscr{B}(X)$ such that

$$\sum_{t \in \mathbb{N}} P^t(\boldsymbol{x}, A)\alpha(t) \geq \nu_\alpha(A) \ \text{, for all } \boldsymbol{x} \in C \text{ and for all } A \in \mathscr{B}(X) \ , \tag{1.6}$$

where $P^0(\boldsymbol{x}, A)$ is defined as a Dirac distribution on $\{\boldsymbol{x}\}$.
Small sets are petite sets; the converse is true for φ-irreducible aperiodic Markov chains [98, Theorem 5.5.7].

1.2.4 Periodicity

Suppose that $\mathbf{\Phi}$ is a ψ-irreducible Markov chain, and take $C \in \mathscr{B}^+(X)$ a ν_m-small set. The period of the Markov chain can be defined as the greatest common divisor of the set

$$E_C = \{k \in \mathbb{N}^* | C \text{ is a } \nu_k\text{-small set with } \nu_k = a_k \nu_m \text{ for some } a_k \in \mathbb{R}_+^*\} \ .$$

According to [98, Theorem 5.4.4] there exists then disjoint sets $(D_i)_{i \in [0..d-1]} \in \mathscr{B}(X)^d$ called a d-cycle such that

1. $P(\boldsymbol{x}, D_{i+1 \mod d}) = 1$ for all $\boldsymbol{x} \in D_i$,
2. $\psi\left(\left(\bigcup_{i=0}^{d-1} D_i\right)^c\right) = 0$.

This d-cycle is maximal in the sense that for any other $\tilde{d}$-cycle $(\tilde{D}_i)_{i \in [0..d-1]}$, $\tilde{d}$ divides d, and if $\tilde{d} = d$ then up to a reordering of indexes, $\tilde{D}_i = D_i$ ψ-almost everywhere.
If $d = 1$, then the Markov chain is called **aperiodic**. For a φ-irreducible aperiodic Markov chain, petite sets are small sets [98, Theorem 5.5.7].

1.2.5 Feller chains and T-chains

These two properties concern the lower semi-continuity of the function $P(\cdot, A) : \boldsymbol{x} \in X \mapsto P(\boldsymbol{x}, A)$, and help to identify the petite sets and small sets of the Markov chain.
A φ-irreducible Markov chain is called a **(weak-)Feller** Markov chain if for all open set $O \in \mathscr{B}(X)$, the function $P(\cdot, O)$ is lower-semi continuous.

If there exists α a distribution on $\mathbb{N}$ and a substochastic transition kernel $T : X \times \mathscr{B}(X) \to \mathbb{R}$ such that

- for all $\boldsymbol{x} \in X$ and $A \in \mathscr{B}(X)$, $K_\alpha(\boldsymbol{x}, A) := \sum_{t\in\mathbb{N}} \alpha(t) P^t(\boldsymbol{x}, A) \geq T(\boldsymbol{x}, A)$,
- for all $\boldsymbol{x} \in X$, $T(\boldsymbol{x}, X) > 0$,

then the Markov chain is called a T-**chain**.
According to [98, Theorem 6.0.1] a φ-irreducible Markov chain for which the support of φ has non-empty interior is a φ-irreducible T-chain. And a φ-irreducible Markov chain is a T-chain if and only if all compact sets are petite sets.

1.2.6 Associated deterministic control model

According to [72, p.24], for $\boldsymbol{\Phi}$ a time-homogeneous Markov chain on an open state space X, there exists an open measurable space Ω, a function $F : X \times \Omega \to X$ and $(\boldsymbol{U}_t)_{t\in\mathbb{N}^*}$ a sequence of i.i.d. random variables valued in Ω such that

$$\Phi_{t+1} = F(\Phi_t, \boldsymbol{U}_{t+1}) \ . \tag{1.7}$$

The transition probability kernel P then writes

$$P(\boldsymbol{x}, A) = \int_\Omega \mathbf{1}_A(F(\boldsymbol{x}, \boldsymbol{u}))\, \mu(d\boldsymbol{u}) \ , \tag{1.8}$$

where μ is the distribution of $\boldsymbol{U}_t$.
Conversely, given a random variable Φ_0 taking values in X, the sequence $(\Phi_t)_{t\in\mathbb{N}}$ can be defined through (1.7), and it is easy to check that it is a Markov chain.
From this function F we can define F^0 as the identity $F^0 : \boldsymbol{x} \in X \mapsto \boldsymbol{x}$, F as F^1 and inductively F^{t+1} for $t \in \mathbb{N}^*$ as

$$F^{t+1}(\boldsymbol{x}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_{t+1}) := F^t(F(\boldsymbol{x}, \boldsymbol{u}_1), \boldsymbol{u}_2, \ldots, \boldsymbol{u}_{t+1}) \ . \tag{1.9}$$

If $\boldsymbol{U}_1$ admits a density p, we can define the **control set** $\Omega_{\boldsymbol{w}} := \{\boldsymbol{u} \in \Omega | p(\boldsymbol{u}) > 0\}$, which allow us to define the **associated deterministic control model**, denoted $CM(F)$, as the deterministic system

$$\boldsymbol{x}_t = F^t(\boldsymbol{x}_0, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_t) \ , \quad \forall\, t \in \mathbb{N} \tag{1.10}$$

where $\boldsymbol{u}_k \in \Omega_{\boldsymbol{w}}$ for all $k \in \mathbb{N}^*$. , and for $\boldsymbol{x} \in X$ we define the set of states reachable from $\boldsymbol{x}$ at time $k \in \mathbb{N}$ from $CM(F)$ as $A_+^0(\boldsymbol{x}) = \{\boldsymbol{x}\}$ when $k = 0$ and otherwise

$$A_+^k(\boldsymbol{x}) := \{F^k(\boldsymbol{x}, \boldsymbol{u}_1, \ldots, \boldsymbol{u}_k) | \boldsymbol{u}_i \in \Omega_{\boldsymbol{w}} \text{ for all } i \in [1..k]\} \ . \tag{1.11}$$

The set of states reachable from $\boldsymbol{x}$ from $CM(F)$ is defined as

$$A_+(\boldsymbol{x}) := \bigcup_{k\in\mathbb{N}} A_+^k(\boldsymbol{x}) \ . \tag{1.12}$$

The control model is said to be **forward accessible** if for all $\boldsymbol{x} \in X$, the set $A_+(\boldsymbol{x})$ has non-empty interior. A point $\boldsymbol{x}^* \in X$ is called a **globally attracting state** if

$$\boldsymbol{x}^* \in \bigcap_{N \in \mathbb{N}^*} \overline{\bigcup_{k=N}^{+\infty} A_+^k(\boldsymbol{y})} \quad \text{for all } \boldsymbol{y} \in X \ . \tag{1.13}$$

In [98, Chapter 7], the function F of (1.7) is supposed C^∞, the random element $\boldsymbol{U}_1$ is assumed to admit a lower semi-continuous density p, and the control model is supposed to be forward accessible. In this context, the Markov chain is shown to be a T-chain [98, Proposition 7.1.5], and in [98, Proposition 7.2.5 and Theorem 7.2.6] φ-irreducibility is proven equivalent to the existence of a globally attracting state. Still in this context, when the control set $\Omega_{\boldsymbol{w}}$ is connected and that there exists a globally attracting state, the aperiodicity of the Markov chain is proven to be implied by the connectedness of the set $\overline{A_+(\boldsymbol{x}^*)}$ [98, Proposition 7.3.4 and Theorem 7.3.5]. Although these results are strong and useful ways to show the irreducibility and aperiodicity or T-chain property of a Markov chain, we cannot apply them on most of the Markov chains studied in Chapter 4, as the transition functions F modelling our Markov chains through (1.7) are not C^∞, but instead are discontinuous due to the selection mechanism in the ESs studied. A part of the contributions of this thesis is to adapt and generalize the results of [98, Chapter 7] to be usable in our problems (see Chapter 3).

1.2.7 Recurrence, Transience and Harris recurrence

A set $A \in \mathscr{B}(X)$ is called **recurrent** if for all $\boldsymbol{x} \in A$, the Markov chain $(\Phi_t)_{t\in\mathbb{N}}$ leaving from $\Phi_0 = \boldsymbol{x}$ will return in average an infinite number of times to A. More formally, A is recurrent if

$$\mathbf{E}\left(\sum_{t\in\mathbb{N}^*} \mathbf{1}_A(\Phi_t) \;\middle|\; \Phi_0 = \boldsymbol{x} \right) = \infty \ , \text{ for all } \boldsymbol{x} \in A \ . \tag{1.14}$$

A ψ-irreducible Markov chain is called **recurrent** if for all $A \in \mathscr{B}^+(X)$, A is recurrent.

The mirrored concept is called transience. A set $A \in \mathscr{B}(X)$ is called **uniformly transient** if there exists $M \in \mathbb{R}$ such that

$$\mathbf{E}\left(\sum_{t\in\mathbb{N}^*} \mathbf{1}_A(\Phi_t) \;\middle|\; \Phi_0 = \boldsymbol{x} \right) \leq M \ , \text{ for all } \boldsymbol{x} \in A \ . \tag{1.15}$$

A ψ-irreducible Markov chain is called **transient** if there exists a countable cover of the state space X by uniformly transient sets. According to [98, Theorem 8.0.1], a ψ-irreducible Markov chain is either recurrent or transient.

A condition stronger than recurrence is Harris recurrence. A set $A \in \mathscr{B}(X)$ is called **Harris recurrent** if for all $\boldsymbol{x} \in A$ the Markov chain leaving from $\boldsymbol{x}$ will return almost surely an infinite number of times to A, that is

$$\Pr\left(\sum_{t\in\mathbb{N}^*} \mathbf{1}_A(\Phi_t) = \infty \;\middle|\; \Phi_0 = \boldsymbol{x} \right) = 1 \ , \text{ for all } \boldsymbol{x} \in A \ . \tag{1.16}$$

A ψ-irreducible Markov chain is called **Harris recurrent** if for all $A \in \mathscr{B}^+(X)$, A is Harris recurrent.

1.2.8 Invariant measure and positivity

A σ-finite measure π on $\mathscr{B}(X)$ is called invariant if

$$\pi(A) = \int_X \pi(d\boldsymbol{x}) P(\boldsymbol{x}, A) \ , \text{ for all } A \in \mathscr{B}(X). \tag{1.17}$$

Therefore if $\Phi_0 \sim \pi$, then for all $t \in \mathbb{N}$, $\Phi_t \sim \pi$.

For $f : X \to \mathbb{R}$ a function, we denote $\pi(f)$ the expected value

$$\pi(f) := \int_X f(\boldsymbol{x}) \pi(d\boldsymbol{x}) \ . \tag{1.18}$$

According to [98, Theorem 10.0.1] a φ-irreducible recurrent Markov chain admits a unique (up to a multiplicative constant) invariant measure. If this measure is a probability measure, we call $\mathbf{\Phi}$ a **positive** Markov chain.

1.2.9 Ergodicity

For ν a signed measure on $\mathscr{B}(X)$ and $f : X \to \mathbb{R}$ a positive function, we define $\|\cdot\|_f$ a norm on signed measures via

$$\|\nu\|_f := \sup_{|g| \leq f} \left| \int_X g(\boldsymbol{x}) \nu(d\boldsymbol{x}) \right| \ . \tag{1.19}$$

Let $f : X \to \mathbb{R}$ be a function lower-bounded by 1. We call $\mathbf{\Phi}$ a **f-ergodic** Markov chain if it is a positive Harris recurrent Markov chain with invariant probability measure π, that $\pi(f)$ is finite, and for any initial condition $\Phi_0 = \boldsymbol{x} \in X$,

$$\|P^t(\boldsymbol{x}, \cdot) - \pi\|_f \underset{t \to +\infty}{\longrightarrow} 0 \ . \tag{1.20}$$

We call $\mathbf{\Phi}$ a **f-geometrically ergodic** Markov chain if it is a positive Harris recurrent Markov chain with invariant probability measure π, that $\pi(f)$ is finite, and if there exists $r_f \in (1, +\infty)$ such that for any initial condition $\Phi_0 = \boldsymbol{x} \in X$,

$$\sum_{t \in \mathbb{N}^*} r_f^t \|P^t(\boldsymbol{x}, \cdot) - \pi\|_f < \infty \ . \tag{1.21}$$

We also call $\mathbf{\Phi}$ a **ergodic** (resp. **geometrically ergodic**) Markov chain if there exists a function $f : X \to \mathbb{R}$ lower bounded by 1 such that $\mathbf{\Phi}$ is f-ergodic (resp. f-geometrically ergodic).

1.2.10 Drift conditions

Drift conditions are powerful tools to show that a Markov chain is transient, recurrent, positive or ergodic. They rely on a potential or drift function $V : X \to \mathbb{R}_+$, and the mean drift

$$\Delta V(\boldsymbol{x}) := \mathbf{E}\left(V(\Phi_{t+1}) \mid \Phi_t = \boldsymbol{x}\right) - V(\boldsymbol{x}) \ . \tag{1.22}$$

A positive drift outside a set $C_V(r) := \{\boldsymbol{x} \in X \mid V(\boldsymbol{x}) \leq r\}$ means that the Markov chain tends to get away from the set, and indicates transience. Formally, for a φ-irreducible Markov chain, if $V : \boldsymbol{X} \to \mathbb{R}_+$ is a bounded function and that there exists $r \in \mathbb{R}_+$ such that both sets $C_V(r)$ and $C_V(r)^c$ are in $\mathscr{B}^+(X)$ and that

$$\Delta V(\boldsymbol{x}) > 0 \text{ for all } \boldsymbol{x} \in C_V(r)^c \ , \tag{1.23}$$

then the Markov chain is transient [98, Theorem 8.4.2].

Conversely, a negative drift is linked to recurrence and Harris recurrence. For a φ-irreducible Markov chain, if there exists a function $V : X \to \mathbb{R}_+$ such that $C_V(r)$ is a petite set for all $r \in \mathbb{R}$, and if there exists a petite set $C \in \mathscr{B}(X)$ such that

$$\Delta V(\boldsymbol{x}) \leq 0 \text{ for all } \boldsymbol{x} \in C^c \ , \tag{1.24}$$

then the Markov chain is Harris recurrent [98, Theorem 9.1.8].

A stronger drift condition ensures the positivity and f-ergodicity of the Markov chain: for a φ-irreducible aperiodic Markov chain and $f : X \to [1, +\infty)$, if there exists a function $V : X \to \mathbb{R}_+$, $C \in \mathscr{B}(X)$ a petite set and $b \in \mathbb{R}$ such that

$$\Delta V(\boldsymbol{x}) \leq -f(\boldsymbol{x}) + b\mathbf{1}_C(\boldsymbol{x}) \ , \tag{1.25}$$

then the Markov chain is positive recurrent with invariant probability measure π, and f-ergodic [98, Theorem 14.0.1].

Finally, a stronger drift condition ensures a geometric convergence of the transition kernel $P^t(\boldsymbol{x}, \cdot)$ to the invariant measure. For a φ-irreducible aperiodic Markov chain, if there exists a function $V : X \to [1, +\infty)$, $C \in \mathscr{B}(X)$ a petite set, $b \in \mathbb{R}$ and $\beta \in \mathbb{R}_+^*$ such that

$$\Delta V(\boldsymbol{x}) \leq -\beta V(\boldsymbol{x}) + b\mathbf{1}_C(\boldsymbol{x}) \ , \tag{1.26}$$

then the Markov chain is positive recurrent with invariant probability measure π, and V-geometrically ergodic [98, Theorem 15.0.1].

1.2.11 Law of Large numbers for Markov chains

Let $\mathbf{\Phi}$ be a positive Harris recurrent Markov chain with invariant measure π, and take $g : X \to \mathbb{R}$ a function such that $\pi(|g|) < \infty$. Then according to [98, Theorem 17.0.1]

$$\frac{1}{t}\sum_{k=1}^{t} g(\Phi_k) \xrightarrow[t\to+\infty]{a.s} \pi(g) \ . \tag{1.27}$$

Chapter 2

Introduction to Black-Box Continuous Optimization

This chapter intends to be a general introduction to black-box continuous optimization by presenting different optimization techniques, problems, and results with a heavier focus on Evolution Strategies. We denote $f : X \subset \mathbb{R}^n \to \mathbb{R}$ the function to be optimized, which we call the **objective function**, and assume w.l.o.g. the problem to be to minimize f[1] by constructing a sequence $(\boldsymbol{x}_t)_{t\in\mathbb{N}} \in X^{\mathbb{N}}$ converging to $\operatorname{argmin}_{x\in X} f(\boldsymbol{x})$[2].

The term black-box means that no information on the function f is available, and although for $\boldsymbol{x} \in X$ we can obtain $f(\boldsymbol{x})$, the calculations behind this are not available. This is a common situation in real-world problems, where $f(\boldsymbol{x})$ may come from a commercial software whose code is unavailable, or may be the result of simulations. We will say that an algorithm is a **black-box**, **zero-order** or **derivative-free** algorithm when it only uses the f-value of $\boldsymbol{x}$. We call an algorithm using the gradient of f (resp. its Hessian) a **first order** algorithm (resp. **second-order** algorithm). We will also say that an algorithm is **function-value free** (FVF) or **comparison-based** if it does not directly use the function value $f(\boldsymbol{x})$, but uses instead how different points are ranked according to their f-values. This notion of FVF is an important property which ensures a certain robustness of an optimization algorithm, and is further developed in 2.4.5.

Section 2.1 will first give some definitions in order to discuss convergence speed in continuous optimization. Then Sections 2.2 and 2.3 will then give a list of well-known deterministic and stochastic optimization algorithms, deterministic and stochastic algorithm requiring different techniques to analyze (the latter requiring the use of probability theory). Section 2.4 will introduce different optimization problems and their characteristics, and Section 2.5 will present results and techniques relating to the convergence of Evolution Strategies.

[1] Maximizing f is equivalent to minimizing $-f$.

[2] Note that in continuous optimization the optimum is usually never found, only approximated.

2.1 Evaluating convergence rates in continuous optimization

In continuous optimization, except for very particular cases, optimization algorithms never exactly find the optimum, contrarily to discrete optimization problems. Instead, at each iteration $t \in \mathbb{N}$ an optimization algorithm produces an estimated solution $\boldsymbol{X}_t$, and the algorithm is considered to solve the problem if the sequence $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ converges to the global optimum $\boldsymbol{x}^*$ of the objective function. To evaluate the convergence speed of the algorithm, one can look at the evolution of the distance between the estimated solution and the optimum, $\|\boldsymbol{X}_t - \boldsymbol{x}^*\|$, or at the average number of iterations required for the algorithm to reach a ball centred on the optimum and of radius $\epsilon \in \mathbb{R}_+^*$. Note that for optimization algorithms (especially in black-box problems) the number of evaluations of f made is an important measure of the computational cost of the algorithm, as the evaluation of the function can be the result of expensive calculations or simulations. And since many algorithms that we consider do multiple function evaluations per iteration, it is therefore often important to consider the converge rate normalized by the number of function evaluations per iteration.

2.1.1 Rates of convergence

Take $(\boldsymbol{x}_t)_{t\in\mathbb{N}}$ a deterministic sequence of real vectors converging to $\boldsymbol{x}^* \in \mathbb{R}^n$. We say that $(\boldsymbol{x}_t)_{t\in\mathbb{N}}$ **converges log-linearly** or **geometrically** to $\boldsymbol{x}^*$ at rate $r \in \mathbb{R}_+^*$ if

$$\lim_{t\to+\infty} \ln \frac{\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}_t - \boldsymbol{x}^*\|} = -r \quad . \tag{2.1}$$

Through Cesàro means[3], this implies that $\lim_{t\to+\infty} \frac{1}{t} \ln(\|\boldsymbol{x}_t - \boldsymbol{x}^*\|) = -r$, meaning that asymptotically, the logarithm of the distance between $\boldsymbol{x}_t$ and the optimum decreases like $-r\,t$.
If (2.1) holds for $r \in \mathbb{R}_-^*$, we say that $(\boldsymbol{x}_t)_{t\in\mathbb{N}}$ **diverges log-linearly** or **geometrically**. If (2.1) holds for $r = +\infty$ then $(\boldsymbol{x}_t)_{t\in\mathbb{N}}$ is said to **converge superlinearly** to $\boldsymbol{x}^*$, and if (2.1) holds for $r = 0$ then $(\boldsymbol{x}_t)_{t\in\mathbb{N}}$ is said to **converge sublinearly**.
In the case of superlinear convergence, for $q \in (1, +\infty)$ we say that $(\boldsymbol{x}_t)_{t\in\mathbb{N}}$ **converges with order** q to $\boldsymbol{x}^*$ at rate $r \in \mathbb{R}_+^*$ if

$$\lim_{t\to+\infty} \ln \frac{\|\boldsymbol{x}_{t+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{x}_t - \boldsymbol{x}^*\|^q} = -r \quad . \tag{2.2}$$

When $q = 2$ we say that the convergence is **quadratic**.
In the case of a sequence of random vectors $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$, t he sequence $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ is said to converge **almost surely** (resp. **in probability**, **in mean**) log-linearly to $\boldsymbol{x}^*$ if the random variable $1/t \ln(\|\boldsymbol{X}_t - \boldsymbol{x}^*\| / \|\boldsymbol{X}_0 - \boldsymbol{x}^*\|)$ converges almost surely (resp. in probability, in mean) to $-r$, with $r \in \mathbb{R}_+^*$. Similarly, we define almost sure divergence and divergence in probability when the random variable $1/t \ln(\|\boldsymbol{X}_t - \boldsymbol{x}^*\| / \|\boldsymbol{X}_0 - \boldsymbol{x}^*\|)$ converges to $r \in \mathbb{R}_+^*$.

[3] The Cesàro means of a sequence $(a_t)_{t\in\mathbb{N}^*}$ are the terms of the sequence $(c_t)_{t\in\mathbb{N}^*}$ where $c_t := 1/t \sum_{i=1}^{t} a_i$. If the sequence $(a_t)_{t\in\mathbb{N}^*}$ converges to a limit l, then so does the sequence $(c_t)_{t\in\mathbb{N}^*}$.

2.1.2 Expected hitting and running time

Take $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ a sequence of random vectors converging to $\boldsymbol{x}^* \in X$. For $\epsilon \in \mathbb{R}_+^*$, the random variable $\tau_\epsilon := \min\{t \in \mathbb{N} | \boldsymbol{X}_t \in B(\boldsymbol{x}^*, \epsilon)\}$ is called the first hitting time of the ball centred in $\boldsymbol{x}$ and of radius ϵ. We define the **expected hitting time** (EHT) as the expected value of the first hitting time. Log-linear convergence at rate r is related to a expected hitting time of $\mathbf{E}(\tau_\epsilon) \sim \ln(1/\epsilon)/r$ when ϵ goes to 0 [67].
Let $\boldsymbol{x}^* \in X$ denote the optimum of a function $f : X \to \mathbb{R}$. We define the running time to a precision $\epsilon \in \mathbb{R}_+^*$ as the random variable $\eta_\epsilon := \min\{t \in \mathbb{N} | |f(\boldsymbol{X}_t) - f(\boldsymbol{x}^*)| \leq \epsilon\}$, and the **expected running time** (ERT) as the expected value of the running time. Although when the objective function is continuous the EHT and ERT are related, it is possible on functions with local optima to have arbitrarily low ERT and high EHT.

2.2 Deterministic Algorithms

In this section we give several classes of deterministic continuous optimization methods. Although this chapter is dedicated to black-box optimization methods, we still present some first and second order methods, as they can be made into zero order methods by estimating the gradients or the Hessian matrices (e.g. through a finite difference method [61]). Furthermore, these methods being widely known and often applied in optimization they are an important comparison point.
We start this section by introducing Newton's method [54] which is a second order algorithm, and Quasi-Newton methods [109, Chapter 6] which are first order algorithms. Then we introduce Trust Region methods [52] which can be derivative-free or first order algorithms. Then we present Pattern Search [115, 136] and Nelder-Mead [108] which are derivative-free methods, with the latter being also function-value free.

2.2.1 Newton's and Quasi-Newton Methods

Inspired from Taylor's expansion, Newton's method [54] is a simple deterministic second order method that can achieve quadratic convergence to a critical point of a C^2 function $f : \mathbb{R}^n \to \mathbb{R}$. Originally, Newton's method is a first order method which converges to a zero of a function $f : \mathbb{R}^n \to \mathbb{R}^n$. To optimize a general C^2 function $f : \mathbb{R}^n \to \mathbb{R}$, Newton's method is instead applied to the function $g : \boldsymbol{x} \in \mathbb{R}^n \mapsto \nabla_{\boldsymbol{x}} f$ to search for points where the gradient is zero, and is therefore used as a second order method. Following this, from an initial point $\boldsymbol{x}_0 \in \mathbb{R}^n$ and $t \in \mathbb{N}$, Newton's method defines $\boldsymbol{x}_{t+1}$ recursively as

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \boldsymbol{H}_f(\boldsymbol{x}_t)^{-1} \nabla_{\boldsymbol{x}_t} f \ , \tag{2.3}$$

where $\boldsymbol{H}_f(\boldsymbol{x})$ is the Hessian matrix of f at $\boldsymbol{x}$. Although the algorithm may converge to saddle points, these can be detected when $\boldsymbol{H}_f(\boldsymbol{x}_t)$ is not positive definite. In order for (2.3) to be well-defined, f needs to be C^2; and if it is C^3 and convex, then quadratic convergence is achieved to the minimum of f [123, Theorem 8.5].

In some cases, computing the gradient or the Hessian of f may be too expensive or not even feasible. They can instead be approximated, which gives a quasi-Newton method. On simple functions quasi-Newton methods are slower than Newton's method but can still, under some conditions, achieve superlinear convergence (see [37]); e.g. sequent method can achieve convergence with order $(1+\sqrt{5})/2$. In general, Eq. (2.3) becomes

$$\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \alpha_t \boldsymbol{p}_t \ , \tag{2.4}$$

where $\boldsymbol{p}_t \in \mathbb{R}^n$ is called the search direction and $\alpha_t \in \mathbb{R}_+^*$ the step-size. The step-size is chosen by doing a line search in the search direction $\boldsymbol{p}_t$, which can be done exactly (e.g. using a conjugate gradient method [110]) or approximately (e.g. using Wolfe conditions [140]). In gradient descent method, the search direction $\boldsymbol{p}_t$ is taken directly as the gradient of f. In BFGS (see [109]), which is the state of the art in quasi-Newton methods, $\boldsymbol{p}_t = \boldsymbol{B}_t^{-1} \nabla_{\boldsymbol{x}_t} f$ where $\boldsymbol{B}_t$ approximates the Hessian of f.
These methods are well-known and often used, and so they constitute an important comparison point for new optimization methods. Also, even when derivatives are not available, if the function to be optimized is smooth enough, approximations of the gradient are good enough for these methods to be effective.

2.2.2 Trust Region Methods

Trust region methods (see [52]) are deterministic methods that approximate the objective function $f : X \subset \mathbb{R}^n \to \mathbb{R}$ by a model function (usually a quadratic function) within an area called the trust region. At each iteration, the trust region is shifted towards the optimum of the current model of f. This shift is limited by the size of the trust region in order to avoid over-estimating the quality of the model and diverging. The size of the trust region is increased when the quality of the model is good, and decreased otherwise. The algorithm may use the gradient of the function to construct the model function [144]. NEWUOA [112] is a derivative-free state-of-the-art trust region method which interpolates a quadratic model using a smaller number of points $m \in [n+2..1/2(n+1)(n+2)]$ (the recommended m-value is $2n+1$) than the $1/2(n+1)(n+2)$ usually used for interpolating quadratic models. The influence of the number of points m used by NEWUOA to interpolate the quadratic model is investigated in [119, 120].

2.2.3 Pattern Search Methods

Pattern search methods (first introduced in [115], [136]) are deterministic function-value free algorithms that improve over a point $\boldsymbol{x}_t \in \mathbb{R}^n$ by selecting a step $\boldsymbol{s}_t \in P_t$, where P_t is subset of $\mathbb{R}^n$ called the pattern, such that $f(\boldsymbol{x}_t + \sigma_t \boldsymbol{s}_t) < f(\boldsymbol{x}_t)$, where $\sigma_t \in \mathbb{R}_+^*$ is called the step-size. If no such point of the pattern exists then $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t$ and the step-size σ_t is decreased by a constant factor, i.e. $\sigma_{t+1} = \theta\sigma_t$ with $\theta \in (0,1)$; otherwise $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \sigma_t \boldsymbol{s}_t$ and the step-size is kept constant. The pattern P_t is defined as the union of the column vectors of a non-singular matrix $\boldsymbol{M}_t$, of its opposite $-\boldsymbol{M}_t$, of the vector $\boldsymbol{0}$ and of an arbitrary number of other vectors of $\mathbb{R}^n$ [136]. Since the matrix $\boldsymbol{M}_t$ has rank n, the vectors of P_t span $\mathbb{R}^n$. The pattern can be

and should be adapted at each iteration: e.g. while a cross pattern (i.e. $\boldsymbol{M}_t = \boldsymbol{Id}_n$) is adapted to a sphere function, it is not for an ellipsoid function with a large condition number (see Section 2.4.2), and even less for a rotated ellipsoid.

2.2.4 Nelder-Mead Method

The Nelder-Mead method introduced in [108] in 1965 is a deterministic function-value free algorithm which evolves a simplex (a polytope with $n+1$ points in a n-dimensional space) to minimize a function $f : X \subset \mathbb{R}^n \to \mathbb{R}$. From a simplex with vertices $(\boldsymbol{x}_i)_{i\in[1..n+1]}$, the algorithm sorts the vertices according to their f-values: $(x_{i:n+1})_{i\in[1..n+1]}$ such that $f(\boldsymbol{x}_{1:n+1}) \leq \ldots \leq f(\boldsymbol{x}_{n+1:n+1})$. Then, denoting $\boldsymbol{x}_c := 1/n\sum_{i=1}^{n} x_{i:n+1}$ the centroid of the n vertices with lowest f-value, it considers three different points on the line between $\boldsymbol{x}_c$ and the vertex with highest f-value $\boldsymbol{x}_{n+1:n+1}$. If none of these points have lower f-value than $\boldsymbol{x}_{n+1:n+1}$, the simplex is reduced by a homothetic transformation with respect to $\boldsymbol{x}_{1:n+1}$ and ratio lower than 1. Otherwise, according to how the f-values of the three points rank with the f-values of the vertices, one of these points replace $\boldsymbol{x}_{n+1:n+1}$ as a vertex of the simplex.
It has been shown that Nelder-Mead algorithm can fail to converge to a stationary point even on strictly convex functions (see [95]). Further discussion about Nelder-Mead algorithm can be found here [142].

2.3 Stochastic algorithms

Stochastic optimization methods use random variables to generate solutions. This make these algorithms naturally equipped to deal with randomness, which can prove useful on difficult functions or in the presence of noise, by for example giving them a chance to escape a local optimum.
In this section we introduce Pure Random Search [146] and Pure Adaptive Search [146], Metropolis-Hastings [41], Simulated Annealing [84], Particle Swarm Optimization [83], Evolutionary Algorithms [26], Genetic Algorithms [73], Differential Evolution [134], Evolution Strategies [117], Natural Evolution Strategies [139] and Information Geometric Optimization [111].

2.3.1 Pure Random Search

Pure Random Search [146] consists in sampling independent random vectors $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ of $\mathbb{R}^n$ from the same distribution P until a stopping criterion is met. The sampling distribution is supposed to be supported by the search space X. The random vector $\boldsymbol{X}_t$ with the lowest f-value is then taken as the solution proposed by the method, i.e. $\boldsymbol{X}_t^{\text{best}} := \text{argmin}_{\boldsymbol{X}\in\{\boldsymbol{X}_k | k\in[0..t]\}} f(\boldsymbol{X})$. While the algorithm is trivial, it is also trivial to show that the sequence $(\boldsymbol{X}_t^{\text{best}})_{t\in\mathbb{N}}$ converges to the global minimum of any continuous function. The algorithm is however very inefficient, converging sublinearly: the expected hitting time for the algorithm to enter a ball of radius $\epsilon \in \mathbb{R}_+^*$ centred around the optimum is proportional to $1/\epsilon^n$. It is therefore a good reminder that convergence in itself is an insufficient criterion to assess the performance of an optimization

algorithm, and any efficient stochastic algorithm using restarts ought to outperform pure random search on most real-world function.

2.3.2 Pure Adaptive Search

Pure Adaptive Search [146] (PAS) is a theoretical algorithm which consists in sampling vectors $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ of $\mathbb{R}^n$ as in PRS, but adding that the support of the distribution from which $\boldsymbol{X}_{t+1}$ is sampled in the strict sub-level set $V_t := \{\boldsymbol{x} \in X | f(\boldsymbol{x}) < f(\boldsymbol{X}_t)\}$. More precisely, denoting P the distribution associated to the PAS that we suppose to be supported by a set $V_0 \subset \mathbb{R}^n$, $\boldsymbol{X}_{t+1} \sim P(\cdot|V_t)$ where $P(\cdot|V_t)$ denotes the probability measure $A \in \mathscr{B}(X) \mapsto P(A \cap V_t)/P(V_t)$. Therefore $(f(\boldsymbol{X}_t))_{t\in\mathbb{N}}$ is a strictly decreasing sequence and the algorithm converges to the minimum of any continuous function. When f is Lipschitz continuous, that the space V_0 is bounded and that P is the uniform distribution on V_0, the running time of PRS with uniform distribution on V_0, η_{PRS}, is exponentially larger than the running time of PAS, η_{PAS}, in the sense that $\eta_{PRS} = \exp(\eta_{PAS} + o(\eta_{PAS}))$ with probability 1 [145, Theorem 3.2].
However, as underlined in [146] simulating the distribution $P(\cdot|\boldsymbol{V}_t)$ in general involves Monte-Carlo sampling or the use of PRS itself, making the algorithm impractical.

2.3.3 Simulated Annealing and Metropolis-Hastings

Here we introduce the Metropolis-Hastings algorithm [41], which uses Monte-Carlo Markov chains to sample random elements from a target probability distribution π supported on R^n, and Simulated Annealing [84] which is an adaptation of the Metropolis-Hastings algorithm as an optimization algorithm.

Metropolis-Hastings

Metropolis-Hastings was first introduced by Metropolis and al. in [97] and extended by Hastings in [71]. Given a function f proportional to the probability density of a distribution π, a point $\boldsymbol{X}_t \in \mathbb{R}^d$ and a conditional symmetric probability density $q(\boldsymbol{x}|\boldsymbol{y})$ (usually taken as a Gaussian distribution with mean $\boldsymbol{y}$ [41]), the Metropolis-Hastings algorithm constructs the random variable $\boldsymbol{X}_{t+1}$ by sampling a candidate $\boldsymbol{Y}_t$ from $q(\cdot|\boldsymbol{X}_t)$, and accepting it as $\boldsymbol{X}_{t+1} = \boldsymbol{Y}_t$ if $f(\boldsymbol{Y}_t) > f(\boldsymbol{X}_t)$, or with probability $f(\boldsymbol{Y}_t)/f(\boldsymbol{X}_t)$ otherwise. If $\boldsymbol{Y}_t$ is rejected, then $\boldsymbol{X}_{t+1} = \boldsymbol{X}_t$. Given $\boldsymbol{X}_0 \in \mathbb{R}^d$, the sequence $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ is a Markov chain, and, given that it is φ-irreducible and aperiodic, it is positive with invariant probability distribution π, and the distribution of $\boldsymbol{X}_t$ converges to π [41].

Simulated Annealing

Simulated Annealing (SA) introduced in [96] in 1953 for discrete optimization problems [84], the algorithm was later extended to continuous problems [34, 53]. SA is an adaptation of the Metropolis-Hastings algorithm which tries to avoid converging to a local and non global minima by having a probability of accepting solutions with higher f-values according to Boltzmann acceptance rule. Denoting $\boldsymbol{X}_t$ the current solution, the algorithm generates a

candidate solution $\boldsymbol{Y}_t$ sampled from a distribution $Q(\cdot|\boldsymbol{X}_t)$. If $f(\boldsymbol{Y}_t) < f(\boldsymbol{X}_t)$ then $\boldsymbol{X}_{t+1} = \boldsymbol{Y}_t$, otherwise $\boldsymbol{X}_{t+1} = \boldsymbol{Y}_t$ with probability $\exp(-(f(\boldsymbol{Y}_t) - f(\boldsymbol{X}_t))/T_t)$ and $\boldsymbol{X}_{t+1} = \boldsymbol{X}_t$ otherwise. The variable T_t is a parameter called the temperature, and decreases to 0 overtime in a process called the cooling procedure, allowing the algorithm to converge.
Although simulated annealing is technically a black-box algorithm, the family of probability distributions $(Q(\cdot|\boldsymbol{x}))_{\boldsymbol{x}\in X}$ and how the temperature changes over time need to be selected according to the optimization problem, making additional information on the objective function f important to the efficiency of the algorithm. Note also that the use of the difference of f-value to compute the probability of taking $\boldsymbol{X}_{t+1} = \boldsymbol{Y}_t$ makes the algorithm not function-value free. SA algorithms can be shown to converge almost surely to the ball of center the optimum of f and radius $\epsilon > 0$, given sufficient conditions on the cooling procedure including that $T_t \geq (1+\mu)N_{f,\epsilon}/\ln(t)$, that the objective function $f : X \to \mathbb{R}$ is continuous, that the distribution $Q(\cdot, \boldsymbol{x})$ is absolutely continuous with respect to the Lebesgue measure for all $\boldsymbol{x} \in X$, and that the search space is compact [91].

2.3.4 Particle Swarm Optimization

Particle Swarm Optimization [83, 49, 132] (PSO) is a FVF optimization algorithm evolving a "swarm", i.e. population of points called particles. It was first introduced by Eberhart and Kennedy in 1995 [83], inspired from the social behaviour of birds or fishes. Take a swarm of particules of size N, and $(\boldsymbol{X}_t^i)_{i\in[1..N]}$ the particles composing the swarm. Each particle $\boldsymbol{X}_t^i$ is attracted towards the best position it has visited, that is $\boldsymbol{p}_t^i := \operatorname{argmin}_{\boldsymbol{x}\in\{\boldsymbol{X}_k^i | k\in[0..t]\}} f(\boldsymbol{x})$, and towards the best position the swarm has visited, that is $\boldsymbol{g}_t := \operatorname{argmin}_{\boldsymbol{x}\in\{\boldsymbol{p}_t^i | i\in[1..N]\}} f(\boldsymbol{x})$, while keeping some of its momentum. More precisely, for $\boldsymbol{V}_t^i$ the velocity of the particle $\boldsymbol{X}_t^i$,

$$\boldsymbol{V}_{t+1}^i = \omega \boldsymbol{V}_t^i + \psi_p \boldsymbol{R}_p \circ (\boldsymbol{p}_t^i - \boldsymbol{X}_t^i) + \psi_g \boldsymbol{R}_g \circ (\boldsymbol{g}_t - \boldsymbol{X}_t^i) \ , \tag{2.5}$$

where ω, ψ_p and ψ_g are real parameters of the algorithm, $\circ$ denote the Hadamard product and $\boldsymbol{R}_p$ and $\boldsymbol{R}_g$ are two independent random vectors, whose coordinates in the canonical basis are independent random variables uniformly distributed in $[0,1]$. Then $\boldsymbol{X}_t^i$ is updated as

$$\boldsymbol{X}_{t+1}^i = \boldsymbol{X}_t^i + \boldsymbol{V}_t^i \ . \tag{2.6}$$

Note that the distribution of $\boldsymbol{R}_p$ and $\boldsymbol{R}_g$ is not rotational invariant, and causes PSO to exploit separability. Although PSO behaves well to ill-conditioning on separable functions, its performances have been shown to be greatly affected when the problem is non-separable (see [70]). Variants of PSO have been developed to avoid these shortcomings [35].

2.3.5 Evolutionary Algorithms

Evolutionary Algorithms [26, 143] (EAs) consist of a wide class of derivative-free optimization algorithms inspired from Darwin's theory of evolution. A set of points, called the population, is evolved using the following scheme: from a population P of $\mu \in \mathbb{N}^*$ points called the parents, a population O of $\lambda \in \mathbb{N}^*$ new points called offsprings is created, and then μ points among

O or $O \cup P$ are selected to create the new parents. To create an offspring in O, an EA can use two or more points from the parent population in a process called recombination, or apply a variation to a parent point due to a random element, which is called mutation. The selection procedure can operate on $O \cup P$ in which case it is called elitist, or on O is which case it is called non-elitist. The selection can choose the best μ points according to their rankings in f-value, or it can use the f-value of a point to compute the chance that this point has to be selected into the new population.

2.3.6 Genetic Algorihms

Genetic Algorithms [104, 60] (GAs) are EAs using mutation and particular recombination operators called crossovers. GAs have first been introduced in [73], where the search space was supposed to be the space of bit strings of a given length $n \in \mathbb{N}^*$ (i.e. $X = \{0,1\}^n$). They have been widely used and represent an important community in discrete optimization. Adaptations of GAs to continuous domains have been proposed in [101, 40]. Taking two points $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$ from the parent population, a crossover operator creates a new points by combining the coordinates of $\boldsymbol{X}_t$ and $\boldsymbol{Y}_t$. To justify the importance of crossovers, GAs rely on the so-called building-block hypothesis, which assumes that the problem can be cut into several lower-order problems that are easier to solve, and that an individual having evolved the structure for one of these low order problem will transmit it to the rest of the population through crossovers. The usefulness of crossovers has long been debated, and it has been suggested that crossovers can be replaced with a mutation operator with large variance. In fact, in [81] it was shown that for some GAs in discrete search spaces, the classic crossover operator is inferior to the headless chicken operator, which consists in doing a crossover of a point with an independently randomly generated point, which can be seen as a mutation. However, it has been proven in [56] that for some discrete problems (here a shortest path problem in graphs), EAs using crossovers can solve these problems better than EAs using pure mutation.

2.3.7 Differential Evolution

Differential Evolution (DE) is a function value free EA introduced by Storn and Price [134]. For each point $\boldsymbol{X}_t$ of its population, it generates a new sample by doing a crossover between this point and the point $\boldsymbol{A}_t + F(\boldsymbol{B}_t - \boldsymbol{C}_t)$, where $\boldsymbol{A}_t$, $\boldsymbol{B}_t$, and $\boldsymbol{C}_t$ are other distinct points randomly taken the population, and $F \in [0,2]$ is called the differentiation weight. If the new sample $\boldsymbol{Y}_t$ has a better fitness than $\boldsymbol{X}_t$, then it replaces $\boldsymbol{X}_t$ in the new population (i.e. $\boldsymbol{X}_{t+1} = \boldsymbol{Y}_t$). The performances of the algorithm highly depends on how the recombination is done and of the value of F [57]. When there is no crossover (i.e. the new sample $\boldsymbol{Y}_t$ is $\boldsymbol{A}_t + F(\boldsymbol{B}_t - \boldsymbol{C}_t)$), the algorithm is rotational-invariant, but otherwise it is not [114, p. 98]. DE is prone to premature convergence and stagnation [87].

2.3.8 Evolution Strategies

Evolution Strategies (ESs) are function value free EAs using mutation, first introduced by Rechenberg and Schwefel in the mid 1960s for continuous optimization [116, 126, 117]. Since ESs are the focus of this work, a more thorough introduction will be given. From a distribution P_{θ_t} valued in $\mathbb{R}^n$, an ES samples $\lambda \in \mathbb{N}^*$ points $(\boldsymbol{Y}_t^i)_{i\in[1..\lambda]}$, and uses the information on the rankings in f-value of the samples to update the distribution P_{θ_t} and other internal parameters of the algorithm. In most cases, the family of distribution $(P_\theta)_{\theta\in\Theta}$ are multivariate normal distributions. A multivariate normal distribution, that we denote $\mathcal{N}(\boldsymbol{X}_t, \boldsymbol{C}_t)$, is parametrized by a mean $\boldsymbol{X}_t$ and a covariance matrix $\boldsymbol{C}_t$; we also add a scaling parameter σ_t called the step size, such that $(\boldsymbol{Y}_t^i)_{i\in[1..\lambda]}$ are sampled from $\sigma_t\mathcal{N}(\boldsymbol{X}_t, \boldsymbol{C}_t)$. Equivalently,

$$\boldsymbol{Y}_t^i = \boldsymbol{X}_t + \sigma_t \boldsymbol{C}_t^{1/2} \boldsymbol{N}_t^i \ , \tag{2.7}$$

where $(\boldsymbol{N}_t^i)_{i\in[1..\lambda]}$ is a sequence of i.i.d. standard multivariate normal vectors that we call **random steps**. The choice of multivariate normal distributions fits exactly to the context of black-box optimization, as multivariate normal distributions are maximum entropy probability distributions, meaning as little assumption as possible on the function f is being made. However, when the problem is not entirely black-box and some information of f is available, other distributions may be considered: e.g. separability can be exploited by distributions having more weight on the axes, such as multivariate Cauchy distributions [64].

The different samples $(\boldsymbol{Y}_t^i)_{i\in[1..\lambda]}$ are ranked according to their f-value. We denote $\boldsymbol{Y}_t^{i:\lambda}$ the sample with the i^{th} lowest f-value among the $(\boldsymbol{Y}_t^i)_{i\in[1..\lambda]}$. This also indirectly defines an ordering on the random steps, and we denote $\boldsymbol{N}_t^{i:\lambda}$ the random step among $(\boldsymbol{N}_t^j)_{j\in[1..\lambda]}$ corresponding to $\boldsymbol{Y}_t^{i:\lambda}$. The ranked samples $(\boldsymbol{Y}_t^{i:\lambda})_{i\in[1..\lambda]}$ are used to update $\boldsymbol{X}_t$, the mean of the sampling distribution, with one of the following strategy [67]:

$$(1,\lambda)\text{-ES:} \qquad \boldsymbol{X}_{t+1} = \boldsymbol{Y}_t^{1:\lambda} = \boldsymbol{X}_t + \sigma_t \boldsymbol{C}_t^{1/2} \boldsymbol{N}_t^{1:\lambda} \ . \tag{2.8}$$

The $(1,\lambda)$-ES is called a **non-elitist** ES.

$$(1+\lambda)\text{-ES:} \qquad \boldsymbol{X}_{t+1} = \boldsymbol{X}_t + \mathbf{1}_{f(\boldsymbol{Y}_t^{1:\lambda})\leq f(\boldsymbol{X}_t)} \sigma_t \boldsymbol{C}_t^{1/2} \boldsymbol{N}_t^{1:\lambda} \ . \tag{2.9}$$

The $(1+\lambda)$-ES is called an **elitist** ES.

$$(\mu/\mu_W,\lambda)\text{-ES:} \qquad \boldsymbol{X}_{t+1} = \boldsymbol{X}_t + \kappa_m \sum_{i=1}^{\mu} w_i (\boldsymbol{Y}_t^{i:\lambda} - \boldsymbol{X}_t) = \boldsymbol{X}_t + \kappa_m \sigma_t \boldsymbol{C}_t^{1/2} \sum_{i=1}^{\mu} w_i \boldsymbol{N}_t^{i:\lambda} \ , \tag{2.10}$$

where $\mu \in [1..\lambda]$, $(w_i)_{i\in[1..\mu]} \in \mathbb{R}^\mu$ are weights such that $\sum_{i=1}^{\mu} w_i = 1$. The parameter $\kappa_m \in \mathbb{R}_+^*$ is called a learning rate, and is usually set to 1. The $(\mu/\mu_W,\lambda)$-ES is said to be with **weighted recombination**. If for all $i \in [1..\mu]$, $w_i = 1/\mu$, the ES is denoted $(\mu/\mu,\lambda)$-ES.

Adaptation of the step-size

For an ES to be efficient, the step-size σ_t has to be adapted. Some theoretical studies [33, 78, 79] consider an ES where the step-size is kept proportional to the distance to the optimum, which is a theoretical ES which can achieve optimal convergence rate on the sphere function [19, Theorem 2] (shown in the case of the isotropic $(1,\lambda)$-ES). Different techniques to adapt the step-size exist; we will present σ-Self-Adaptation [129] (σSA) and Cumulative Step-size Adaptation [66] (CSA), the latter being used in the state-of-the-art algorithm CMA-ES [66].

Self-Adaptation The mechanism of σSA to adapt the covariance matrix of the sampling distribution was first introduced by Schwefel in [127]. In σSA, the sampling of the new points $\boldsymbol{Y}_t^i$ is slightly different from Eq. (2.7). Each new sample $\boldsymbol{Y}_t^i$ is coupled with a step-size $\sigma_t^i := \sigma_t \exp(\tau \xi_t^i)$, where $\tau \in \mathbb{R}_+^*$ and $(\xi_t^i)_{t\in\mathbb{N}, i\in[1..\lambda]}$ is a sequence of i.i.d. random variables, usually standard normal variables [130]. The samples $\boldsymbol{Y}_t^i$ are then defined as

$$\boldsymbol{Y}_t^i := \boldsymbol{X}_t + \sigma_t^i \boldsymbol{C}_t^{1/2} \boldsymbol{N}_t^i \ , \tag{2.11}$$

where $(\boldsymbol{N}_t^i)_{i\in[1..\lambda]}$ is a i.i.d. sequence of random vectors with multivariate standard normal distribution. Then $\sigma_t^{i:\lambda}$ is defined as the step-size associated to the sample with the i^{th} lowest value, $\boldsymbol{Y}_t^{i:\lambda}$. The step-size is then adapted as $\sigma_{t+1} = \sigma_t^{1:\lambda}$ for a $(1,\lambda)$-ES, or $\sigma_{t+1} = 1/\mu \sum_{i=1}^{\mu} \sigma_t^i$ in the case of weighted recombination with weights $w_i = 1$ for all $i \in [1..\mu]$. Note that using an arithmetic mean to recombine the step-sizes (which are naturally geometric) creates a bias towards larger step-size values.
The indirect selection for the step-size raises some problems, as raised in [63]: on a linear function, since $\boldsymbol{N}_t^i$ and $-\boldsymbol{N}_t^i$ are as likely to be sampled, the i^{th} best sample $\boldsymbol{Y}_t^{i:\lambda}$ and the i^{th} worst sample $\boldsymbol{Y}_t^{\lambda-i:\lambda}$ are as likely to be generated by the same step-size, and therefore there is no correlation between the step-size and the ranking. In [68] σSA is analysed and compared with other step-size adaptation mechanisms on the linear, sphere, ellipsoid, random fitness and stationary sphere functions.

Cumulative Step-size Adaptation In Cumulative Step-size Adaptation (CSA), which is detailed in [66], for a $(\mu/\mu_W,\lambda)$-ES the difference between the means of the sampling distribution at iteration t and $t+1$ is renormalized as $\Delta_t := \sqrt{\mu_w}\boldsymbol{C}_t^{-1/2}(\boldsymbol{X}_{t+1}-\boldsymbol{X}_t)/\sigma_t$ where $\mu_w = 1/\sum_{i=1}^{\mu} w_i^2$ and $(w_i)_{i\in[1..\mu]}$ are the weights defined in page 21. If the objective function ranks the samples uniformly randomly, this renormalization makes Δ_t distributed as a standard normal multivariate vector. The variable Δ_t is then added to a variable $\boldsymbol{p}_{t+1}^{\sigma}$ called an evolution path following

$$\boldsymbol{p}_{t+1}^{\sigma} = (1-c_\sigma)\boldsymbol{p}_t^{\sigma} + \sqrt{c_\sigma(2-c_\sigma)}\sqrt{\mu_w}\boldsymbol{C}_t^{-1/2}\frac{\boldsymbol{X}_{t+1}-\boldsymbol{X}_t}{\sigma_t} \ . \tag{2.12}$$

The coefficients in (2.12) are chosen such that if $\boldsymbol{p}_t^{\sigma} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Id}_n)$ and if f ranks the samples uniformly randomly, then $\Delta_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Id}_n)$ and $\boldsymbol{p}_{t+1}^{\sigma} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{Id}_n)$. The variable $c_\sigma \in (0,1]$ is called the cumulation parameter, and determines the "memory" of the evolution path, with

the importance of a step Δ_0 decreasing in $(1-c_\sigma)^t$. The "memory" of the evolution path is about $1/c_\sigma$.
The step-size is then adapted depending on the length of the evolution path. If the evolution path is longer (resp. shorter) than the expected length of a standard normal multivariate vector, the step-size is increased (resp. decreased) as follow:

$$\sigma_{t+1} = \sigma_t \exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|\boldsymbol{p}^\sigma_{t+1}\|}{\mathbf{E}(\|\mathcal{N}(\mathbf{0}, \boldsymbol{Id}_n)\|)} - 1\right)\right) \ . \tag{2.13}$$

The variable d_σ determines the variations of the step-size. Usually d_σ is taken as 1.

Adaptation of the covariance matrix

To be able to solve ill-conditioned or not separable functions, evolution strategies need to adapt the covariance matrix $\boldsymbol{C}_t$, which can be done with the state-of-the-art algorithm Covariance Matrix Adaptation (CMA) [66]. CMA adapts the step-size by using CSA, and uses another evolution path $\boldsymbol{p}_t$ to adapt the covariance matrix:

$$\boldsymbol{p}_{t+1} = (1-c)\boldsymbol{p}_t + \sqrt{\mu_w c(2-c)}\frac{\boldsymbol{X}_{t+1} - \boldsymbol{X}_t}{\sigma_t} \ , \tag{2.14}$$

where $c \in (0,1]$. The evolution path $\boldsymbol{p}_t$ is similar to $\boldsymbol{p}^\sigma_t$ with added information on the covariance matrix.
The covariance matrix is then updated as follow:

$$\boldsymbol{C}_{t+1} = (1-c_1-c_\mu)\boldsymbol{C}_t + \underbrace{c_1 \boldsymbol{p}_t \boldsymbol{p}_t^T}_{\text{rank-1 update}} + \underbrace{c_\mu \sum_{i=1}^{\mu} w_i \frac{(\boldsymbol{Y}_t^{i:\lambda} - \boldsymbol{X}_t)(\boldsymbol{Y}_t^{i:\lambda} - \boldsymbol{X}_t)^T}{\sigma_t^2}}_{\text{rank-}\mu\text{ update}} \ , \tag{2.15}$$

where $(c_1, c_\mu) \in (0,1]^2$ and $c_1 + c_\mu \leq 1$. The update associated to c_1 is called the rank-one update, and bias the sampling distribution in the direction of $\boldsymbol{p}_t$. The other is called the rank-μ update, and bias the sampling distribution in the direction of the best sampled points of this iteration.

2.3.9 Natural Evolution Strategies and Information Geometry Optimization

ESs can be viewed as stochastic algorithms evolving a population of points defined on the search space X. In order to optimize a function f, the population needs to converge to the optimum of f. And in order for this process to be efficient, the sampling distribution used to evolve the population needs to be adapted as well throughout the optimization.
A new paradigm is proposed with Estimation of Distribution Algorithms [88]: an ES can be said to evolve a probability distribution among a family of distribution $(P_\theta)_{\theta\in\Theta}$ parametrized by $\theta \in \Theta$. The current probability distribution P_{θ_t} represents the current estimation of where optimal values of f lies. Hence to optimize a function f, the mass of the probability distribution is expected to concentrate around the optimum.
In this perspective, theoretically well-founded optimization algorithms can be defined[139,

111] through stochastic gradient ascent or descent on the Riemannian manifold $(P_\theta)_{\theta\in\Theta}$ by using a natural gradient [4] which is adapted to the Riemannian metric structure of the manifold $(P_\theta)_{\theta\in\Theta}$. Also, interestingly, as shown in [3, 59] the $(\mu/\mu_W,\lambda)$-CMA-ES defined in 2.3.8 using rank-μ update (i.e. setting $c_\sigma = 0$, $\sigma_0 = 1$ and $c_1 = 0$) can be connected to a natural gradient ascent on the Riemannian manifold $(P_\theta)_{\theta\in\Theta}$.

Natural Evolution Strategies

Given a family of probability distributions, Natural Evolution Strategies [139, 138, 59] (NESs) indirectly minimize a function $f : \mathbb{R}^n \to \mathbb{R}$ by minimizing the criterion

$$\tilde{J}(\theta) := \int_{\mathbb{R}^n} f(\boldsymbol{x}) P_\theta(d\boldsymbol{x}) \ . \tag{2.16}$$

Minimizing this criterion involves concentrating the distribution P_θ around the global minima of f. To minimize $\tilde{J}(\theta)$, a straightforward gradient descent

$$\theta_{t+1} = \theta_t - \eta \nabla_{\theta_t} \tilde{J}(\theta_t) \tag{2.17}$$

could be considered, where $\eta \in \mathbb{R}_+^*$ is a learning rate. Using the so called log-likelihood trick, it can be shown that

$$\nabla_\theta \tilde{J}(\theta) = \int_{\mathbb{R}^n} f(\boldsymbol{x}) \nabla_\theta \ln\left(P_\theta(\boldsymbol{x})\right) P_\theta(d\boldsymbol{x}) \ , \tag{2.18}$$

which can be used to estimate $\nabla_\theta \tilde{J}(\theta)$ as $\nabla_\theta^{\text{est}} \tilde{J}(\theta)$ via

$$\nabla_\theta^{\text{est}} \tilde{J}(\theta) = \frac{1}{\lambda} \sum_{i=1}^{\lambda} f\left(\boldsymbol{Y}^i\right) \nabla_\theta \ln\left(P_\theta\left(\boldsymbol{Y}^i\right)\right) \ , \quad \text{where } \left(\boldsymbol{Y}^i\right)_{i\in[1..\lambda]} \text{ i.i.d. and } \boldsymbol{Y}^i \sim P_\theta \ . \tag{2.19}$$

However, as the authors of [138] stress out, the algorithm defined through (2.17) is not invariant to a change of parametrization of the distribution. To correct this, NESs use the natural gradient proposed in [4] which is invariant to changes of parametrization of the distribution. The direction of the natural gradient $\tilde{\nabla}_\theta \tilde{J}(\theta)$ can be computed using the Fisher information matrix $\boldsymbol{F}(\theta)$ via

$$\tilde{\nabla}_\theta \tilde{J}(\theta) := \boldsymbol{F}(\theta)^{-1} \nabla_\theta \tilde{J}(\theta) \ , \tag{2.20}$$

where the Fisher information matrix is defined as

$$\boldsymbol{F}(\theta) := \int_{\mathbb{R}^n} \nabla_\theta \ln\left(P_\theta(\boldsymbol{x})\right) \nabla_\theta \ln\left(P_\theta(\boldsymbol{x})\right)^T P_\theta(d\boldsymbol{x}) \ . \tag{2.21}$$

Combining (2.20), (2.19) gives the formulation of NESs which update the distribution parameter θ_t through a stochastic natural gradient descent

$$\theta_{t+1} = \theta_t - \eta \boldsymbol{F}(\theta_t)^{-1} \nabla_{\theta_t}^{\text{est}} \tilde{J}(\theta_t) \ . \tag{2.22}$$

Note that the Fisher information matrix can be approximated as done in [139]. However, in [3, 59] expressions of the Fisher information matrix for multivariate Gaussian distribution are given.

The criterion $\tilde{J}(\theta)$ is not invariant to the composition of f by strictly increasing transformations (see 2.4.5), and therefore the algorithm defined in (2.22) is not either. In [138] following [111], in order for the NES to be invariant under the composition of f by strictly increasing transformations, the gradient $\nabla_\theta \tilde{J}(\theta)$ is estimated through the rankings of the different samples $(\boldsymbol{Y}^i)_{i\in[1..\lambda]}$ instead of through their f-value, i.e.

$$\nabla_\theta^{\text{est},2}\tilde{J}(\theta) = \frac{1}{\lambda}\sum_{i=1}^{\lambda} w_i \nabla_\theta \ln\left(P_\theta\left(\boldsymbol{Y}^{i:\lambda}\right)\right) \ , \tag{2.23}$$

where $(\boldsymbol{Y}^i)_{i\in[1..\lambda]}$ is a i.i.d. sequence of random elements with distribution P_θ and $\boldsymbol{Y}^{i:\lambda}$ denotes the element of the sequence $(\boldsymbol{Y}^i)_{i\in[1..\lambda]}$ with the i^{th} lowest f-value, and $(w_i)_{i\in[1..\lambda]} \in \mathbb{R}^\lambda$ is a decreasing sequence of weight such that $\sum_{i=1}^{\lambda} |w_i| = 1$. The approximated gradient $\nabla_\theta^{\text{est},2}\tilde{J}(\theta)$ can be used in (2.22) instead of $\nabla_\theta^{\text{est}}\tilde{J}(\theta)$ to make NES invariant with respect to the composition of f by strictly increasing transformations.

When the probability distribution family $(P_\theta)_{\theta\in\Theta}$ is the multivariate Gaussian distributions, an NES with exponential parametrization of the covariance matrix results results in eXponential NES [59] (xNES).

Information Geometry Optimization

Information Geometry Optimization [111] (IGO) offers another way to turn a family of probabilities $(P_\theta)_{\theta\in\Theta}$ into an optimization algorithm. Instead of using $\tilde{J}(\theta)$ of (2.16) as in NES, IGO considers a criterion invariant to the composition of f by strictly increasing transformations

$$J_{\theta_t}(\theta) := \int_{\mathbb{R}^n} W_{\theta_t}^f(\boldsymbol{x}) P_\theta(d\boldsymbol{x}) \ , \tag{2.24}$$

where $W_{\theta_t}^f$, the weighted quantile function, is a transformation of f using P_{θ_t}-quantiles $q_{\theta_t}^{\leq}$ and $q_{\theta_t}^{<}$ defined as

$$q_{\theta_t}^{\leq}(\boldsymbol{x}) := \Pr\left(f(\boldsymbol{Y}) \leq f(\boldsymbol{x}) | \boldsymbol{Y} \sim P_\theta\right) \tag{2.25}$$

$$q_{\theta_t}^{<}(\boldsymbol{x}) := \Pr\left(f(\boldsymbol{Y}) < f(\boldsymbol{x}) | \boldsymbol{Y} \sim P_\theta\right) \tag{2.26}$$

and which define $W_{\theta_t}^f$ as

$$W_{\theta_t}^f(\boldsymbol{x}) := \begin{cases} w\left(q_{\theta_t}^{\leq}(\boldsymbol{x})\right) & \text{if } q_{\theta_t}^{\leq}(\boldsymbol{x}) = q_{\theta_t}^{<}(\boldsymbol{x}) \\ \frac{1}{q_{\theta_t}^{\leq}(\boldsymbol{x}) - q_{\theta_t}^{<}(\boldsymbol{x})} \int_{q_{\theta_t}^{<}(\boldsymbol{x})}^{q_{\theta_t}^{\leq}(\boldsymbol{x})} w(q)dq & \text{otherwise,} \end{cases} \tag{2.27}$$

where the function $w : [0,1] \to \mathbb{R}$ is any non-increasing function. Note that small f-values correspond to high values of $W_{\theta_t}^f$. Hence minimizing f translates into maximizing $J_{\theta_t}(\theta)$ over

Θ.
In order to estimate $W_{\theta_t}^f$, λ points $(\boldsymbol{Y}^i)_{i\in[1..\lambda]}$ are sampled independently from P_{θ_t} and ranked according to their f-value. We define their rankings through the function $\mathrm{rk}^< : \boldsymbol{y} \in \{\boldsymbol{Y}^i | i \in [1..\lambda]\} \mapsto \#\{j \in [1..\lambda] | f(\boldsymbol{Y}^j) < f(\boldsymbol{y})\}$, and then we define $\hat{w}_i$ as

$$\hat{w}_i\left(\left(\boldsymbol{Y}^j\right)_{j\in[1..\lambda]}\right) := \frac{1}{\lambda} w\left(\frac{\mathrm{rk}^<(\boldsymbol{Y}^i) + \frac{1}{2}}{\lambda}\right) \tag{2.28}$$

where $w : [0,1] \to \mathbb{R}$ is the same function as in (2.27). The IGO algorithm with parametrization θ, sample size $\lambda \in \mathbb{N}^*$ and step-size $\delta t \in \mathbb{R}_+^*$ is then defined as a stochastic natural gradient ascent via the update

$$\theta_{t+\delta t} = \theta_t + \delta t \boldsymbol{F}(\theta_t)^{-1} \frac{1}{\lambda} \sum_{i=1}^{\lambda} \hat{w}_i\left(\left(\boldsymbol{Y}^j\right)_{j\in[1..\lambda]}\right) \nabla_\theta \ln\left(P_\theta\left(\boldsymbol{Y}^i\right)\right)\Big|_{\theta=\theta_t} , \tag{2.29}$$

where $\boldsymbol{F}(\theta_t)$ is the Fisher information matrix defined in (2.21), and $(\boldsymbol{Y}^j)_{j\in[1..\lambda]}$ are i.i.d. random elements with distribution P_{θ_t}. Note that the estimate of $W_{\theta_t}^f$, $\hat{w}_i$, is also invariant to the composition of f by strictly increasing transformations, which makes IGO invariant to the composition of f by strictly increasing transformations. Note that as shown in [111, Theorem 6], $1/\lambda \sum_{i=1}^{\lambda} \hat{w}_i((\boldsymbol{Y}^j)_{j\in[1..\lambda]}) \nabla_\theta \ln(P_\theta(\boldsymbol{Y}^i))\big|_{\theta=\theta_t}$ is a consistent estimator of $\nabla_\theta J_{\theta_t}(\theta)\big|_{\theta=\theta_t}$.
IGO offers a large framework for optimization algorithms. As shown in [111, Proposition 20], IGO for multivariate Gaussian distributions corresponds to the $(\mu/\mu_W, \lambda)$-CMA-ES with rank-μ update (i.e. $c_1 = 0$, $c_\sigma = 1$). IGO can also be used in discrete problems, and as shown in [111, Proposition 19], for Bernoulli distributions IGO corresponds to the Population-Based Incremental Learning [27].

2.4 Problems in Continuous Optimization

Optimization problems can be characterized by several features that can greatly impact the behaviour of optimization algorithms on such problems, thus proving to be potential sources of difficulty. We first identify some of these features, then discuss functions that are important representatives of these features or that relate to optimization problems in general. Some algorithms can be insensitive to specific types of difficulty, which we will discuss through the invariance of these algorithms to a class of functions.

2.4.1 Features of problems in continuous optimization

Following [22], we give here a list of important features impacting the difficulty of optimization problems. For some of the difficulties, we also give examples of algorithms impacted by the difficulty, and techniques or algorithms that alleviate the difficulty.
A well-known, albeit ill-defined, source of difficulty is **ruggedness**. We call a function rugged when its graph is rugged, and the more complex or rugged this graph is, the more information is needed to correctly infer the shape of the function, and so the more expensive it gets to

optimize the function. This ruggedness may stem from the presence of many local optima (which is called **multi-modality**), the presence of **noise** (meaning that the evaluation of a point $\boldsymbol{x} \in X$ by f is perturbated by a random variable, so two evaluations of the same point may give two different f-values), or the function being **not differentiable** or even **not continuous**. Noise is a great source of difficulty, and appears in many real-world problems. We develop it further in Section 2.4.4. The non-differentiability or continuity of the function is obviously a problem for algorithms relying on such properties, such as first order algorithms like gradient based methods. When the gradient is unavailable, these algorithms may try to estimate it (e.g. through a finite difference method [89]), but these methods are sensitive to noise or discontinuities. In contrast, as developed in Section 2.4.5, function-free value algorithms are in a certain measure resilient to discontinuities. Multi-modality is also a great source of difficulty. A multi-modal function can trap an optimization algorithm in a local minimum, which then needs to detect it to get outside of the local minimum. This is usually done simply by restarting the algorithm at a random location (see [107] and [94, Chapter 12] for more on restarts). To try to avoid falling in a local optima, an algorithm can increase the amount of information it acquires at each iteration (e.g. increase of population in population-based algorithms). How large should the increment be is problem dependent, so some algorithms adapt this online over each restart (e.g. IPOP-CMA-ES [23]).

The **dimension** of the search space X is a well known source of difficulty. The "curse of dimensionality" refers to the fact that volumes grow exponentially with the dimension, and so the amount of points needed to achieve a given density in a volume also grows exponentially. Also, algorithms that update full $n \times n$ matrices, such as BFGS (see 2.2.1) or CMA-ES (see 2.3.8) typically perform operations such as matrices multiplication or inversion that scale at least quadratically with the dimension. So in very high dimension (which is called large-scale) the time needed to evaluate the objective function can become negligible compared to the time for internal operations of these algorithms, such as matrices multiplication, inversion or eigen values decomposition. In a large-scale context, these algorithms therefore use sparse matrices to alleviate this problem (see [90] for BFGS, or [92] for CMA-ES).

Ill-conditioning is another common difficulty. For a function whose level sets are close to an ellipsoid, the conditioning can be defined as the ratio between the largest and the smallest axis of the ellipsoid. A function is said ill-conditioned when the conditioning is large (typically larger than 10^5). An isotropic ES (i.e. whose sampling distribution has covariance matrix $\boldsymbol{Id}_n$, see Section 2.3.8) will be greatly slowed down. Algorithms must be able to gradually learn the local conditioning of the function through second order models approximating the Hessian or its inverse (as in BFGS or CMA-ES).

A less known source of difficulty is **non-separability**. A function f with global optimum $\boldsymbol{x}^* = (x_1^*, \ldots, x_n^*) \in \mathbb{R}^n$ is said separable if for any $i \in [1..n]$ and any $(a_j)_{j \in [1..n]} \in \mathbb{R}^n$, $x_i^* = \operatorname{argmin}_{x \in \mathbb{R}} f(a_1, \ldots, a_{i-1}, x, a_{i+1}, \ldots, a_n)$. This implies that the problem can be solved by solving n one-dimensional problems, and that the coordinate system is well adapted to the problem. Many algorithms assume the separability of the function (e.g. by manipulating vectors coordinate-wise), and their performances can hence be greatly affected when the function is not separable.

Constraints are another source of difficulty, especially as many optimization algorithms are tailored with unconstrained optimization in mind. While any restriction of the search space from $\mathbb{R}^n$ to one of its subset is a constraint, constraints are usually described through two sequences of functions $(g_i)_{i\in[1..r]}$ and $(h_i)_{i\in[1..r]}$, the inequality constraints and equality constraints. The constrained optimization problem then reads

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$$

$$\text{subject to } g_i(\boldsymbol{x}) \geq 0 \text{ for } i \in [1..r] \text{ and}$$
$$h_i(\boldsymbol{x}) = 0 \text{ for } i \in [1..s] \quad .$$

Constraints are an important problem in optimization, and many methods have been developed to deal with them [100, 51, 109]. This subject is developed further in this section.

2.4.2 Model functions

In order to gain insight in an optimization algorithm, it is often useful to study its behaviour on different test functions which represent different situations and difficulties an algorithm may face in real-world problems. Important classes of test functions include

- **Linear functions**: If the algorithm admits a step-size σ_t, linear functions model when the step-size is small compared to the distance to the optimum. The level sets of the objective function may then locally be approximated by hyperplanes, which corresponds to the level sets of a linear function. Since a linear function has no optimum, we say that an optimization algorithm solves this function if the sequence $(f(\boldsymbol{X}_t))_{t\in\mathbb{N}}$ diverges to $+\infty$, where $\boldsymbol{X}_t$ is the solution recommended by the algorithm at step t. Linear functions need to be solved efficiently for an algorithm using a step-size to be robust with regards to the initialization.

- **Sphere function**: The sphere function is named after the shape of its level sets and is usually defined as

  $$f_{sphere} : \boldsymbol{x} \in \mathbb{R}^n \mapsto \|\boldsymbol{x}\|^2 = \sum_{i=1}^{n} [\boldsymbol{x}]_i^2 \quad .$$

  The sphere function model an optimal situation where the algorithm is close to an optimum of a convex, separable and well conditioned problem. Studying an algorithm on the sphere function tells how fast we can expect an algorithm to converge in the best case. The isotropy and regularity properties of the sphere function also make theoretical analysis of optimization algorithms easier, and so they have been the subject of many studies [33, 18, 78, 79].

- **Ellipsoid functions**: Ellipsoid functions are functions of the form

  $$f_{ellipsoid} : \boldsymbol{x} \in \mathbb{R}^n \mapsto \boldsymbol{x}^T \boldsymbol{O}^T \boldsymbol{D}\boldsymbol{O}\boldsymbol{x} \quad ,$$

where $\boldsymbol{D}$ is a diagonal matrix and O is an orthogonal matrix, and so the level sets are ellipsoids. Denoting a_i the eigenvalues of $\boldsymbol{D}$, the number $\max_{i\in[1..n]} a_i / \min_{i\in[1..n]} a_i$ is the condition number. When $\boldsymbol{O} = \boldsymbol{Id}_n$ and with a large condition number, ellipsoid functions are ill-conditioned separable sphere functions, making them interesting functions to study the impact of ill-conditioning on the convergence of an algorithm. When the matrix $\boldsymbol{O}^T\boldsymbol{D}\boldsymbol{O}$ is non diagonal and has a high condition number, the ill-conditioning combined with the rotation makes the function non-separable. Using ellipsoids with both $O^T DO$ diagonal or non-diagonal and high condition number can therefore give a measure of the impact of non-separability on an algorithm.

- **Multimodal functions**: Multimodal functions are very diverse in shape. Multimodal functions may display a general structure leading to the global optimum, such as the Rastrigin function [106]

$$f_{\text{rastrigin}} := 10n + \sum_{i=1}^{n} [\boldsymbol{x}]_i^2 + 10\cos(2\pi[\boldsymbol{x}]_i) \ .$$

The global structure of $f_{\text{rastrigin}}$ is given by $\sum_{i=1}^{n}[\boldsymbol{x}]_i^2$, while many local optima are created by $10\cos(2\pi[\boldsymbol{x}]_i)$. In some functions, such as the bi-Rastrigin Lunacek function [55]

$$f_{\text{lunacek}} := \min\left\{\sum_{i=1}^{n}([\boldsymbol{x}]_i - \mu_1)^2,\ \ dn + s\sum_{i=1}^{n}([\boldsymbol{x}]_i - \mu_2)^2\right\} + 10\sum_{i=1}^{n}(1-\cos(2\pi[\boldsymbol{x}]_i)) \ ,$$

where $(\mu_1, d, s) \in \mathbb{R}^3$ and $\mu_2 = -\sqrt{\mu_1^2 - d/s}$, this general structure is actually a trap. Others display little general structure and algorithms need to fall in the right optimum. These functions can be composed by a diagonal matrix and/or rotations to further study the effect of ill-conditioning and non-separability on the performances of optimization algorithms.

2.4.3 Constrained problems

In constrained optimization, an algorithm has to optimize a real-valued function f defined on a subset of $\mathbb{R}^n$ which is usually defined by inequality functions $(g_i)_{i\in[1..r]}$ and equality functions $(h_i)_{i\in[1..s]}$. The problem for minimization then reads

$$\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x})$$

$$\text{subject to } g_i(\boldsymbol{x}) \geq 0 \text{ for } i \in [1..r] \text{ and}$$
$$h_i(\boldsymbol{x}) = 0 \text{ for } i \in [1..s]$$

Constraints can be linear or non-linear. Linear constraints appear frequently as some variables are required to be positive or bounded. When all coordinates are bounded, the problem is said to be box constrained. Constraints can also be hard (solutions are not allowed to violate the constraints) or soft (violation is possible but penalized). The set of points for which the constraints are satisfied is called the **feasible set**. Note that an equality constraint $h(\boldsymbol{x}) = 0$

can be modelled by two inequality constraints $h(\boldsymbol{x}) \geq 0$ and $-h(\boldsymbol{x}) \geq 0$, so for simplicity of notations we consider in the following only inequality constraints.
In the case of constrained problems, necessary conditions on a C^1 objective function for the minimality of $f(\boldsymbol{x}^*)$, such as $\nabla_{\boldsymbol{x}^*} f = \mathbf{0}$, do not hold. Indeed, an optimum $\boldsymbol{x}^*$ can be located on constraint boundaries. Instead, Karush-Kuhn-Tucker (KKT) conditions [82, 86] offer necessary first order conditions for the minimality of $f(\boldsymbol{x}^*)$.
Real world problems often impose constraints on the problem, but many continuous optimization algorithms are designed for unconstrained problems [128, 28]. For some optimization algorithms a version for box constraints has been specifically developed (e.g. BOBYQA [113] for NEWUOA [112], L-BFGS-B [38] for L-BFGS [90]). In general, many techniques have been developed to apply these algorithms to constrained problems, and a lot of investigation has been done on the behaviour of different algorithms coupled with different constraint-handling methods, on different search functions [103, 50, 100, 6, 124, 109].
An overview of constraint-handling methods for Evolutionary Algorithms has been concluded in [51, 100]. Since ESs, which are Evolutionary Algorithms, are the focus of this thesis, following [51, 100] we present a classification of constraint handling methods for Evolutionary Algorithms:

- **Resampling**: if new samples are generated through a random variable that has positive probability of being in the feasible set, then if it is not feasible it can be resampled until it lies in the feasible set. Although this method is simple to code, resampling can be computationally expensive, or simply infeasible with equality constraints.

- **Penalty functions**: penalty functions transform the constrained problem in an unconstrained one by adding a component to the objective function which penalizes points close to the constraint boundary and unfeasible points [109, Chapter 15,17][133]. The problem becomes $\min_{\boldsymbol{x}\in\mathbb{R}^n} f(\boldsymbol{x}) + p(\boldsymbol{x})/\mu$ where p is the penalty function and $\mu \in \mathbb{R}_+^*$ is the penalty parameter and determines the importance of not violating the constraint, and the constrained problem can be solved by solving the unconstrained one with decreasing values of μ [109, Chapter 15]. The penalty parameter is often adapted throughout the optimization (see e.g. [93, 65]). Generally, $p(\boldsymbol{x}) = 0$ if $\boldsymbol{x}$ is feasible [100], although for barrier methods unfeasible solutions are given an infinite fitness value, and $p(\boldsymbol{x})$ increases as $\boldsymbol{x}$ goes near the constraints boundaries [109, 133]. Usually the function p is a function of the distance to the constraint, or a function of the amount of violated constraints [133]. A well-known penalty function is the augmented Lagrangian [29] which combine quadratic penalty functions [137] with Lagrange multipliers from the KKT conditions into $p(\boldsymbol{x}) = \sum_{i=1}^{r} p_i(\boldsymbol{x})$ where $p_i(\boldsymbol{x}) = -\lambda_t^i g_i(\boldsymbol{x}) + g_i(\boldsymbol{x})^2/(2\mu)$ if $g_i(\boldsymbol{x}) - \mu\lambda_i \leq 0$, and $p_i(\boldsymbol{x}) = -\mu\lambda_i^2/2$ otherwise. The coefficients $(\lambda_i)_{i\in[1..r]}$ are estimates of the Lagrangian multipliers of the KKT conditions, and are adapted through $\lambda_i \leftarrow \max(\lambda_i - g_i(\boldsymbol{x})/\mu, 0)$.

- **Repairing**: repairing methods replace unfeasible points with feasible points, e.g. by projecting the unfeasible point to the nearest constraint boundary [6]. See [124] for a survey of repair methods.

- **Special operators or representations**: these methods ensures that new points cannot be unfeasible by changing how the points are sampled directly in the algorithm, or finding a representation mapping the feasible space X to $\mathbb{R}^d$ [105, 85, 102]. In [85], the feasible space is mapped to a n-dimensional cube (which corresponds to $\mathbb{R}^n$ with specific linear constraints), and in [105] the feasible space constrained by linear functions is mapped to the unconstrained space $\mathbb{R}^n$. Resampling and repair can also be considered as special operators.

- **Multiobjective optimization**: contrarily to penalty functions where the objective function and constraint functions are combined into a new objective function, the constrained problem can be seen instead as a problem where both the objective function and the violation of the constraints are optimized as a multiobjective problem (see [99] for a survey).

2.4.4 Noisy problems

A function is said noisy when the reevaluation of the f-value of a point $\boldsymbol{x}$ can lead to a different value. Noisy functions are important to study as many real-world problems contain some noise due to the imperfection of measurements, data, or because simulations are used to obtain a value of the function to be optimized. For $\boldsymbol{x} \in X$, the algorithm does not have direct access to $f(\boldsymbol{x})$, but instead the algorithm queries a random variable $F(\boldsymbol{x})$. Different distributions for $F(\boldsymbol{x})$ have been considered [17], and correspond to different noise models, e.g.

$$\begin{aligned}
\textbf{Additive noise } [80]: \; & F(\boldsymbol{x}) \overset{d}{=} f(\boldsymbol{x}) + N \\
\textbf{Multiplicative noise } [5]: \; & F(\boldsymbol{x}) \overset{d}{=} f(\boldsymbol{x})(1+N) \\
\textbf{Actuator noise } [131]: \; & F(\boldsymbol{x}) \overset{d}{=} f(\boldsymbol{x}+\boldsymbol{N}) \;\; ,
\end{aligned}$$

where N and $\boldsymbol{N}$ are random elements. When N is a standard normal variable, the noise is called Gaussian noise [80]. Other distributions for N have been studied in [12], such as Cauchy distributions in [11].
The inaccuracy of the information acquired by an optimization algorithm on a noisy function (and so, the difficulty induced by the noise) is directly connected to the variation of f-value respectively to the variance of the noise, called the signal-to-noise ratio [65]. In fact, for additive noise on the sphere function where this ratio goes to 0 when the algorithm converges to the optimum, it has been shown in [17] that ESs do not converge log-linearly to the minimum. An overview of different techniques to reduce the influence of the noise is realized in [80]. The variance of the noise can be reduced by a factor $\sqrt{k}$ by resampling k times the same point. The number of times a point is resampled can be determined by a statistical test [39], and for EAs displaying a population of points, which point should be resampled can be chosen using the ranking of the points [1]. Another method to smooth the noise is to construct a surrogate model from the points previously evaluated [125, 36], which can average the effect of the noise. Population based algorithms, such as EAs, are naturally resilient to noise [5], and a higher

population size implicitly reduces the noise [8]. For an ES, increasing only the population size λ is inferior to using resampling [62], but increasing both λ and μ is superior [5] when the step-size is appropriately adapted.

2.4.5 Invariance to a class of transformations

Invariances [69, 70, 25] are strong properties that can make an algorithm insensitive to some difficulties. They are therefore important indicators of the robustness of an algorithm, which is especially useful in black-box optimization where the algorithms need to be effective on a wide class of problems.

An algorithm is said to be invariant to a class of transformations $\mathscr{C}$ if for all functions f and any transformation $g \in \mathscr{C}$, the algorithm behaves the same on f and $g \circ f$ or $f \circ g$, depending on the domain of g. More formally following [69], let $\mathscr{H} : \{X \to \mathbb{R}\} \to 2^{\{X \to \mathbb{R}\}}$ be a function which maps a function $f : X \to \mathbb{R}$ to a set of functions, let S denote the state space of an algorithm $\mathscr{A}$, and $\mathscr{A}_f : S \to S$ be an iteration of $\mathscr{A}$ under an objective function f. The algorithm $\mathscr{A}$ is called **invariant under** $\mathscr{H}$ if for all $f : X \to \mathbb{R}$ and $h \in \mathscr{H}(f)$ there exists a bijection $T_{f,h} : S \to S$ such that

$$\mathscr{A}_h \circ T_{f,h}(s) = T_{f,h}(s) \circ \mathscr{A}_f \quad . \tag{2.30}$$

A basic invariance is **invariance to translations**, which is expected of any optimization algorithm. An important invariance shared by all FVF algorithms is the **invariance to strictly increasing functions**. This implies that a FVF algorithm can optimize just as well a smooth function than its composition with any non-convex, non-differentiable or non-continuous function, which indicates robustness against rugged functions [58]. Another important invariance is the **invariance to rotations**. This allows a rotation invariant algorithm to have the same performances on an ellipsoid and a rotated ellipsoid, showing robustness on non-separable functions.

The No Free Lunch theorem [141] states (for discrete optimization) that improvement over a certain class of functions is offset by lesser performances on another class of functions. Algorithms exploiting a particular property of a function may improve their performances when the objective function has this property, at the cost invariance and of their performances on other functions. For example, algorithms exploiting separability are not invariant to rotations. In [70] CMA-ES (see 2.3.8) is shown to be invariant to rotations, while the performances of PSO (see 2.3.4) are shown to be greatly impacted on ill-conditioned non-separable functions. In [21] the dependence of BFGS (see 2.2.1), NEWUOA (see 2.2.2), CMA-ES and PSO on ill-conditioning and separability is investigated.

2.5 Theoretical results and techniques on the convergence of Evolution Strategies

We will present a short overview of theoretical results on ESs. Most theoretical studies on ESs are focused on isotropic ESs (that is the covariance matrix of their sampling distribution is equal to the identity matrix throughout the optimization).
Almost sure convergence of elitist ESs with constant step-size (or non-elitist ESs in a bounded search space) has been shown in [121][20] on objective functions with bounded sublevel sets $E_\epsilon := \{\boldsymbol{x} \in X | f(\boldsymbol{x}) \leq \epsilon\}$. However constant step-size implies a long expected hitting time of the order of $1/\epsilon^n$ to reach an ϵ-ball around the optimum [20], which is comparable with Pure Random Search and therefore too slow to be practically relevant. Note that when using step-size adaptation, ESs are not guaranteed convergence, and the $(1+1)$-ES using the so-called $1/5$ success rule has been shown with probability 1 to not converge to the optimum of a particular multi-modal function [122]. Similarly, on a linear function with a linear constraint, a $(1,\lambda)$-CSA-ES and a $(1,\lambda)$-σSA-ES can converge log-linearly [14, 15, 6], while on a linear function divergence is required. In constrained problems, the constraint handling mechanism can be critical to the convergence or divergence of the algorithm: for any value of the population size λ or of the cumulation parameter c a $(1,\lambda)$-CSA-ES using resampling can fail on a linear function with a linear problem, while for a high enough value of λ or low enough value of c a $(1,\lambda)$-CSA-ES using repair appears to solve any linear function with a linear constraint [6].
The convergence rate of ESs using step-size adaptation has been empirically observed to be log-linear on many problems. It has been shown in [135] that comparison based algorithms which use a bounded number of comparisons between function evaluations cannot converge faster than log-linearly. More precisely, the expected hitting time of a comparison based algorithm into a ball $B(\boldsymbol{x}^*, \epsilon)$ (where $\boldsymbol{x}^*$ is the optimum of f) is lower bounded by $n\ln(1/\epsilon)$ when $\epsilon \to 0$. And more specifically, the expected hitting time of any isotropic $(1,\lambda)$ and $(1+\lambda)$-ESs is lower bounded by $bn\ln(1/\epsilon)\lambda\ln(\lambda)$ when $\epsilon \to 0$ where $b \in \mathbb{R}_+^*$ is a proportionality constant [76, 77]. On the sphere function and some ellipsoid functions for a $(1+1)$-ES using the so-called $1/5$-success rule, the expected number of function evaluations required to decrease the approximation error $f(\boldsymbol{X}_0) - f(\boldsymbol{x}^*)$ by a factor 2^{-t} where t is polynomial in n has been shown to be $\Theta(tn)$ [74, 75].
Besides studies on the expected hitting time of ESs, a strong focus has been put in proofs of log-linear convergence, estimations of the convergence rates and the dependence between the convergence rate and the parameters of an algorithm. Note that the estimation of convergence rates or the investigation of their dependency with other parameters often involve the use of Monte-Carlo simulations. For $(\Phi_t)_{t\in\mathbb{N}}$ a positive Markov chain valued on X with invariant measure π and $h : X \to \mathbb{R}$ a function, the fact that a Monte-Carlo simulation$1/t\sum_{k=0}^{t-1} h(\Phi_k)$ converge independently of their initialisation to $\mathbf{E}_\pi(h(\Phi_0))$ is implied by the h-ergodicity of $(\Phi_t)_{t\in\mathbb{N}}$, which is therefore a crucial property. In many theoretical work on ESs this property is assumed, although as presented in 1.2.9 Markov chain theory provides tools to show ergodicity.
We will start this chapter by introducing in 2.5.1 the so-called progress rate, which can be used to obtain quantitative estimates of lower bounds on the convergence rate , and results

obtained through it. Then in 2.5.2 we will present results obtained by analysing ESs using the theory of Markov chains. And then in 2.5.3 we present ordinary differential equations underlying the IGO algorithm presented in 2.3.9.

2.5.1 Progress rate

The normalized progress rate [30]) is a measurement over one iteration of an ES, defined as the dimension of the search space n multiplied by the expected improvement in the distance to the optimum normalized by the current distance to the optimum, knowing $\boldsymbol{X}_t$ the current mean of the sampling distribution and $\boldsymbol{S}_t$ the other parameters of the algorithm or of the problem; that is

$$\varphi^* = n\mathbf{E}\left(\frac{\|\boldsymbol{X}_t - \boldsymbol{x}^*\| - \|\boldsymbol{X}_{t+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{X}_t - \boldsymbol{x}^*\|} \middle| \boldsymbol{X}_t,\ \boldsymbol{S}_t \right) \ . \tag{2.31}$$

The fact that the normalized progress rate is a measurement over one iteration links the normalized progress rate with the convergence of ESs where the step-size is kept proportional to the distance to the optimum (see [19]). On the sphere function for isotropic ESs, φ^* depends of the distance to the optimum normalized by the step-size. Thus the normalized progress rate is usually expressed as a function of the normalized step-size $\sigma^* = n\sigma_t / \|\boldsymbol{X}_t - \boldsymbol{x}^*\|$ [30], which is a constant when the step-size is kept proportional to the distance to the optimum. This has been used in [30, 117, 31] to define an optimal step-size as the value of σ^* that maximizes the normalized progress rate, and to study how the progress rate changes with σ^*. Similarly, it has been used to define optimal values for other parameters of the algorithm, such as μ/λ for the $(\mu/\mu, \lambda)$-ES [31], as the values maximizing the progress rate. Through different approximations, the dependence of the progress rate on these values is investigated [30, 31].
The progress rate lower bounds the convergence rate of ESs. Indeed, take $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ the sequence of vectors corresponding to the mean of the sampling distribution of an ES, and suppose that the sequence $(\|\boldsymbol{X}_t - \boldsymbol{x}^*\|)_{t\in\mathbb{N}}$ converges in mean log-linearly to the rate $r \in \mathbb{R}_+^*$. Since for $x \in \mathbb{R}_+^*$, $1 - x \leq -\ln(x)$, we have

$$\begin{aligned}
\varphi^* &= n\left(1 - \mathbf{E}\left(\frac{\|\boldsymbol{X}_{t+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{X}_t - \boldsymbol{x}^*\|} \middle| \boldsymbol{X}_t,\ \boldsymbol{S}_t \right)\right) \\
&\leq -n\ln\left(\mathbf{E}\left(\frac{\|\boldsymbol{X}_{t+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{X}_t - \boldsymbol{x}^*\|} \middle| \boldsymbol{X}_t,\ \boldsymbol{S}_t \right)\right) \\
&\leq -n\mathbf{E}\left(\ln\left(\frac{\|\boldsymbol{X}_{t+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{X}_t - \boldsymbol{x}^*\|} \right) \middle| \boldsymbol{X}_t,\ \boldsymbol{S}_t \right) = nr \ ,
\end{aligned}$$

so the progress rate is a lower bound to the convergence rate multiplied by n, and a positive progress rate implies that $\mathbf{E}(\ln(\|\boldsymbol{X}_{t+1} - \boldsymbol{x}^*\| / \|\boldsymbol{X}_t - \boldsymbol{x}^*\|))$ converges to a negative value. However, suppose that $\|\boldsymbol{X}_{t+1} - \boldsymbol{x}^*\| / \|\boldsymbol{X}_t - \boldsymbol{x}^*\| \sim \exp(\mathcal{N}(0,1) - a)$ for $a \in \mathbb{R}_+^*$. Then if a is small enough, then $\mathbf{E}(\|\boldsymbol{X}_{t+1} - \boldsymbol{x}^*\| / \|\boldsymbol{X}_t - \boldsymbol{x}^*\|) \geq 1$ which imply a negative progress rate, while $\mathbf{E}(\ln(\|\boldsymbol{X}_{t+1} - \boldsymbol{x}^*\| / \|\boldsymbol{X}_t - \boldsymbol{x}^*\|)) < 0$ which implies log-linear convergence; hence a negative progress rate does not imply divergence [19]. The progress rate is therefore not a tight lower bound of the convergence rate of ESs.

To correct this, the log-progress rate $\varphi^*_{\ln}$ [19] can be considered. It is defined as

$$\varphi^*_{\ln} := n\mathbf{E}\left(\ln\left(\frac{\|\boldsymbol{X}_{t+1}-\boldsymbol{x}^*\|}{\|\boldsymbol{X}_t-\boldsymbol{x}^*\|}\right)\middle|\boldsymbol{X}_t,\boldsymbol{S}_t\right) \tag{2.32}$$

By definition, the log-progress rate is equal to the expected value of the convergence rate of ESs where the step-size is kept proportional to the optimum, which as shown in [19] consists in a tight lower bound of the convergence rate of ESs. Furthermore, on the sphere function for a $(1,\lambda)$-ES the normalized progress rate and the log-progress rate coincide when the dimension goes to infinity [19, Theorem 1], which makes high dimension an important condition for the accuracy of results involving the normalized progress rate.
Extensive research has been conducted on the progress rate, which give quantitative lower bounds (i.e. that can be precisely estimated) to the convergence rate in many different scenarios [67]. The $(1+1)$-ES on the sphere function [117], sphere function with noise [10], the $(\mu/\mu,\lambda)$-ES on the sphere function [30, 31] which gives when $n\to\infty$ an optimal ratio μ/λ of 0.27 for the sphere function, sphere function with noise [9]. Different step-size adaptation mechanisms have also been studied where the normalized step-size is assumed to reach a stationary distribution, and where its expected value under the stationary distribution is approximated and compared to the optimal step-size. This has been realized for CSA (see 2.3.8) on the sphere [7] and ellipsoid functions [13], or for σSA (see (see 2.3.8)) on the linear [63] and sphere [32] functions.

2.5.2 Markov chain analysis of Evolution Strategies

Markov chain theory was first used to study the log-linear convergence of ESs in [33], which proves the log-linear convergence on the sphere function of a $(1,\lambda)$-ES where the step-size is kept proportional to the distance to the optimum. It also analyses the $(1,\lambda)$-σSA-ES on the sphere function and assumes the positivity and Harris recurrence of the Markov chain involved, from which it deduces the log-linear convergence of the algorithm. A full proof of the positivity and Harris recurrence of a Markov chain underlying the $(1,\lambda)$-σSA-ES, and so of the linear-convergence of a $(1,\lambda)$-σSA-ES on the sphere function is then realized in [18]. In [79] a scale-invariant $(1+1)$-ES on a sphere function with multiplicative noise is proven to converge log-linearly almost surely if and only if the support of the noise is a subset of $\mathbb{R}^*_+$. All of these studies use a similar methodology which is introduced in [25]. The paper [25] proposes a methodology to analyse comparison-based algorithms adapting a step-size, such as ESs, on scaling invariant functions. Scaling invariant functions are a wide class of functions which includes the sphere, the ellipsoid and the linear functions. A function $f:\mathbb{R}^n\to\mathbb{R}$ is called scaling invariant with respect to $\boldsymbol{x}^*\in\mathbb{R}^n$ if

$$f(\boldsymbol{x})\leq f(\boldsymbol{y})\Leftrightarrow f(\boldsymbol{x}^*+\rho(\boldsymbol{x}-\boldsymbol{x}^*))\leq f(\boldsymbol{x}^*+\rho(\boldsymbol{y}-\boldsymbol{x}^*))\ ,\quad \text{for all } (\boldsymbol{x},\boldsymbol{y})\in\mathbb{R}^n\times\mathbb{R}^n,\ \rho\in\mathbb{R}^*_+\ . \tag{2.33}$$

Scaling invariant functions are useful to consider in the context of comparison-based algorithms (such as ESs), as the fact that they are comparison based makes them invariant to any rescaling of the search space around $\boldsymbol{x}^*$. Note that, as shown in [25], a function which is scaling

invariant with respect to $\boldsymbol{x}^*$ cannot have any strict local optima except for $\boldsymbol{x}^*$ (and $\boldsymbol{x}^*$ may not be a local optima, e.g. for linear functions). A more structured class of scaling invariant functions is positively homogeneous functions: a function $f : \mathbb{R}^n \to \mathbb{R}$ is called positively homogeneous with degree $\alpha > 0$ if

$$f(\rho \boldsymbol{x}) = |\rho|^\alpha f(\boldsymbol{x}) \quad \text{for all } \rho > 0 \text{ and } \boldsymbol{x} \in \mathbb{R}^n. \tag{2.34}$$

As shown in [25] the class of scaling invariant functions is important for ESs as, under a few assumptions, on scaling invariant functions the sequence $(\boldsymbol{X}_t/\sigma_t)_{t\in\mathbb{N}}$ is a time-homogeneous Markov chain. Proving that this Markov chain is positive and Harris recurrent can be used to show the linear convergence or divergence of the ES. The methodology proposed in [25] is used in [24] to show the log-linear convergence of a $(1+1)$-ES with a step-size adaptation mechanism called the one-fifth success rule [117] on positively homogeneous functions.

2.5.3 IGO-flow

Let $(P_\theta)_{\theta\in\Theta}$ denote a family of probability distributions parametrized by $\theta \in \Theta$. The IGO-flow [111] is the set of continuous-time trajectories on the parameter space Θ defined by the ordinary differential equation

$$\frac{d\theta_t}{dt} = \boldsymbol{F}(\theta_t)^{-1} \int_{\mathbb{R}^n} W^f_{\theta_t}(\boldsymbol{x})\, \nabla_\theta \ln\left(P_\theta(\boldsymbol{x})\right)|_{\theta=\theta_t}\, P_{\theta_t}(d\boldsymbol{x}) \ , \tag{2.35}$$

where $\boldsymbol{F}(\theta_t)$ is the Fisher information matrix defined in (2.21), and $W^f_{\theta_t}$ is the weighted quantile function defined in (2.27). IGO algorithms defined in 2.3.9 are a time discretized version of the IGO-flow, where $W^f_{\theta_t}(\boldsymbol{x})$ and the gradient $\nabla_\theta \ln(P_\theta(\boldsymbol{x}))|_{\theta=\theta_t}$ are estimated using a number $\lambda \in \mathbb{N}^*$ of samples $(\boldsymbol{Y}^i)_{i\in[1..\lambda]}$ i.i.d. with distribution P_{θ_t} through the consistent estimator $1/\lambda \sum_{i=1}^{\lambda} \hat{w}_i((\boldsymbol{Y}^j)_{j\in[1..\lambda]})\, \nabla_\theta \ln(P_\theta(\boldsymbol{Y}^i))\big|_{\theta=\theta_t}$ (see [111, Theorem 6]), with $\hat{w}_i$ defined in (2.28). IGO algorithms offer through the IGO-flow a theoretically tractable model. In [2] the IGO-flow for multivariate Gaussian distributions with covariance matrix equal to $\sigma_t \boldsymbol{Id}_n$ has been shown to locally converge on C^2 functions with Λ_n-negligible level sets to critical points of the objective function that admit a positive definite Hessian matrix; this holds under the assumption that (i) the function w used in (2.27) is non-increasing, Lipschitz-continuous and that $w(0) > w(1)$; and (ii) the standard deviation σ_t diverges log-linearly on the linear function. Furthermore, as the $(\mu/\mu_W, \lambda)$-CMA-ES with rank-μ update (i.e. $c_\sigma = 1$, $c_1 = 0$, see 2.3.8) and the xNES described in 2.3.9 have both been shown to be connected with IGO for multivariate Gaussian distributions (see [111, Proposition 20, Proposition 21], results in the IGO-flow framework have impact on the CMA-ES and the NES.

Chapter 3

Contributions to Markov Chain Theory

In in this chapter we present a model for Markov chains for which we derive sufficient conditions to prove that a Markov chain is a φ-irreducible aperiodic T-chain and that compact sets are small sets for the chain. Similar results using properties of the underlying deterministic control model as presented in 1.2.6 have been previously derived in [98, Chapter 7]. These results are placed in a context where the Markov chain studied $\Phi = (\Phi_t)_{t\in\mathbb{N}}$, valued on a state space X which is a open subset of $\mathbb{R}^n$, can be defined through

$$\Phi_{t+1} = G(\Phi_t, \boldsymbol{U}_{t+1}) \ , \tag{3.1}$$

where $G : X \times \mathbb{R}^p \to X$ is a measurable function that we call the transition function, and $(\boldsymbol{U}_t)_{t\in\mathbb{N}^*}$ is a i.i.d. sequence of random elements valued in $\mathbb{R}^p$. To obtain the results of [98, Chapter 7] the transition function G is assumed to be C^∞ and the random element $\boldsymbol{U}_1$ is assumed to admit a lower semi-continuous density p. However the transition functions as described in (3.1) of most of the Markov chains that we study in the context of ESs are not C^∞, and not even continuous due to the selection mechanism in ESs, and so the results of [98, Chapter 7] cannot be applied to most of our problems.
However, we noticed in our problems the existence of $\alpha : X \times \mathbb{R}^p \to O$ a measurable function where O is an open subset of $\mathbb{R}^m$, such that there exists a C^∞ function $F : X \times O \to X$ for which we can define our Markov chain through

$$\Phi_{t+1} = F(\Phi_t, \alpha(\Phi_t, \boldsymbol{U}_{t+1})) \ . \tag{3.2}$$

With this new model where the function α is typically discontinuous, and the sequence $(\boldsymbol{W}_{t+1})_{t\in\mathbb{N}} = (\alpha(\Phi_t, \boldsymbol{U}_{t+1}))_{t\in\mathbb{N}}$ is typically not i.i.d., we give sufficient conditions related to the ones of [98, Chapter 7] to prove that a Markov chain is φ-irreducible, aperiodic T-chain and that compact sets are small sets. These conditions are

1. the transition function F is C^1,
2. for all $\boldsymbol{x} \in X$ the random element $\alpha(\boldsymbol{x}, \boldsymbol{U}_1)$ admits a density $p_{\boldsymbol{x}}$,
3. the function $(\boldsymbol{x}, \boldsymbol{w}) \mapsto p_{\boldsymbol{x}}(\boldsymbol{w})$ is lower semi-continuous,

4. there exists $\boldsymbol{x}^* \in X$ a strongly globally attracting state, $k \in \mathbb{N}^*$ and $\boldsymbol{w}^* \in O_{\boldsymbol{x}^*,k}$ such that $F^k(\boldsymbol{x}^*, \cdot)$ is a submersion at $\boldsymbol{w}^*$.

The set $O_{\boldsymbol{x}^*,k}$ is the support of the conditional density of $(\boldsymbol{W}_t)_{t\in[1..k]}$ knowing that $\Phi_0 = \boldsymbol{x}^*$; F^k is the k-steps transition function inductively defined by $F^1 = F$ and $F^{t+1}(\boldsymbol{x}, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_{t+1}) = F^t(F(\boldsymbol{x}, \boldsymbol{w}_1), \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{t+1})$; and the concept of strongly globally attracting states is introduced in the paper presented in this chapter, namely $\boldsymbol{x}^* \in X$ is called a strongly globally attracting state if

$$\forall \boldsymbol{y} \in X,\ \forall \epsilon > 0,\ \exists t_{\boldsymbol{y},\epsilon} \in \mathbb{N}^* \text{ such that } \forall t \geq t_{\boldsymbol{y},\epsilon},\ A^t_+(\boldsymbol{y}) \cap B(\boldsymbol{x}^*, \epsilon) \neq \emptyset\ , \tag{3.3}$$

with $A^t_+(\boldsymbol{y})$ the set of states reachable at time t from $\boldsymbol{y}$, as defined in (1.11).
To appreciate these results it is good to know that proving the irreducibility and aperiodicity of some Markov chains exhibited in [25] used to be a ad-hoc and tedious process, in some cases very long and difficult[1], while proving so is now relatively trivial.
We present this new model and these conditions in the following paper, and in the same paper we use these conditions to show the φ-irreducibility, aperiodicity and the property that compact sets are small sets, for Markov chains underlying the so-called xNES algorithm [59] with identity covariance matrix on scaling invariant functions, and for the $(1, \lambda)$-CSA-ES algorithm on a linear constrained problem with the cumulation parameter c_σ equal to 1, which were problems we could not solve before these results.

3.1 Paper: Verifiable Conditions for Irreducibility, Aperiodicity and T-chain Property of a General Markov Chain

The following paper [42] will soon be submitted to Bernoulli, and presents sufficient conditions for the irreducibility, aperiodicity, T-chain property and the property that compact sets are petite sets for a Markov chain, and then presents some applications of these conditions to problems involving ESs as mentioned in the beginning of this chapter. The different ideas and proofs in this work are a contribution of the first author. The second author gave tremendous help to give the paper the right shape, and to proof read as well as discuss the different ideas and proofs.

[1] Anne Auger, private communication, 2013.

arXiv: math.PR/0000000

Verifiable Conditions for Irreducibility, Aperiodicity and T-chain Property of a General Markov Chain

ALEXANDRE CHOTARD[1]
ANNE AUGER[1]

[1] *TAO Team - Inria Saclay - Île-de-France Université Paris-Sud, LRI. Rue Noetzlin, Bât. 660, 91405 ORSAY Cedex - France E-mail:* alexandre.chotard@gmail.com; anne.auger@inria.fr

We consider in this paper Markov chains on a state space being an open subset of $\mathbb{R}^n$ that obey the following general non linear state space model: $\Phi_{t+1} = F\left(\Phi_t, \alpha(\Phi_t, \mathbf{U}_{t+1})\right), t \in \mathbb{N}$, where $(\mathbf{U}_t)_{t\in\mathbb{N}^*}$ (each $\mathbf{U}_t \in \mathbb{R}^p$) are i.i.d. random vectors, the function α, taking values in $\mathbb{R}^m$, is a measurable typically discontinuous function and $(\mathbf{x}, \mathbf{w}) \mapsto F(\mathbf{x}, \mathbf{w})$ is a C^1 function. In the spirit of the results presented in the chapter 7 of the Meyn and Tweedie book on *"Markov Chains and Stochastic Stability"*, we use the underlying deterministic control model to provide sufficient conditions that imply that the chain is a φ-irreducible, aperiodic T-chain with the support of the maximality irreducibility measure that has a non empty interior. Our results rely on the coupling of the functions F and α: we assume that for all $\mathbf{x}$, $\alpha(\mathbf{x}, \mathbf{U}_1)$ admits a lower semi-continuous density and then pass the discontinuities of the overall update function $(\mathbf{x}, \mathbf{u}) \mapsto F(\mathbf{x}, \alpha(\mathbf{x}, \mathbf{u}))$ into the density while the function $(\mathbf{x}, \mathbf{w}) \mapsto F(\mathbf{x}, \mathbf{w})$ is assumed C^1. In contrast, using previous results on our modelling would require to assume that the function $(\mathbf{x}, \mathbf{u}) \mapsto F(\mathbf{x}, \alpha(\mathbf{x}, \mathbf{u}))$ is C^∞.

We introduce the notion of a *strongly globally attracting state* and we prove that if there exists a strongly globally attracting state and a time step k, such that we find a k-path such that the k^{th} transition function starting from $\mathbf{x}^*$, $F^k(\mathbf{x}^*, .)$, is a submersion at this k-path, the the chain is a φ-irreducible, aperiodic, T-chain.

We present two applications of our results to Markov chains arising in the context of adaptive stochastic search algorithms to optimize continuous functions in a black-box scenario.

Keywords: Markov Chains, Irreducibility, Aperiodicity, T-chain, Control model, Optimization.

Contents

1. Introduction

Let X be an open subset of $\mathbb{R}^n$ and O an open subset of $\mathbb{R}^m$ equipped with their Borel sigma-algebra $\mathcal{B}(X)$ and $\mathcal{B}(O)$ for n, m two integers. This paper considers Markov chains $\Phi = (\Phi_t)_{t\in\mathbb{N}}$ defined on X via a multidimensional non-linear state space model

$$\Phi_{t+1} = G\left(\Phi_t, \mathbf{U}_{t+1}\right), \quad t \in \mathbb{N} \tag{1}$$

where $G : X \times \mathbb{R}^p \to X$ (for $p \in \mathbb{N}$) is a measurable function ($\mathbb{R}^p$ being equipped of the Borel sigma-algebra) and $(\mathbf{U}_t)_{t\in\mathbb{N}^*}$ is an i.i.d. sequence of random vectors valued in $\mathbb{R}^p$ and defined on a probability space $(\Omega, \mathcal{A}, \mathbb{P})$ independent of Φ_0 also defined on the same probability space, and valued in X. In addition, we assume that Φ admits an alternative representation under the form

$$\Phi_{t+1} = F\left(\Phi_t, \alpha(\Phi_t, \mathbf{U}_{t+1})\right) \ , \tag{2}$$

where $F : X \times O \to X$ is in a first time assumed measurable, but will typically be C^1 unless explicitly stated and $\alpha : X \times \mathbb{R}^p \to O$ is measurable and can typically be discontinuous. The functions F, G and α are connected via $G(\mathbf{x}, \mathbf{u}) = F(\mathbf{x}, \alpha(\mathbf{x}, \mathbf{u}))$ for any $\mathbf{x}$ in X and $\mathbf{u} \in \mathbb{R}^p$ such that G can also be typically *discontinuous*.

Deriving φ-irreducibility and aperiodicity of a general chain defined via (1) can sometimes be relatively challenging. An attractive way to do so is to investigate the underlying deterministic *control model* and use the results presented in [8, Chapter 7] that connect properties of the control model to the irreducibility and aperiodicity of the chain. Indeed, it is typically easy to manipulate deterministic trajectories and prove properties related to this deterministic path. Unfortunately, the conditions developed in [8, Chapter 7] assume in particular that G is C^∞ and $\mathbf{U}_t$ admits a lower semi-continuous density such that they cannot be applied to settings where G is discontinuous.

In this paper, following the approach to investigate the underlying control model for chains defined with (2), we develop general conditions that allow to easily verify φ-irreducibility, aperiodicity, the fact that the chain is a T-chain and identify that compact sets are small sets for the chain. Our approach relies on the fundamental assumptions that while α can be discontinuous, given $\mathbf{x} \in X$, $\alpha(\mathbf{x}, \mathbf{U})$ for $\mathbf{U}$ distributed as $\mathbf{U}_t$ admits a density $p_{\mathbf{x}}(\mathbf{w})$ where $\mathbf{w} \in O$ such that $p(\mathbf{x}, \mathbf{w}) = p_{\mathbf{x}}(\mathbf{w})$ is lower semi-continuous. Hence we "pass" the discontinuity of G coming from the discontinuity of α into this density.

The model (2) is motivated by Markov chains arising in the stochastic black-box optimization context. Generally, Φ_t represents the state of a stochastic algorithm, for instance mean and covariance matrix of a multivariate normal distribution used to sample

candidate solutions, $\mathbf{U}_{t+1}$ contains the random inputs to sample the candidate solutions and $\alpha(\Phi_t, \mathbf{U}_{t+1})$ models the selection of candidate solutions according to a black-box function $f : \mathbb{R}^d \to \mathbb{R}$ to be optimized. This selection step is usually discontinuous as points having similar function values can stem from different sampled vectors $\mathbf{U}_{t+1}$ pointing to different solutions $\alpha(\Phi_t, \mathbf{U}_{t+1})$ belonging however to the same level set. The function F corresponds then to the update of the state of the algorithm given the selected solutions, and this update can be chosen to be at least C^1. Some more detailed examples will be presented in Section 4. For some specific functions to be optimized, proving the *linear convergence* of the optimization algorithm can be done by investigating stability properties of a Markov chain underlying the optimization algorithm and following (2) [1, 2, 3]. Aperiodicity and φ-irreducibility are then two basic properties that generally need to be verified. This verification can turn out to be very challenging without the results developed in this paper. In addition, Foster-Lyapunov drift conditions are usually used to prove properties like Harris-recurrence, positivity or geometric ergodicity. Those drift conditions hold outside *small sets*. It is thus necessary to identify some small sets for the Markov chains.

Overview of the main results and structure of the paper The results we present stating the φ-irreducibility of a Markov chain defined via (2) uses the concept of global attractiveness of a state–also used in [8]–that is a state that can be approached infinitely close from any initial state. We prove in Theorem 2 that if F is C^1 and the density $p_{\mathbf{x}}(\mathbf{w})$ is lower semi-continuous, then the existence of a globally attractive state $\mathbf{x}^*$ for which at some point in time, say k, we have a deterministic path such that the k^{th} transition function starting from $\mathbf{x}^*$, $F^k(\mathbf{x}^*, .)$ is a submersion at this path, implies the φ-irreducibility of the chain. If we moreover assume that F is C^∞, we can transfer the Theorem 7.2.6 of [8] to our setting and show that if the model is *forward accessible*, then φ-irreducibility is equivalent to the existence of a globally attracting state.

To establish the aperiodicity, we introduce the notion of a *strongly globally attracting state* that is, informally speaking, a globally attracting state, $\mathbf{x}^*$, for which for any initial state and any distance $\epsilon > 0$, there exists a time step, say t_y, such that we find for all time step larger than t_y a deterministic path that puts the chain within distance ϵ of $\mathbf{x}^*$. We then prove in Theorem 3 that under the same conditions than for the φ-irreducibility but holding at a strongly globally attracting state (instead of only a globally attracting state), the chain is φ-irreducible and aperiodic.

Those two theorems contain the main ingredients to prove the main theorem of the paper, Theorem 1, that under the same conditions than for the aperiodicity states that the chain is a φ-irreducible aperiodic T-chain for which compact sets are small sets.

This paper is structured as follows. In Section 2, we introduce and remind several definitions related to the Markov chain model of the paper needed all along the paper. We also present a series of technical results that are necessary in the next sections. In Section 3 we present the main result, i.e. Theorem 1, that states sufficient conditions for a Markov chain to be a φ-irreducible aperiodic T-chain for which compact sets are small sets. This result is a consequence of the propositions established in the subsequent subsections, namely Theorem 2 for the φ-irreducibility, Theorem 3 for the aperiodicity

and Proposition 5 for the weak-Feller property. We also derive intermediate propositions and corollaries that clarify the connection between our results and the ones of [8, Chapter 7] (Proposition 3, Corollary 1) and that characterize the support of the maximal irreducibility measure (Proposition 4). We present in Section 4 two applications of our results. We detail two homogeneous Markov chains associated to two adaptive stochastic search algorithms aiming at optimizing continuous functions, sketch why establishing their irreducibility, aperiodicity and identifying some small sets is important while explaining why existing tools cannot be applied. We then illustrate how the assumptions of Theorem 1 can be easily verified and establish thus that the chains are φ-irreducible, aperiodic, T-chains for which compact sets are small sets.

Notations

For A and B subsets of X, $A \subset B$ denotes that A is included in B ($\subsetneq$ denotes the strict inclusion). We denote $\mathbb{R}^n$ the set of n-dimensional real vectors, $\mathbb{R}_+$ the set of non-negative real numbers, $\mathbb{N}$ the set of natural numbers $\{0, 1, \ldots\}$, and for $(a, b) \in \mathbb{N}^2$, $[a..b] = \bigcup_{i=a}^{b} \{i\}$. For $A \subset \mathbb{R}^n$, A^* denotes $A \backslash \mathbf{0}$. For X a metric space, $\mathbf{x} \in X$ and $\epsilon > 0$, $B(\mathbf{x}, \epsilon)$ denotes the open ball of center $\mathbf{x}$ and radius ϵ. For $X \subset \mathbb{R}^n$ a topological space, $\mathcal{B}(X)$ denotes the Borel σ-algebra on X. We denote Λ_n the Lebesgue measure on $\mathbb{R}^n$, and for $B \in \mathcal{B}(\mathbb{R}^n)$, μ_B denotes the trace-measure $A \in \mathcal{B}(\mathbb{R}^n) \mapsto \Lambda_n(A \cap B)$. For $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^n$, $\mathbf{x}.\mathbf{y}$ denotes the scalar product of $\mathbf{x}$ and $\mathbf{y}$, and $[\mathbf{x}]_i$ denotes the i^{th} coordinate of the vector $\mathbf{x}$ and $\mathbf{x}^T$ denotes the transpose of the vector. For a function $f : X \to \mathbb{R}^n$, we say that f is C^p if f is continuous, and its k-first derivatives exist and are continuous. For $f : X \to \mathbb{R}^n$ a differentiable function and $\mathbf{x} \in X$, $D_{\mathbf{x}} f$ denotes the differential of f at $\mathbf{x}$. A multivariate distribution with mean vector zero and covariance matrix identity is called a *standard multivariate normal distribution*, a standard normal distribution correspond to the case of the dimension 1. We use the notation $\mathcal{N}(0, \mathrm{I}_n)$ for a indicating the standard multivariate normal distribution where I_n is the identity matrix in dimension n. We use the acronym i.i.d. for independent identically distributed.

2. Definitions and Preliminary Results

The random vectors defined in the previous section are assumed measurable with respect to the Borel σ-algebras of their codomain. We denote for all t, the random vector $\alpha(\Phi_t, \mathbf{U}_{t+1})$ of O as $\mathbf{W}_{t+1}$, i.e.

$$\mathbf{W}_{t+1} := \alpha(\Phi_t, \mathbf{U}_{t+1}) \tag{3}$$

such that Φ satisfies

$$\Phi_{t+1} = F\left(\Phi_t, \mathbf{W}_{t+1}\right) \quad . \tag{4}$$

Given $\Phi_t = \mathbf{x}$, the vector $\mathbf{W}_t$ is assumed absolutely continuous with distribution $p_{\mathbf{x}}(\mathbf{w})$. The function $p(\mathbf{x}, \mathbf{w}) = p_{\mathbf{x}}(\mathbf{w})$ will be assumed lower semi-continuous in the whole paper.

We remind the definition of a substochastic transition kernel as well as of a transition kernel. Let $K : X \times \mathcal{B}(X) \to \mathbb{R}_+$ such that for all $A \in \mathcal{B}(X)$, the function $\mathbf{x} \in X \mapsto K(\mathbf{x}, A)$ is a non-negative measurable function, and for all $\mathbf{x} \in X$, $K(\mathbf{x}, \cdot)$ is a measure on $\mathcal{B}(X)$. If $K(\mathbf{x}, X) \leq 1$ then K is called a *substochastic transition kernel*, and if $K(\mathbf{x}, X) = 1$ then K is called a *transition kernel*.

Given F and $p_{\mathbf{x}}$ we define for all $\mathbf{x} \in X$ and all $A \in \mathcal{B}(X)$

$$P(\mathbf{x}, A) = \int 1_A(F(\mathbf{x}, \mathbf{w}))p_{\mathbf{x}}(\mathbf{w})d\mathbf{w} \ . \tag{5}$$

Then the function $\mathbf{x} \in X \mapsto P(\mathbf{x}, A)$ is measurable for all $A \in \mathcal{B}(X)$ (as a consequence of Fubini's theorem) and for all $\mathbf{x}$, $P(\mathbf{x}, .)$ defines a measure on $(X, \mathcal{B}(X))$. Hence $P(\mathbf{x}, A)$ defines a transition kernel. It is immediate to see that this transition kernel corresponds to the transition kernel of the Markov chain defined in (2) or (4).

For $\mathbf{x} \in X$, we denote $O_{\mathbf{x}}$ the set of $\mathbf{w}$ such that $p_{\mathbf{x}}$ is strictly positive, i.e.

$$O_{\mathbf{x}} := \{\mathbf{w} \in O | p_{\mathbf{x}}(\mathbf{w}) > 0\} = p_{\mathbf{x}}^{-1}((0, +\infty)) \tag{6}$$

that we call support of $p_{\mathbf{x}}$[1]. Similarly to [8, Chapter 7] we consider the recursive functions F^t for $t \in \mathbb{N}^*$ such that $F^1 := F$ and for $\mathbf{x} \in X$ and $(\mathbf{w}_i)_{i \in [1..t+1]} \in O^{t+1}$

$$F^{t+1}(\mathbf{x}, \mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{t+1}) := F\left(F^t(\mathbf{x}, \mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_t), \mathbf{w}_{t+1}\right) \ . \tag{7}$$

The function F^t is connected to the Markov chain $\Phi = (\Phi_t)_{t \in \mathbb{N}}$ defined via (4) in the following manner

$$\Phi_t = F^t(\Phi_0, \mathbf{W}_1, \ldots, \mathbf{W}_t) \ . \tag{8}$$

In addition, we define $p_{\mathbf{x},t}$ as $p_{\mathbf{x}}$ for $t = 1$ and for $t > 1$

$$p_{\mathbf{x},t}((\mathbf{w}_i)_{i \in [1..t]}) := p_{\mathbf{x},t-1}((\mathbf{w}_i)_{i \in [1..t-1]}) p_{F^{t-1}(\mathbf{x}, (\mathbf{w}_i)_{i \in [1..t-1]})}(\mathbf{w}_t) \ , \tag{9}$$

that is

$$p_{\mathbf{x},t}((\mathbf{w}_i)_{i \in [1..t]}) = p_{\mathbf{x}}(\mathbf{w}_1) p_{F(\mathbf{x}, \mathbf{w}_1)}(\mathbf{w}_2) \ldots p_{F^{t-1}(\mathbf{x}, \mathbf{w}_1, \ldots, \mathbf{w}_{t-1})}(\mathbf{w}_t) \ . \tag{10}$$

Then $p_{\mathbf{x},t}$ is measurable as the composition and product of measurable functions. Let $O_{\mathbf{x},t}$ be the support of $p_{\mathbf{x},t}$

$$O_{\mathbf{x},t} := \{\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_t) \in O^t | p_{\mathbf{x},t}(\mathbf{w}) > 0\} = p_{\mathbf{x},t}^{-1}((0, +\infty)) \ . \tag{11}$$

Then by the measurability of $p_{\mathbf{x},t}$, $O_{\mathbf{x},t}$ is a Borel set of O^t (endowed with the Borel σ-algebra). Note that $O_{\mathbf{x},1} = O_{\mathbf{x}}$.

Given $\Phi_0 = \mathbf{x}$, the function $p_{\mathbf{x},t}$ is the joint probability distribution function of $(\mathbf{W}_1, \ldots, \mathbf{W}_t)$.

[1] Note that the support is often defined as the closure of what we call support here.

Since $p_{\mathbf{x},t}$ is the joint probability distribution of $(\mathbf{W}_1, \ldots, \mathbf{W}_t)$ given $\Phi_0 = \mathbf{x}$ and because Φ_t is linked to F^t via (8), the t-steps transition kernel P^t of Φ writes

$$P^t(\mathbf{x}, A) = \int_{O_{\mathbf{x},t}} \mathbf{1}_A \left(F^t(\mathbf{x}, \mathbf{w}_1, \ldots, \mathbf{w}_t) \right) p_{\mathbf{x},t}(\mathbf{w}) \mathrm{d}\mathbf{w} \ , \tag{12}$$

for all $\mathbf{x} \in X$ and all $A \in \mathcal{B}(X)$.

The deterministic system with trajectories

$$\mathbf{x}_t = F^t(\mathbf{x}_0, \mathbf{w}_1, \ldots, \mathbf{w}_t) = F^t(\mathbf{x}_0, \mathbf{w})$$

for $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_t) \in O_{\mathbf{x},t}$ and for any $t \in \mathbb{N}^*$ is called the *associated control model* and is denoted CM(F). Using a similar terminology to Meyn and Tweedie's [8], we say that $O_{\mathbf{x}}$ is a control set for CM(F). We introduce the notion of t-steps path from a point $\mathbf{x} \in X$ to a set $A \in \mathcal{B}(X)$ as follows:

Definition 1 (t-steps path)**.** *For $\mathbf{x} \in X$, $A \in \mathcal{B}(X)$ and $t \in \mathbb{N}^*$, we say that $\mathbf{w} \in O^t$ is a t-steps path from $\mathbf{x}$ to A if $\mathbf{w} \in O_{\mathbf{x},t}$ and $F^t(\mathbf{x}, \mathbf{w}) \in A$.*

Similarly to chapter 7 of Meyn-Tweedie, we define

$$A_+^k(\mathbf{x}) := \{F^k(\mathbf{x}, \mathbf{w}) | \mathbf{w} \in O_{\mathbf{x},k}\}$$

that is the set of all states that can be reached from $\mathbf{x}$ after k steps.

Note that this definition depends on the probability density function $p_{\mathbf{x}}$ that determines the set of control sequences $\mathbf{w} = (\mathbf{w}_1, \ldots, \mathbf{w}_k)$ via the definition of $O_{\mathbf{x},k}$. More precisely, several density functions equal almost everywhere can be associated to a same random vector $\alpha(\mathbf{x}, \mathbf{U}_1)$. However, they can generate different sets $A_+^k(\mathbf{x})$.

Following [8], the set of states that can be reached starting from $\mathbf{x}$ at some time in the future from $\mathbf{x}$ is defined as

$$A_+(\mathbf{x}) = \bigcup_{k=0}^{+\infty} A_+^k(\mathbf{x}) \ .$$

The associated control model CM(F) is *forward accessible* if for all $\mathbf{x}$, $A_+(\mathbf{x})$ has non empty interior [6].

Finally, a point $\mathbf{x}^*$ is called a *globally attracting state* if for all $\mathbf{y} \in X$,

$$\mathbf{x}^* \in \bigcap_{N=1}^{+\infty} \overline{\bigcup_{k=N}^{+\infty} A_+^k(\mathbf{y})} := \Omega_+(\mathbf{y}) \ . \tag{13}$$

Although in general $\Omega_+(\mathbf{y}) \neq \overline{A_+(\mathbf{y})}$, these two sets can be used to define globally attracting states, as shown in the following proposition.

Proposition 1. *A point $\mathbf{x}^* \in X$ is a globally attracting state if and only if for all $\mathbf{y} \in X$, $\mathbf{x}^* \in \overline{A_+(\mathbf{y})}$.*

Equivalently, a point $\mathbf{x}^* \in X$ is a globally attracting state if and only if for all $\mathbf{y} \in X$ and any $U \in \mathcal{B}(X)$ neighbourhood of $\mathbf{x}^*$, there exists $t \in \mathbb{N}^*$ such that there exists a t-steps path from $\mathbf{y}$ to U.

Proof. Let us prove the first equivalence. Let $\mathbf{x}^*$ be a globally attracting state. According to (13), $\mathbf{x}^* \in \overline{\bigcup_{k=1}^{+\infty} A_+^k(\mathbf{y})} = \overline{A_+(\mathbf{y}) \backslash \{\mathbf{y}\}} \subset \overline{A_+(\mathbf{y})}$, so $\mathbf{x}^* \in \overline{A_+(\mathbf{y})}$.

Let $\mathbf{x}^*$ such that for all $\mathbf{y} \in X$, $\mathbf{x}^* \in \overline{A_+(\mathbf{y})}$. We want to show that for all $\mathbf{y} \in X$, $\mathbf{x}^* \in \bigcap_{N=1}^{+\infty} \overline{\bigcup_{k=N}^{+\infty} A_+^k(\mathbf{y})}$, so that for all $N \in \mathbb{N}^*$, $\mathbf{x}^* \in \overline{\bigcup_{k=N}^{+\infty} A_+^k(\mathbf{y})}$. Let $N \in \mathbb{N}^*$. Note that for any $\tilde{\mathbf{y}} \in A_+^N(\mathbf{y})$, $\overline{\bigcup_{k=N}^{+\infty} A_+^k(\mathbf{y})} \supset \overline{A_+(\tilde{\mathbf{y}})}$. And by hypothesis, $\mathbf{x}^* \in \overline{A_+(\tilde{\mathbf{y}})}$ so $\mathbf{x}^* \in \overline{\bigcup_{k=N}^{+\infty} A_+^k(\mathbf{y})}$.

For the first implication of the second equivalence, let us take U a neighbourhood of $\mathbf{x}^*$, and suppose that $\mathbf{x}^*$ is a globally attracting state, which as we showed in the first part of this proof, implies that for all $\mathbf{y} \in X$, $\mathbf{x}^* \in \overline{A_+(\mathbf{y})}$. This implies the existence of a sequence $(\mathbf{y}_k)_{k\in\mathbb{N}}$ of points of $A_+(\mathbf{y})$ converging to $\mathbf{x}^*$. Hence there exists a $k \in \mathbb{N}$ such that $\mathbf{y}_k \in U$, and since $\mathbf{y}_k \in A_+(\mathbf{y})$, then either there exists $t \in \mathbb{N}^*$ such that there is a t-steps path from $\mathbf{y}$ to $\mathbf{y}_k \in U$, or either $\mathbf{y}_k = \mathbf{y}$. In the latter case, we can take any $\mathbf{w} \in O_{\mathbf{y}}$ and consider $F(\mathbf{y}, \mathbf{w})$: from what we just showed, either there exists $t \in \mathbb{N}^*$ and $\mathbf{u}$ a t-steps path from $F(\mathbf{y}, \mathbf{w})$ to U, in which case $(\mathbf{w}, \mathbf{u})$ is a $t+1$-steps path from $\mathbf{y}$ to U; either $F(\mathbf{y}, \mathbf{w}) \in U$, in which case $\mathbf{w}$ is a 1-step path from $\mathbf{y}$ to U.

Now suppose that for all $\mathbf{y} \in X$ and U neighbourhood of $\mathbf{x}^*$, there exists $t \in \mathbb{N}^*$ such that there exists a t-steps path from $\mathbf{y}$ to U. Let $\mathbf{w}_k$ be a t_k-steps path from $\mathbf{y}$ to $B(\mathbf{x}^*, 1/k)$, and $\mathbf{y}_k$ denote $F^{t_k}(\mathbf{y}, \mathbf{w}_k)$. Then since $\mathbf{y}_k \in A_+(\mathbf{y})$ for all $k \in \mathbb{N}^*$ and that the sequence $(\mathbf{y}_k)_{k\in\mathbb{N}^*}$ converges to $\mathbf{x}^*$, we do have $\mathbf{x}^* \in \overline{A_+(\mathbf{y})}$, which according to what we previously proved, prove that $\mathbf{x}^*$ is a globally attracting state. □

The existence of a globally attracting state is linked in [8, Proposition 7.2.5] with φ-irreducibility. We will show that this link extends to our context.

We now define the notion of *strongly globally attractive state* that is needed for our result on the aperiodicity. More precisely we define:

Definition 2 (Strongly globally attracting state)**.** *A point $\mathbf{x}^* \in X$ is called a strongly globally attracting state if for all $\mathbf{y} \in X$, for all $\epsilon \in \mathbb{R}_+^*$, there exists $t_{\mathbf{y},\epsilon} \in \mathbb{N}^*$ such that for all $t \geq t_{\mathbf{y},\epsilon}$, there exists a t-steps path from $\mathbf{y}$ to $B(\mathbf{x}^*, \epsilon)$. Equivalently, for all $(\mathbf{y}, \epsilon) \in X \times \mathbb{R}_+^*$*

$$\exists\, t_{\mathbf{y},\epsilon} \in \mathbb{N}^* \text{such that } \forall\, t \geq t_{\mathbf{y},\epsilon}, A_+^t(\mathbf{y}) \cap B(\mathbf{x}^*, \epsilon) \neq \emptyset \ . \tag{14}$$

The following proposition connects globally and strongly globally attracting states.

Proposition 2. *Let $\mathbf{x}^* \in X$ be a strongly globally attracting state, then $\mathbf{x}^*$ is a globally attracting state.*

Proof. We will show its contrapositive: if $\mathbf{x}^*$ is not a globally attracting state, then according to (13) there exists $\mathbf{y} \in X$, $N \in \mathbb{N}^*$ and $\epsilon \in \mathbb{R}_+^*$ such that for all $k \geq N$, $B(\mathbf{x}^*, \epsilon) \cap A_+^k(\mathbf{y}) = \emptyset$. This holds for all $k \geq N$, and therefore with (14), for all $k \geq t_{\mathbf{y},\epsilon}$ which contradicts (14).

□

Our aim is to derive conditions for proving φ-irreducibility, aperiodicity and prove that compacts of X are small sets. We remind below the formal definitions associated to those notions as well as the definition of a weak Feller chain and a T-chain. A Markov chain Φ is φ-irreducibile if there exists a measure φ on $\mathcal{B}(X)$ such that for all $A \in \mathcal{B}(X)$

$$\varphi(A) > 0 \Rightarrow \sum_{t=1}^{\infty} P^t(\mathbf{x}, A) > 0 \text{ for all } \mathbf{x} \ . \tag{15}$$

A set C is *small* if there exists $t \geq 1$ and a non-trivial measure ν_t on $\mathcal{B}(X)$ such that for all $\mathbf{z} \in C$

$$P^t(\mathbf{z}, A) \geq \nu_t(A)\,, A \in \mathcal{B}(X) \ . \tag{16}$$

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

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

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

The transition kernel of Φ is acting on bounded functions $f : X \to \mathbb{R}$ via the following operator

$$Pf(\mathbf{x}) \mapsto \int f(\mathbf{y}) P(\mathbf{x}, d\mathbf{y}), \mathbf{x} \in X \ . \tag{17}$$

Let $\mathcal{C}(X)$ be the class of bounded continuous functions from X to $\mathbb{R}$, then Φ is *weak Feller* if P maps $\mathcal{C}(X)$ to $\mathcal{C}(X)$. This definition is equivalent to $P1_O$ is lower semicontinuous for every open set $O \in \mathcal{B}(X)$.

Let a be a probability distribution on $\mathbb{N}$, we denote

$$K_a : (\mathbf{x}, A) \in X \times \mathcal{B}(X) \mapsto \sum_{i \in \mathbb{N}} a(i) P^i(\mathbf{x}, A) \tag{18}$$

the transition kernel, the associated Markov chain being called the K_a chain with *sampling distribution* a. When a satisfy the geometric distribution

$$a_\epsilon(i) = (1 - \epsilon)\epsilon^i \tag{19}$$

for $i \in \mathbb{N}$, then the transition kernel K_{a_ϵ} is called the resolvent. If there exists a sub-stochastic transition kernel T satisfying

$$K_a(\mathbf{x}, A) \geq T(\mathbf{x}, A)$$

for all $\mathbf{x} \in X$ and $A \in \mathcal{B}(X)$ with $T(\cdot, A)$ a lower semi-continuous function, then T is called a *continuous component* of K_a ([8, p.124]). If there exists a sampling distribution a, T a continuous component of K_a, and that $T(\mathbf{x}, X) > 0$ for all $\mathbf{x} \in X$, then the Markov chain Φ is called a *T-chain* ([8, p.124]). We say that $B \in \mathcal{B}(X)$ is *uniformly accessible using a* from $A \in \mathcal{B}(X)$ if there exists $\delta \in \mathbb{R}_+^*$ such that

$$\inf_{\mathbf{x} \in A} K_a(\mathbf{x}, B) > \delta \ ,$$

which is written as $A \overset{a}{\rightsquigarrow} B$ ([8, p.116]).

2.1. Technical results

We present in this section a series of technical results that will be needed to establish the main results of the paper.

Lemma 1. *Let $A \in \mathcal{B}(X)$ with X an open set of $\mathbb{R}^n$. If for all $\mathbf{x} \in A$ there exists $V_\mathbf{x}$ an open neighbourhood of $\mathbf{x}$ such that $A \cap V_\mathbf{x}$ is Lebesgue negligible, then A is Lebesgue negligible.*

Proof. For $\mathbf{x} \in A$, let $r_\mathbf{x} > 0$ be such that $B(\mathbf{x}, r_\mathbf{x}) \subset V_\mathbf{x}$, and take $\epsilon > 0$. The set $\overline{\bigcup_{\mathbf{x} \in A} B(\mathbf{x}, r_\mathbf{x}/2) \cap B(\mathbf{0}, \epsilon)}$ is closed and bounded, so it is a compact, and $\bigcup_{\mathbf{x} \in A} V_\mathbf{x}$ is an open cover of this compact. Hence we can extract a finite subcover $(V_{\mathbf{x}_i})_{i \in I}$, and so $\bigcup_{i \in I} V_{\mathbf{x}_i} \supset A \cap B(\mathbf{0}, \epsilon)$. Hence, it also holds that $A \cap B(\mathbf{0}, \epsilon) = \bigcup_{i \in I} A \cap B(\mathbf{0}, \epsilon) \cap V_{\mathbf{x}_i}$. Since by assumption $\Lambda_n(A \cap V_{\mathbf{x}_i}) = 0$, from the sigma-additivity property of measures we deduce that $\Lambda_n(A \cap B(\mathbf{0}, \epsilon)) = 0$. So with Fatou's lemma $\int_X \mathbf{1}_A(\mathbf{x}) d\mathbf{x} \leq \liminf_{k \to +\infty} \int_X \mathbf{1}_{A \cap B(\mathbf{0}, k)}(\mathbf{x}) d\mathbf{x} = 0$, which shows that $\Lambda_n(A) = 0$. □

Lemma 2. *Suppose that $F : X \times O \to X$ is C^p for $p \in \mathbb{N}$, then for all $t \in \mathbb{N}^*$, $F^t : X \times O^t \to X$ defined as in (7) is C^p.*

Proof. By hypothesis, $F^1 = F$ is C^p. Suppose that F^t is C^p. Then the function $h : (\mathbf{x}, (\mathbf{w}_i)_{i \in [1..t+1]}) \in X \times O^{t+1} \mapsto (F^t(\mathbf{x}, \mathbf{w}_1, \ldots, \mathbf{w}_t), \mathbf{w}_{t+1})$ is C^p, and so is $F^{t+1} = F \circ h$. □

Lemma 3. *Suppose that the function $p : (\mathbf{x}, \mathbf{w}) \in X \times O \mapsto p_\mathbf{x}(\mathbf{w}) \in \mathbb{R}_+$ is lower semi-continuous and the function $F : (\mathbf{x}, \mathbf{w}) \in X \times O \mapsto F(\mathbf{x}, \mathbf{w}) \in X$ is continuous, then for all $t \in \mathbb{N}^*$ the function $(\mathbf{x}, \mathbf{w}) \in X \times O^t \mapsto p_{\mathbf{x},t}(\mathbf{w})$ defined in (9) is lower semi-continuous.*

Proof. According to Lemma 2, F^t is continuous. By hypothesis, the function p is lower semi-continuous, which is equivalent to the fact that $p^{-1}((a, +\infty))$ is an open set for all $a \in \mathbb{R}$. Let $t \in \mathbb{N}^*$. Suppose that $(\mathbf{x}, \mathbf{w}) \in X \times O^t \mapsto p_{\mathbf{x},t}(\mathbf{w})$ is lower semi-continuous. Let $a \in \mathbb{R}$, then the set $B_{a,t} := \{(\mathbf{x}, \mathbf{w}) \in X \times O^t | p_{\mathbf{x},t}(\mathbf{w}) > a\}\}$ is an open set. We will show that then $B_{a,t+1}$ is also an open set.

First, suppose that $a > 0$. With (9),

$$\begin{aligned} B_{a,t+1} &= \{(\mathbf{x}, \mathbf{w}, \mathbf{u}) \in X \times O^t \times O | p_{\mathbf{x},t}(\mathbf{w}) p_{F^t(\mathbf{x},\mathbf{w})}(\mathbf{u}) > a\} \\ &= \bigcup_{b \in \mathbb{R}_+^*} \{(\mathbf{x}, \mathbf{w}, \mathbf{u}) \in B_{b,t} \times O | p_{F^t(\mathbf{x},\mathbf{w})}(\mathbf{u}) > a/b\} \\ &= \bigcup_{b \in \mathbb{R}_+^*} \{(\mathbf{x}, \mathbf{w}, \mathbf{u}) \in B_{b,t} \times O | (F^t(\mathbf{x}, \mathbf{w}), \mathbf{u}) \in B_{a/b,1}\} \end{aligned}$$

The function F^t being continuous and $B_{a/b,1}$ being an open set, the set $B^F_{a/b,t+1} := \{(\mathbf{x}, \mathbf{w}, \mathbf{u}) \in X \times O^t \times O | (F^t(\mathbf{x}, \mathbf{w}), \mathbf{u}) \in B_{a/b,1}\}$ is also an open set. Therefore and as $B_{b,t}$ is an open set so is the set $(B_{b,t} \times O) \cap B^F_{a/b,t+1}$ for any $b \in \mathbb{R}$, and hence so is $B_{a,t+1} = \bigcup_{b \in \mathbb{R}_+^*} (B_{b,t} \times O) \cap B^F_{a/b,t+1}$.

If $a = 0$, note that $p_{\mathbf{x},t}(\mathbf{w}) p_{F^t(\mathbf{x},\mathbf{w})}(\mathbf{u}) > 0$ is equivalent to $p_{\mathbf{x},t}(\mathbf{w}) > 0$ and $p_{F^t(\mathbf{x},\mathbf{w})}(\mathbf{u}) > 0$; hence $B_{0,t+1} = \{(\mathbf{x}, \mathbf{w}, \mathbf{u}) \in B_{0,t} \times O | (F^t(\mathbf{x}, \mathbf{w}), \mathbf{u}) \in B_{0,1}\}$, so the same reasoning holds.

If $a < 0$, then $B_{a,t+1} = X \times O^{t+1}$ which is an open set.

So we have proven that for all a, $B_{a,t+1}$ is an open set and hence $(\mathbf{x}, \mathbf{w}) \in X \times O^{t+1} \mapsto p_{\mathbf{x},t+1}(\mathbf{w})$ is lower semi-continuous. □

Lemma 4. *Suppose that the function $F : X \times O \to X$ is C^0, and that the function $p : (\mathbf{x}, \mathbf{w}) \mapsto p_{\mathbf{x}}(\mathbf{w})$ is lower semi-continuous. Then for any $\mathbf{x}^* \in X$, $t \in \mathbb{N}^*$, $\mathbf{w}^* \in O_{\mathbf{x}^*,t}$ and V an open neighbourhood of $F^t(\mathbf{x}^*, \mathbf{w}^*)$, $P^t(\mathbf{x}^*, V) > 0$.*

Proof. Since F is C^0, from Lemma 2 F^t is also C^0. Similarly, since p is lower semi-continuous, according to Lemma 3 so is the function $(\mathbf{x}, \mathbf{w}) \mapsto p_{\mathbf{x},t}(\mathbf{w})$, and so the set $O_{\mathbf{x},t} = p^{-1}_{\mathbf{x},t}((0, +\infty))$ is open for all $\mathbf{x}$ and thus also for $\mathbf{x} = \mathbf{x}^*$. Let $B_V := \{\mathbf{w} \in O_{\mathbf{x}^*,t} | F^t(\mathbf{x}^*, \mathbf{w}) \in V\}$. Since F^t is continuous and $O_{\mathbf{x}^*,t}$ is open, the set B_V is open, and as $\mathbf{w}^* \in B_V$, it is non-empty. Furthermore

$$\begin{aligned} P^t(\mathbf{x}^*, V) &= \int_{O_{\mathbf{x}^*,t}} \mathbf{1}_V(F^t(\mathbf{x}^*, \mathbf{w})) p_{\mathbf{x}^*,t}(\mathbf{w}) d\mathbf{w} \\ &= \int_{B_V} p_{\mathbf{x}^*,t}(\mathbf{w}) d\mathbf{w} \ . \end{aligned}$$

As $p_{\mathbf{x}^*,t}$ is a strictly positive function over $B_V \subset O_{\mathbf{x}^*,t}$, and that B_V has positive Lebesgue measure, $P^t(\mathbf{x}^*, V) > 0$. □

The following lemma establishes useful properties on a C^1 function $f : X \times O \to X$ for which there exists $\mathbf{x}^* \in X$ and $\mathbf{w}^* \in O$ such that $f(\mathbf{x}^*, \cdot)$ is a submersion at $\mathbf{w}^*$, and show in particular that a limited inverse function theorem and implicit function theorem can be expressed for submersions. These properties rely on the fact that a submersion can be seen locally as the composition of a diffeomorphism by a projection, as shown in [10].

Lemma 5. *Let $f : X \times O \to X$ be a C^1 function where $X \subset \mathbb{R}^n$ and $O \subset \mathbb{R}^m$ are open sets with $m \geq n$. If there exists $\mathbf{x}^* \in X$ and $\mathbf{w}^* \in O$ such that $f(\mathbf{x}^*, \cdot)$ is a submersion at $\mathbf{w}^*$, then*

1. *there exists N an open neighbourhood of $(\mathbf{x}^*, \mathbf{w}^*)$ such that for all $(\mathbf{y}, \mathbf{u}) \in N$, $f(\mathbf{y}, \cdot)$ is a submersion at $\mathbf{u}$,*
2. *there exists $U_{\mathbf{w}^*} \subset O$ an open neighbourhood of $\mathbf{w}^*$, and $V_{f(\mathbf{x}^*, \mathbf{w}^*)}$ a neighbourhood of $f(\mathbf{x}^*, \mathbf{w}^*)$, such that $V_{f(\mathbf{x}^*, \mathbf{w}^*)}$ equals to the image of $\mathbf{w} \in U_{\mathbf{w}^*} \mapsto f(\mathbf{x}^*, \mathbf{w})$, i.e. $V_{f(\mathbf{x}^*, \mathbf{w}^*)} = f(\mathbf{x}^*, U_{\mathbf{w}^*})$,*
3. *there exists g a C^1 function from $\tilde{V}_{\mathbf{x}^*}$ an open neighbourhood of $\mathbf{x}^*$ to $\tilde{U}_{\mathbf{w}^*}$ an open neighbourhood of $\mathbf{w}^*$ such that for all $\mathbf{y} \in \tilde{V}_{\mathbf{x}^*}$*

$$f(\mathbf{y}, g(\mathbf{y})) = f(\mathbf{x}^*, \mathbf{w}^*) \ .$$

Proof. Let $(\mathbf{e}_i)_{i \in [1..m]}$ be the canonical basis of $\mathbb{R}^m$ and let us denote $f = (f_1, \ldots, f_n)^T$ the representation of f (in the canonical basis of $\mathbb{R}^n$). Similarly, $\mathbf{u} \in O$ writes in the canonical basis $\mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_m)^T$.

We start by proving the second point of the lemma. Since $f(\mathbf{x}^*, \cdot)$ is a submersion at $\mathbf{w}^*$, the matrix composed by the vectors $(D_{\mathbf{w}^*} f(\mathbf{x}^*, \cdot)(\mathbf{e}_i))_{i \in [1..m]}$ is of full rank n, hence there exists σ a permutation of $[1..m]$ such that the vectors $(D_{\mathbf{w}^*} f(\mathbf{x}^*, \cdot)(\mathbf{e}_{\sigma(i)}))_{i \in [1..n]}$ are linearly independent. We suppose that σ is the identity (otherwise we consider a reordering of the basis $(\mathbf{e}_i)_{i \in [1..m]}$ via σ). Let

$$h_{\mathbf{x}^*} : \mathbf{u} = (\mathbf{u}_1, \ldots, \mathbf{u}_m)^T \in O \mapsto (f_1(\mathbf{x}^*, \mathbf{u}), \ldots, f_n(\mathbf{x}^*, \mathbf{u}), \mathbf{u}_{n+1}, \ldots, \mathbf{u}_m)^T \in \mathbb{R}^m \ .$$

The Jacobian matrix of $h_{\mathbf{x}^*}$ taken at the vector $\mathbf{w}^*$ writes

$$Jh_{\mathbf{x}^*}(\mathbf{w}^*) = \begin{pmatrix} \nabla_{\mathbf{w}} f_1(\mathbf{x}^*, \mathbf{w})^T \\ \vdots \\ \nabla_{\mathbf{w}} f_n(\mathbf{x}^*, \mathbf{w})^T \\ E_{n+1} \\ \vdots \\ E_m \end{pmatrix}$$

where $E_i \in \mathbb{R}^m$ is the (line) vector with a 1 at position i and zeros everywhere else. The matrix of the differential of $(D_{\mathbf{w}^*} f(\mathbf{x}^*, \cdot)$ expressed in the canonical basis correspond to the n first lines of the above Jacobian matrix, such that the matrix $(D_{\mathbf{w}^*} f(\mathbf{x}^*, \cdot)(\mathbf{e}_i))_{i \in [1..m]}$ corresponds to the n times n first block. Hence the Jacobian matrix $Jh_{\mathbf{x}^*}(\mathbf{w}^*)$ is invertible. In addition, $h_{\mathbf{x}^*}$ is C^1. Therefore we can apply the inverse function theorem to $h_{\mathbf{x}^*}$: there exists $U_{\mathbf{w}^*} \subset O$ a neighbourhood of $\mathbf{w}^*$ and $V_{h_{\mathbf{x}^*}(\mathbf{w}^*)}$ a neighbourhood of $h_{\mathbf{x}^*}(\mathbf{w}^*)$ such that $h_{\mathbf{x}^*}$ is a bijection from $U_{\mathbf{w}^*}$ to $V_{h_{\mathbf{x}^*}(\mathbf{w}^*)}$. Let π_n denote the projection

$$\pi_n : \mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)^T \in \mathbb{R}^m \mapsto (\mathbf{y}_1, \ldots, \mathbf{y}_n)^T \in \mathbb{R}^n \ .$$

Then $f(\mathbf{x}^*, \mathbf{u}) = \pi_n \circ h_{\mathbf{x}^*}(\mathbf{u})$ for all $\mathbf{u} \in O$, and so $f(\mathbf{x}^*, U_{\mathbf{w}^*}) = \pi_n(V_{h_{\mathbf{x}^*}(\mathbf{w}^*)})$. The set $V_{h_{\mathbf{x}^*}(\mathbf{w}^*)}$ being an open set, so is $V_{f(\mathbf{x}^*, \mathbf{w}^*)} := \pi_n(V_{h_{\mathbf{x}^*}(\mathbf{w}^*)})$ which is therefore an open

neighbourhood of $f(\mathbf{x}^*, \mathbf{w}^*) = \pi_n \circ h_{\mathbf{x}^*}(\mathbf{w}^*)$, that satisfies $V_{f(\mathbf{x}^*,\mathbf{w}^*)}=f(\mathbf{x}^*, U_{\mathbf{w}^*})$, which shows 2.

We are now going to prove the first point of the lemma. Since f is C^1, the coefficients of the Jacobian matrix of $h_{\mathbf{x}^*}$ at $\mathbf{w}^*$ are continuous functions of $\mathbf{x}^*$ and $\mathbf{w}^*$, and as the Jacobian determinant is a polynomial in those coefficients, it is also a continuous function of $\mathbf{x}^*$ and $\mathbf{w}^*$. The Jacobian determinant of $h_{\mathbf{x}^*}$ at $\mathbf{w}^*$ being non-zero (since we have seen when proving the second point above that the Jacobian matrix at $\mathbf{w}^*$ is invertible), the continuity of the Jacobian determinant implies the existence of N an open neighbourhood of $(\mathbf{x}^*, \mathbf{w}^*)$ such that for all $(\mathbf{y}, \mathbf{u}) \in N$, the Jacobian determinant of $h_{\mathbf{y}}$ at $\mathbf{u}$ is non-zero. Since the matrix $(D_{\mathbf{u}} f(\mathbf{y}, .)(\mathbf{e}_i))_{1\leq i\leq m}$ corresponds to the n times n first block of the Jacobian matrix $Jh_{\mathbf{y}}(\mathbf{u})$, it is invertible which shows that $D_{\mathbf{u}} f(\mathbf{y}, .)$ is of rank n which proves that $f(\mathbf{y}, \cdot)$ is a submersion at $\mathbf{u}$ for all $(\mathbf{y}, \mathbf{u}) \in N$, which proves 1.

We may also apply the implicit function theorem to the function $(\mathbf{y}, \mathbf{u}) \in (\mathbb{R}^n \times \mathbb{R}^m) \mapsto h_{\mathbf{y}}(\mathbf{u}) \in \mathbb{R}^m$: there exists g a C^1 function from $\tilde{V}_{\mathbf{x}^*}$ an open neighbourhood of $\mathbf{x}^*$ to $\tilde{U}_{\mathbf{w}^*}$ a open neighbourhood of $\mathbf{w}^*$ such that $h_{\mathbf{y}}(\mathbf{u}) = h_{\mathbf{x}^*}(\mathbf{w}^*) \Leftrightarrow \mathbf{u} = g(\mathbf{y})$ for all $(\mathbf{y}, \mathbf{u}) \in \tilde{V}_{\mathbf{x}^*} \times \tilde{U}_{\mathbf{w}^*}$. Then $f(\mathbf{y}, g(\mathbf{y})) = \pi_n \circ h_{\mathbf{y}}(g(\mathbf{y})) = \pi_n \circ h_{\mathbf{x}^*}(\mathbf{w}^*) = f(\mathbf{x}^*, \mathbf{w}^*)$, proving 3. □

The following lemma is a generalization of [8, Proposition 7.1.4] to our setting.

Lemma 6. *Suppose that F is C^∞ and that the function $(\mathbf{x}, \mathbf{w}) \mapsto p_{\mathbf{x}}(\mathbf{w})$ is lower semi-continuous. Then the control model is forward accessible if and only if for all $\mathbf{x} \in X$ there exists $t \in \mathbb{N}^*$ and $\mathbf{w} \in O_{\mathbf{x},t}$ such that $F^t(\mathbf{x}, \cdot)$ is a submersion at $\mathbf{w}$.*

Proof. Suppose that the control model is forward accessible. Then, for all $\mathbf{x} \in X$, $A_+(\mathbf{x})$ is not Lebesgue negligible. Since $\sum_{i\in\mathbb{N}} \Lambda_n(A^i_+(\mathbf{x})) \geq \Lambda_n(A_+(\mathbf{x})) > 0$, there exists $i \in \mathbb{N}^*$ such that $\Lambda_n(A^i_+(\mathbf{x})) > 0$ ($i \neq 0$ because $A^0_+(\mathbf{x}) = \{\mathbf{x}\}$ is Lebesgue negligible). Suppose that for all $\mathbf{w} \in O_{\mathbf{x},i}$, $\mathbf{w}$ is a critical point for $F^i(\mathbf{x}, \cdot)$, that is the differential of $F^i(\mathbf{x}, \cdot)$ in $\mathbf{w}$ is not surjective. According to Lemma 2 the function F^t is C^∞, so we can apply Sard's theorem [13, Theorem II.3.1] to $F^i(\mathbf{x}, \cdot)$ which implies that the image of the critical points is Lebesgue negligible, hence $F^i(\mathbf{x}, O_{\mathbf{x},t}) = A^i_+(\mathbf{x})$ is Lebesgue negligible. We have a contradiction, so there exists $\mathbf{w} \in O_{\mathbf{x},i}$ for which $F^i(\mathbf{x}, \cdot)$ is a submersion at $\mathbf{w}$.

Suppose now that for all $\mathbf{x} \in X$, there exists $t \in \mathbb{N}^*$ and $\mathbf{w} \in O_{\mathbf{x},t}$ such that $F^t(\mathbf{x}, \cdot)$ is a submersion at $\mathbf{w}$ and let us prove that the control model is forward accessible. Since the function $(\mathbf{x}, \mathbf{w}) \mapsto p_{\mathbf{x}}(\mathbf{w})$ is lower continuous and that F is continuous, according to Lemma 3, then $p_{\mathbf{x},t}$ is lower semi-continuous and hence $O_{\mathbf{x},t}$ is an open set. Then according to Lemma 5, point 2) applied to the function F^t restricted to the open set $X \times O_{\mathbf{x},t}$, there exists $U_{\mathbf{w}} \subset O_{\mathbf{x},t}$ and $V_{F^t(\mathbf{x},\mathbf{w})}$ non-empty open sets such that $F^t(\mathbf{x}, U_{\mathbf{w}}) \supset V_{F^t(\mathbf{x},\mathbf{w})}$. Since $A_+(\mathbf{x}) \supset F^t(\mathbf{x}, O_{\mathbf{x},t}) \supset F^t(\mathbf{x}, U_{\mathbf{w}})$, $A_+(\mathbf{x})$ has non-empty interior for all $\mathbf{x} \in X$, meaning the control model is forward accessible. □

The following lemma treats of the preservation of Lebesgue null sets by a locally Lipschitz continuous function on spaces of equal dimension.

Lemma 7. *(From [7, Corollary 5.9]) Take U an open set of $\mathbb{R}^n$ and $f : U \to \mathbb{R}^n$ a locally Liptschiz-continuous function. Take $A \subset U$ a set of zero Lebesgue measure. Then its image $f(A)$ is also of zero Lebesgue measure.*

Lemma 7 requires the dimensions of the domain and codomain to be equal. When the dimension of the domain is lower or equal than the dimension of the codomain, a generalization of Lemma 7 is presented in [11] for the preimage of sets via submersions. The authors of [11] investigate the so-called 0-property: a continuous function $f : Z \subset \mathbb{R}^m \to X \subset \mathbb{R}^n$ has the 0-property if the preimage of any set of Lebesgue measure 0 has Lebesgue measure 0. They show in [11, Theorem 2 and Theorem 3] that if f is a continuous function and that for almost all $\mathbf{z} \in Z$ it is a submersion at $\mathbf{z}$, then is has the 0-property. They also show in [11, Theorem 1] that for f a C^r function with $r \geq m-n+1$ (this inequality coming from Sard's theorem [13, Theorem II.3.1]), then the 0-property is equivalent to f being a submersion at $\mathbf{z}$ for almost all $\mathbf{z} \in Z$. In the following lemma, we establish conditions for a function f to have a stronger form of 0-property, for which the preimage of a set has Lebesgue measure 0 if and only if the set has measure 0.

Lemma 8. *Let $g : Z \subset \mathbb{R}^m \to X \subset \mathbb{R}^n$ be a C^1 function where Z and X are open sets. Let $A \in \mathcal{B}(X)$ and let us assume that for almost all $\mathbf{z} \in g^{-1}(A)$, g is a submersion at $\mathbf{z}$, i.e. the differential of g at $\mathbf{z}$ is surjective (which implies that $m \geq n$).*

Then (i) $\Lambda_n(A) = 0$ implies that $\Lambda_m(g^{-1}(A)) = 0$, and (ii) if $A \subset g(Z)$ and if g is a submersion at $\mathbf{z}$ for all $\mathbf{z} \in g^{-1}(A)$, then $\Lambda_n(A) = 0$ if and only if $\Lambda_m(g^{-1}(A)) = 0$.

Proof. This first part of the proof is similar to the proof of Lemma 5. Let $N \in \mathcal{B}(Z)$ be a Λ_m-negligible set such that g is a submersion at all points of $g^{-1}(A)\backslash N$, and take $\mathbf{z} \in g^{-1}(A)\backslash N$ and $(\mathbf{e}_i)_{i\in[1..m]}$ the canonical basis of $\mathbb{R}^m$. For $\mathbf{y} \in \mathbb{R}^m$, we denote $\mathbf{y} = (\mathbf{y}_1, \ldots, \mathbf{y}_m)^T$ its expression in the canonical basis. In the canonical basis of $\mathbb{R}^n$ we denote $g(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_n(\mathbf{x}))^T$. Since g is a submersion at $\mathbf{z}$, $D_{\mathbf{z}}g$ the differential of g at $\mathbf{z}$ has rank n so there exists a permutation $\sigma_{\mathbf{z}} : [1..m] \to [1..m]$ such that the matrix formed by the vectors $(D_{\mathbf{z}}g(\mathbf{e}_{\sigma(i)}))_{i\in[1..n]}$ has rank n. We assume that this permutation is the identity (otherwise we consider a reordering of the canonical basis via σ). Let

$$h_{\mathbf{z}} : \mathbf{y} \in \mathbb{R}^m \mapsto (g_1(\mathbf{y}), \ldots, g_n(\mathbf{y}), \mathbf{y}_{n+1}, \ldots, \mathbf{y}_m)^T$$

Similarly as in the proof of Lemma 5, by expressing the differential of $h_{\mathbf{z}}$ in the basis $(\mathbf{e}_i)_{i\in[1..m]}$ we can see that the Jacobian determinant of $h_{\mathbf{z}}$ equals to the determinant of the matrix composed of the vectors $(D_{\mathbf{z}}g(\mathbf{e}_i))_{i\in[1..n]}$, which is non-zero, multiplied by the determinant of the identity matrix, which is one. Hence the Jacobian determinant of $h_{\mathbf{z}}$ is non-zero, and so we can apply the inverse function theorem to $h_{\mathbf{z}}$ (which inherits the C^1 property from g). We hence obtain that there exists $U_{\mathbf{z}}$ an open neighbourhood of $\mathbf{z}$, $V_{h_{\mathbf{z}}(\mathbf{z})}$ an open neighbourhood of $h_{\mathbf{z}}(\mathbf{z})$ such that the function $h_{\mathbf{z}}$ is a diffeomorphism from $U_{\mathbf{z}}$ to $V_{h_{\mathbf{z}}(\mathbf{z})}$. Then, denoting π_n the projection

$$\pi_n : \mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_m)^T \in \mathbb{R}^m \mapsto (\mathbf{z}_1, \ldots, \mathbf{z}_n)^T \quad ,$$

we have $g(\mathbf{u}) = \pi_n \circ h_{\mathbf{z}}(\mathbf{u})$ for all $\mathbf{u} \in Z$.

Then $g^{-1}(A) \cap U_{\mathbf{z}} = h_{\mathbf{z}}^{-1} \circ \pi_n^{-1}(A) \cap h_{\mathbf{z}}^{-1}(V_{h_{\mathbf{z}}(\mathbf{z})}) = h_{\mathbf{z}}^{-1}(A \times \mathbb{R}^{m-n} \cap V_{h_{\mathbf{z}}(\mathbf{z})})$. Since $h_{\mathbf{z}}$ is a diffeomorphism from $U_{\mathbf{z}}$ to $V_{h_{\mathbf{z}}(\mathbf{z})}$, $h_{\mathbf{z}}$ and $h_{\mathbf{z}}^{-1}$ are locally Lipschitz continuous. So we can use Lemma 7 with $h_{\mathbf{z}}^{-1}$ and its contrapositive with $h_{\mathbf{z}}$ and obtain that $\Lambda_m(A \times \mathbb{R}^{m-n} \cap V_{h_{\mathbf{z}}(\mathbf{z})}) = 0$ if and only if $\Lambda_m(h_{\mathbf{z}}^{-1}(A \times \mathbb{R}^{m-n} \cap V_{h_{\mathbf{z}}(\mathbf{z})})) = 0$, which implies that

$$\Lambda_m(A \times \mathbb{R}^{m-n} \cap V_{h_{\mathbf{z}}(\mathbf{z})}) = 0 \text{ if and only if } \Lambda_m(g^{-1}(A) \cap U_{\mathbf{z}}) = 0 \ . \tag{20}$$

If $\Lambda_n(A) = 0$ then $\Lambda_m(A \times \mathbb{R}^{m-n}) = 0$ and thus $\Lambda_m(A \times \mathbb{R}^{m-n} \cap V_{h_{\mathbf{z}}(\mathbf{z})}) = 0$ which in turns implies with (20) that $\Lambda_m(g^{-1}(A) \cap U_{\mathbf{z}}) = 0$. This latter statement holds for all $\mathbf{z} \in g^{-1}(A) \backslash N$, which with Lemma 1 implies that $\Lambda_m(g^{-1}(A) \backslash N) = 0$, and since N is a Lebesgue negligible set $\Lambda_m(g^{-1}(A)) = 0$. We have then proven the statement (i) of the lemma.

We will now prove the second statement. Suppose that $\Lambda_n(A) > 0$, so there exists $\mathbf{x} \in A$ such that for all $\epsilon > 0$, $\Lambda_n(B(\mathbf{x}, \epsilon) \cap A) > 0$ (this is implied by the contrapositive of Lemma 1). Assume that $A \subset g(Z)$, i.e. g is surjective on A, then there exists $\mathbf{z} \in Z$ such that $g(\mathbf{z}) = \mathbf{x}$. Since in the second statement we suppose that g is a submersion at $\mathbf{u}$ for all $\mathbf{u} \in g^{-1}(A)$, we have that g is a submersion at $\mathbf{z}$, and so $h_{\mathbf{z}}$ is a diffeomorphism from $U_{\mathbf{z}}$ to $V_{h_{\mathbf{z}}(\mathbf{z})}$ and (20) holds. Since $V_{h_{\mathbf{z}}(\mathbf{z})}$ is an open neighbourhood of $h_{\mathbf{z}}(\mathbf{z}) = (g(\mathbf{z}), \mathbf{z}_{n+1}, \ldots, \mathbf{z}_m)$, there exists (r_1, r_2) such that $B(g(\mathbf{z}), r_1) \times B((\mathbf{z}_i)_{i \in [n+1..m]}, r_2) \subset V_{h_{\mathbf{z}}(\mathbf{z})}$. Since $\Lambda_m(A \times \mathbb{R}^{m-n} \cap B(\mathbf{x}, r_1) \times B((\mathbf{z}_i)_{i \in [n+1..m]}, r_2)) = \Lambda_m((A \cap B(\mathbf{x}, r_1)) \times B((\mathbf{z}_i)_{i \in [n+1..m]}, r_2)) > 0$, we have $\Lambda_m(A \times \mathbb{R}^{m-n} \cap V_{h_{\mathbf{z}}(\mathbf{z})}) > 0$. This in turn implies through (20) that $\Lambda_m(g^{-1}(A) \cap U_{\mathbf{z}}) > 0$ and thus $\Lambda_m(g^{-1}(A)) > 0$. We have thus proven that if $\Lambda_n(A) > 0$ then $\Lambda_m(g^{-1}(A)) > 0$, which proves the lemma. □

3. Main Results

We present here our main result. Its proof will be established in the following subsections.

Theorem 1. *Let $\Phi = (\Phi_t)_{t \in \mathbb{N}}$ be a time-homogeneous Markov chain on an open state space $X \subset \mathbb{R}^n$, defined via*

$$\Phi_{t+1} = F(\Phi_t, \alpha(\Phi_t, \mathbf{U}_{t+1})) \tag{21}$$

where $(\mathbf{U}_t)_{t \in \mathbb{N}^*}$ is a sequence of i.i.d. random vectors in $\mathbb{R}^p$, $\alpha : X \times \mathbb{R}^p \to O$ and $F : X \times O \to X$ are two measurable functions with O an open subset of $\mathbb{R}^m$. For all $\mathbf{x} \in X$, we assume that $\alpha(\mathbf{x}, \mathbf{U}_1)$ admits a probability density function that we denote $\mathbf{w} \in O \mapsto p_{\mathbf{x}}(\mathbf{w})$. We define the function $F^t : X \times O^t \to X$ via (7), the probability density function $p_{\mathbf{x},t}$ via (9), and the sets $O_{\mathbf{x}}$ and $O_{\mathbf{x},t}$ via (6) and (11). For $B \in \mathcal{B}(X)$, we denote μ_B the trace measure $A \in \mathcal{B}(X) \mapsto \Lambda_n(A \cap B)$, where Λ_n denotes the Lebesgue measure on $\mathbb{R}^n$. Suppose that

1. *the function $(\mathbf{x}, \mathbf{w}) \in (X \times O) \mapsto F(\mathbf{x}, \mathbf{w})$ is C^1,*
2. *the function $(\mathbf{x}, \mathbf{w}) \in (X \times O) \mapsto p_{\mathbf{x}}(\mathbf{w})$ is lower semi-continuous,*

3. *there exists* $\mathbf{x}^* \in X$ *a strongly globally attracting state,* $k \in \mathbb{N}^*$ *and* $\mathbf{w}^* \in O_{\mathbf{x}^*,k}$ *such that the function* $\mathbf{w} \in O^k \mapsto F^k(\mathbf{x}^*, \mathbf{w})$ *is a submersion at* $\mathbf{w}^*$.

Then there exists B_0 *a non-empty open subset of* $A_+^k(\mathbf{x}^*)$ *containing* $F^k(\mathbf{x}^*, \mathbf{w}^*)$ *such that* Φ *is a* μ_{B_0}*-irreducible aperiodic T-chain, and compacts sets of* X *are small sets.*

Before to provide the proof of this theorem, we discuss its assumptions with respect to the chapter 7 of the Meyn and Tweedie book. Results similar to Theorem 1 are presented in [8, Chapter 7]. The underlying assumptions there translate to our setting as (i) the function $p(\mathbf{x}, \mathbf{w})$ is independent of $\mathbf{x}$, that is $(\mathbf{x}, \mathbf{w}) \mapsto p(\mathbf{x}, \mathbf{w}) = p(\mathbf{w})$, (ii) $\mathbf{w} \mapsto p(\mathbf{w})$ is lower semi-continuous, F is C^∞. In contrast, in our context we do not need $p(\mathbf{x}, \mathbf{w})$ to be independent of $\mathbf{x}$, we need the function $(\mathbf{x}, \mathbf{w}) \mapsto p_{\mathbf{x}}(\mathbf{w})$ to be lower semi-continuous, and we need F to be C^1 rather than C^∞. In [8], assuming (i) and (ii) and the forward accessibility of the control model, the Markov chain is proved to be a T-chain [8, Proposition 7.1.5]; this property is then used to prove that the existence of a globally attracting state is equivalent to the φ-irreducibility of the Markov chain [8, Proposition 7.2.5 and Theorem 7.2.6]. The T-chain property is a strong property and in our context, we prove in Proposition 3 that if Φ is a T-chain, then we also get the equivalence between φ-irreducibility and the existence of a globally attracting state. We develop another approach in Lemma 9, relying on the submersion property of point 3) of Theorem 1 rather than on the T-chain property. This approach is used in Theorem 2 to prove that the existence of a globally attracting state $\mathbf{x}^* \in X$ for which there exists $k \in \mathbb{N}^*$ and $\mathbf{w}^* \in O_{\mathbf{x}^*,k}$ such that $F^k(\mathbf{x}^*, \cdot)$ is a submersion at $\mathbf{w}^*$ implies the φ-irreducibility of the Markov chain. The approach developed in Lemma 9 allows for a finer control of the transition kernel than with the T-chain property, which is then used to get aperiodicity in Theorem 3 by assuming the existence of a strongly attracting state on which the submersion property of 3) of Theorem 1 holds. In the applications of Section 4, the existence of a strongly attracting state is immediately derived from the proof of the existence of a globally attracting state. In contrast in [8, Theorem 7.3.5], assuming (i), (ii), the forward accessibility of the control model, the existence of a globally attracting state $\mathbf{x}^*$ and the connexity of $O_{\mathbf{x}}$, aperiodicity is proven to be equivalent to the connexity of $\overline{A_+(\mathbf{x}^*)}$.

Proof. (of Theorem 1) From Theorem 3, there exists B_0 a non-empty open subset of $A_+^k(\mathbf{x}^*)$ containing $F^k(\mathbf{x}^*, \mathbf{w}^*)$ such that Φ is a μ_{B_0}-irreducible aperiodic chain. With Proposition 5 the chain is also weak Feller. Since B_0 is a non-empty open set supp μ_{B_0} has non empty interior, so from [8, Theorem 6.0.1] with (iii) Φ is a μ_{B_0}-irreducible T-chain and with (ii) compact sets are petite sets. Finally, since the chain is μ_{B_0}-irreducible and aperiodic, with [8, Theorem 5.5.7] petite sets are small sets. □

Assuming that F is C^∞ we showed in Lemma 6 that the forward accessibility of the control model is equivalent to assuming that for all $\mathbf{x} \in X$ there exists $t \in \mathbb{N}^*$ and $\mathbf{w} \in O_{\mathbf{x},t}$ such that $F^t(\mathbf{x}, \cdot)$ is a submersion at $\mathbf{w}$, which satisfies a part of condition 3. of Theorem 1. Hence, we can use Lemma 6 and Theorem 1 to derive Corollary 1.

Corollary 1. *Suppose that*

1. *the function* $(\mathbf{x}, \mathbf{w}) \mapsto F(\mathbf{x}, \mathbf{w})$ *is* C^∞,
2. *the function* $(\mathbf{x}, \mathbf{w}) \mapsto p_{\mathbf{x}}(\mathbf{w})$ *is lower semi-continuous,*
3. *the control model CM(F) is forward accessible,*
4. *there exists* $\mathbf{x}^*$ *a strongly globally attracting state.*

Then there exists B_0 *a non-empty open subset of* $A_+^k(\mathbf{x}^*)$ *containing* $F^k(\mathbf{x}^*, \mathbf{w}^*)$ *such that* Φ *is a* μ_{B_0}*-irreducible aperiodic T-chain, and compacts sets of* X *are small sets.*

Proof. From Lemma 6, the second part of the assumption 3. of Theorem 1 is satisfied such that the conclusions of Theorem 1 hold. □

3.1. φ-Irreducibility

When (i) the function $(\mathbf{x}, \mathbf{w}) \mapsto p(\mathbf{x}, \mathbf{w})$ is independent of $\mathbf{x}$, that is $p(\mathbf{x}, \mathbf{w}) = p(\mathbf{w})$, (ii) the function $\mathbf{w} \mapsto p(\mathbf{w})$ for all $\mathbf{x} \in X$ is lower semi-continuous, (iii) F is C^∞ and (iv) the control model is forward accessible, it is shown in [8, Proposition 7.1.5] that Φ is a T-chain. This is a strong property that is then used to show the equivalence of the existence of a globally attracting state and the φ-irreducibility of the Markov chain Φ in [8, Theorem 7.2.6]. In our context where the function $(\mathbf{x}, \mathbf{w}) \mapsto p(\mathbf{x}, \mathbf{w})$ varies with $\mathbf{x}$, the following proposition shows that the equivalence still holds assuming that the Markov chain Φ is a T-chain.

Proposition 3. *Suppose that*

1. *the Markov chain* Φ *is a* T*-chain,*
2. *the function* F *is continuous,*
3. *the function* $(\mathbf{x}, \mathbf{w}) \mapsto p_{\mathbf{x}}(\mathbf{w})$ *is lower semi-continuous*

Then the Markov chain Φ *is* φ*-irreducible if and only if there exists* $\mathbf{x}^*$ *a globally attracting state.*

Proof. Suppose that there exists $\mathbf{x}^*$ a globally attracting state. Since Φ is a T-chain, there exists a a sampling distribution such that K_a possesses a continuous component T such that $T(\mathbf{x}, X) > 0$ for all $\mathbf{x} \in X$.

Take $A \in \mathcal{B}(X)$ such that $T(\mathbf{x}^*, A) > 0$ (such a A always exists because we can for instance take $A = X$). The function $T(\cdot, A)$ being lower semi-continuous, there exists $\delta > 0$ and $r > 0$ such that for all $\mathbf{y} \in B(\mathbf{x}^*, r)$, $T(\mathbf{y}, A) > \delta$, hence $B(\mathbf{x}^*, r) \overset{a}{\rightsquigarrow} A$. Since $\mathbf{x}^*$ is a globally attracting state, for all $\mathbf{y} \in X$, $\mathbf{x}^* \in \overline{\bigcup_{k\in\mathbb{N}^*} A_+^k(\mathbf{y})}$ so there exists points of $\bigcup_{k\in\mathbb{N}^*} A_+^k$ arbitrarily close to $\mathbf{x}^*$. Hence there exists $t_{\mathbf{y}}$ and $\mathbf{w} \in O_{\mathbf{y},t_{\mathbf{y}}}$ such that $F^{t_{\mathbf{y}}}(\mathbf{y}, \mathbf{w}) \in B(\mathbf{x}^*, r)$. Furthermore, since $O_{\mathbf{y},t_{\mathbf{y}}}$ is an open set (by the lower semi-continuity of $p_{\mathbf{x},t_{\mathbf{y}}}(\cdot)$ which in turn is implied by the lower semi-continuity of the function $(\mathbf{x}, \mathbf{w}) \mapsto p_{\mathbf{x}}(\mathbf{w})$, the continuity of F with Lemma 3) and $F^{t_{\mathbf{y}}}(\mathbf{y}, \cdot)$ is continuous (as implied by the continuity of F with Lemma 2) the set $E := \{\mathbf{u} \in O_{\mathbf{y},t_{\mathbf{y}}} | F^{t_{\mathbf{y}}}(\mathbf{y}, \mathbf{u}) \in B(\mathbf{x}^*, r)\}$

is an open set, and as $\mathbf{w} \in E$ it is non empty. Since $P^{t_\mathbf{y}}(\mathbf{y}, B(\mathbf{x}^*, r)) = \int_E p_{\mathbf{y},t_\mathbf{y}}(\mathbf{u})d\mathbf{u}$ and that $p_{\mathbf{y},t_\mathbf{y}}(\mathbf{u}) > 0$ for all $\mathbf{u} \in E \subset O_{\mathbf{y},t_\mathbf{y}}$, $P^{t_\mathbf{y}}(\mathbf{y}, B(\mathbf{x}^*, r)) > 0$ as the integral of a positive function over a set of positive Lebesgue measure is positive. Hence $K_{a_\epsilon}(\mathbf{y}, B(\mathbf{x}^*, r)) > 0$ (where K_{a_ϵ} is the transition kernel defined in (18) with the geometric distribution (19)), and so $\{\mathbf{y}\} \overset{a_\epsilon}{\rightsquigarrow} B(\mathbf{x}^*, r)$. Hence with [8, Lemma 5.5.2] $\{\mathbf{y}\} \overset{a*a_\epsilon}{\rightsquigarrow} A$ which implies that for some $t \in \mathbb{N}^*$, $P^t(\mathbf{y}, A) > 0$. Therefore, $T(\mathbf{x}^*, A) > 0$ implies that $\sum_{t\in\mathbb{N}^*} P^t(\mathbf{y}, A) > 0$ for all $\mathbf{y} \in X$. And since $T(\mathbf{x}^*, X) > 0$, $T(\mathbf{x}^*, \cdot)$ is not a trivial measure, so the Markov chain Φ is $T(\mathbf{x}^*, \cdot)$-irreducible.

Suppose that Φ is φ-irreducible, then φ is non-trivial and according to Proposition 4 any point of $\operatorname{supp}\varphi$ is a globally attracting state, so there exists a globally attracting state. □

Although the T-chain property allows for a simple proof of the equivalence between the existence of a globally attracting state and the φ-irreducibility of the Markov chain. The T-chain property is not needed for Theorem 2, which instead relies on the following lemma. Interestingly, not using the T-chain in the lemma allows some control on the transition kernel, which is then used for Theorem 3 for aperiodicity.

Lemma 9. *Let $A \in \mathcal{B}(X)$ and suppose that*

1. *the function F is C^1,*
2. *the function $(\mathbf{x}, \mathbf{w}) \mapsto p_\mathbf{x}(\mathbf{w})$ is lower semi-continuous,*
3. *there exists $\mathbf{x}^* \in X$, $k \in \mathbb{N}^*$ and $\mathbf{w}^* \in O_{\mathbf{x}^*,k}$ such that $F^k(\mathbf{x}^*, \cdot)$ is a submersion at $\mathbf{w}^*$.*

Then there exists $B_0 \subset A^k_+(\mathbf{x}^*)$ a non-empty open set containing $F^k(\mathbf{x}^*, \mathbf{w}^*)$ and such that for all $\mathbf{z} \in B_0$, there exists $U_{\mathbf{x}^*}$ an open neighbourhood of $\mathbf{x}^*$ that depends on $\mathbf{z}$ and having the following property: for $\mathbf{y} \in X$ if there exists a t-steps path from $\mathbf{y}$ to $U_{\mathbf{x}^*}$, then let $A \in \mathcal{B}(X)$

$$P^{t+k}(\mathbf{y}, A) = 0 \Rightarrow \exists\, V_\mathbf{z} \text{ an open neighbourhood of } \mathbf{z} \text{ such that } \Lambda_n(V_\mathbf{z} \cap A) = 0 \quad (22)$$

or equivalently,

$$\text{for all } V_\mathbf{z} \text{ open neighbourhood of } \mathbf{z}, \Lambda_n(V_\mathbf{z} \cap A) > 0 \Rightarrow P^{t+k}(\mathbf{y}, A) > 0\ . \quad (23)$$

Proof. (i) We will need through this proof a set $N = N_1 \times N_2$ which is an open neighbourhood of $(\mathbf{x}^*, \mathbf{w}^*)$, such that for all $(\mathbf{x}, \mathbf{w}) \in N$ we have $p_{\mathbf{x},k}(\mathbf{w}) > 0$ and that $F^k(\mathbf{x}, \cdot)$ is a submersion at $\mathbf{w}$. To obtain N, first let us note that since F is C^1, according to Lemma 7 so is F^t for all $t \in \mathbb{N}^*$; and since the function $(\mathbf{x}, \mathbf{w}) \mapsto p_\mathbf{x}(\mathbf{w})$ is lower semi-continuous, according to Lemma 3 so is the function $(\mathbf{x}, \mathbf{w}) \mapsto p_{\mathbf{x},t}(\mathbf{w})$ for all $t \in \mathbb{N}^*$. Hence the set $\{(\mathbf{x}, \mathbf{w}) \in X \times O^t | p_{\mathbf{x},k}(\mathbf{w}) > 0\}$ is an open set, and since $\mathbf{w}^* \in O_{\mathbf{x}^*,k}$, there exists $M_1 \times M_2$ a neighbourhood of $(\mathbf{x}^*, \mathbf{w}^*)$ such that for all $(\mathbf{x}, \mathbf{w}) \in M_1 \times M_2$, $p_{\mathbf{x},k}(\mathbf{w}) > 0$. Furthermore, according to point 1. of Lemma 5, there exists $\tilde{M} = \tilde{M}_1 \times \tilde{M}_2$ an open neighbourhood of $(\mathbf{x}^*, \mathbf{w}^*)$ such that for all $(\mathbf{x}, \mathbf{w}) \in \tilde{M}_1 \times \tilde{M}_2$, $F^k(\mathbf{x}, \cdot)$ is a submersion at $\mathbf{w}$. Then the set $N := M \cap \tilde{M}$ has the desired property.

(ii) We now prove that for all $\mathbf{y} \in X$, U any open neighbourhood of $\mathbf{x}^*$ and $A \in \mathcal{B}(X)$, if there exists $\mathbf{v}$ a t-steps path from $\mathbf{y}$ to U and if $P^{t+k}(\mathbf{y}, A) = 0$ then there exists $\mathbf{x}_0 \in U$ such that $P^k(\mathbf{x}_0, A) = 0$. Indeed, U being open containing $F^t(\mathbf{y}, \mathbf{v})$ there exists $\epsilon > 0$ such that $B(F^t(\mathbf{y}, \mathbf{v}), \epsilon) \subset U$, and by continuity of $F^t(\mathbf{y}, \cdot)$, there exists $\eta > 0$ such that $F^t(\mathbf{y}, B(\mathbf{v}, \eta)) \subset B(F^t(\mathbf{y}, \mathbf{v}), \epsilon) \subset U$; furthermore, $P^{t+k}(\mathbf{y}, A) = 0$ implies that

$$P^{t+k}(\mathbf{y}, A) = \int_{O_{\mathbf{y},t}} p_{\mathbf{y},t}(\mathbf{u}) P^k(F^t(\mathbf{y}, \mathbf{u}), A) d\mathbf{u} = 0 \ .$$

Since for all $\mathbf{u} \in O_{\mathbf{y},t}$, $p_{\mathbf{y},t}(\mathbf{u}) > 0$, this implies that for almost all $\mathbf{u} \in O_{\mathbf{y},t}$, $P^k(F^t(\mathbf{y}, \mathbf{u}), A) = 0$. Since $\mathbf{v} \in O_{\mathbf{y},t}$, the set $O_{\mathbf{y},t} \cap B(\mathbf{v}, \eta)$ is a non-empty open set and therefore has positive Lebesgue measure; so there exists $\mathbf{u}_0 \in O_{\mathbf{y},t} \cap B(\mathbf{v}, \eta)$ such that $P^k(F^t(\mathbf{y}, \mathbf{u}_0), A) = 0$. Let $\mathbf{x}_0$ denote $F^t(\mathbf{y}, \mathbf{u}_0)$. By choice of η, we also have $\mathbf{x}_0 \in F^t(\mathbf{y}, B(\mathbf{v}, \eta)) \subset U$.

(iii) Now let us construct the set B_0 mentioned in the lemma. We consider the function F^k restricted to $X \times N_2$. According to assumption 3. and (i), we have $\mathbf{x}^* \in X$ and $\mathbf{w}^*$ in N_2 such that $F^k(\mathbf{x}^*, .)$ is a submersion at $\mathbf{w}^*$. Hence using point 2. of Lemma 5 on the function F^k restricted to $X \times N_2$, we obtain that there exists $V_{\mathbf{w}^*} \subset N_2$ an open neighbourhood of $\mathbf{w}^*$ and $U_{F^k(\mathbf{x}^*, \mathbf{w}^*)}$ an open neighbourhood of $F^k(\mathbf{x}^*, \mathbf{w}^*)$ such that $U_{F^k(\mathbf{x}^*, \mathbf{w}^*)} \subset F^k(\mathbf{x}^*, V_{\mathbf{w}^*})$. We take $B_0 = U_{F^k(\mathbf{x}^*, \mathbf{w}^*)}$ and will prove in what follows that it satisfies the properties announced. Note that since $B_0 \subset F^k(\mathbf{x}^*, V_{\mathbf{w}^*})$, that $V_{\mathbf{w}^*} \subset N_2$ and that $\mathbf{x}^* \in N_1$, $V_{\mathbf{x}^*} \subset O_{\mathbf{x}^*,k}$ and so $B_0 \subset A^k_+(\mathbf{x}^*)$.

(iv) Now, for $\mathbf{z} \in B_0$, let us construct the set $U_{\mathbf{x}^*}$ mentioned in the lemma. We will make it so that there exists a C^1 function g valued in O and defined on a set containing $U_{\mathbf{x}^*}$, such that $F^k(\mathbf{x}, g(\mathbf{x})) = \mathbf{z}$ for all $\mathbf{x} \in U_{\mathbf{x}^*}$. First, since $\mathbf{z} \in B_0$ and $B_0 = U_{F^k(\mathbf{x}^*, \mathbf{w}^*)} \subset F^k(\mathbf{x}^*, V_{\mathbf{w}^*})$, there exists $\mathbf{w}_\mathbf{z} \in V_{\mathbf{w}^*}$ such that $F^k(\mathbf{x}^*, \mathbf{w}_\mathbf{z}) = \mathbf{z}$. Since $V_{\mathbf{w}^*} \subset N_2$, the function $F^k(\mathbf{x}^*, \cdot)$ is a submersion at $\mathbf{w}_\mathbf{z}$, so we can apply point 3. of Lemma 5 to the function F^k restricted to $X \times N_2$: there exists g a C^1 function from $\tilde{U}^g_{\mathbf{x}^*}$ an open neighbourhood of $\mathbf{x}^*$ to $\tilde{V}^g_{\mathbf{w}_\mathbf{z}} \subset N_2$ an open neighbourhood of $\mathbf{w}_\mathbf{z}$ such that for all $\mathbf{x} \in \tilde{U}^g_{\mathbf{x}^*}$, $F^k(\mathbf{x}, g(\mathbf{x})) = F^k(\mathbf{x}^*, \mathbf{w}_\mathbf{z}) = \mathbf{z}$. We now take $U_{\mathbf{x}^*} := \tilde{U}^g_{\mathbf{x}^*} \cap N_1$; it is an open neighbourhood of $\mathbf{x}^*$ and for all $\mathbf{x} \in U_{\mathbf{x}^*}$, $F^k(\mathbf{x}, g(\mathbf{x})) = \mathbf{z}$.

(v) We now construct the set $V_\mathbf{z}$. For $\mathbf{y} \in X$, if there exists a t-steps path from $\mathbf{y}$ to $U_{\mathbf{x}^*}$ and that $P^{t+k}(\mathbf{y}, A) = 0$, then we showed in (ii) that there exists $\mathbf{x}_0 \in U_{\mathbf{x}^*}$ such that $P^k(\mathbf{x}_0, A) = 0$. Since $\mathbf{x}_0 \in U_{\mathbf{x}^*} \subset \tilde{U}^g_{\mathbf{x}^*} \cap N_1$ and that $g(\mathbf{x}_0) \in \tilde{V}^g_{\mathbf{w}_\mathbf{z}} \subset N_2$, the function $F^k(\mathbf{x}_0, \cdot)$ is a submersion at $g(\mathbf{x}_0)$. Therefore, we can apply point 2) of Lemma 5 to F restricted to $X \times N_2$, and so there exists $U_{g(\mathbf{x}_0)} \subset N_2$ an open neighbourhood of $g(\mathbf{x}_0)$ and $V_\mathbf{z}$ an open neighbourhood of $F^k(\mathbf{x}_0, g(\mathbf{x}_0)) = \mathbf{z}$ such that $V_\mathbf{z} \subset F^k(\mathbf{x}_0, U_{g(\mathbf{x}_0)})$.

(vi) Finally we will show that $\Lambda_n(V_\mathbf{z} \cap A) = 0$. Let $\tilde{B} := \{\mathbf{w} \in U_{g(\mathbf{x}_0)} | F^k(\mathbf{x}_0, \mathbf{w}) \in V_\mathbf{z} \cap A\}$. Then

$$P^k(\mathbf{x}_0, A) = \int_{O_{\mathbf{x}_0,k}} \mathbf{1}_A(F^k(\mathbf{x}_0, \mathbf{w})) p_{\mathbf{x}_0,k}(\mathbf{w}) d\mathbf{w} \geq \int_{\tilde{B}} p_{\mathbf{x}_0,k}(\mathbf{w}) d\mathbf{w} \ ,$$

so $\int_{\tilde{B}} p_{\mathbf{x}_0,k}(\mathbf{w}) d\mathbf{w} = 0$. As $\mathbf{x}_0 \in U_{\mathbf{x}^*} \subset N_1$ and $\tilde{B} \subset U_{g(\mathbf{x}_0)} \subset N_2$, $p_{\mathbf{x}_0,k}(\mathbf{w}) > 0$ for all $\mathbf{w} \in \tilde{B}$, which implies with the fact that $\int_{\tilde{B}} p_{\mathbf{x}_0,k}(\mathbf{w}) d\mathbf{w} = 0$ that $\tilde{B}$ is Lebesgue

negligible. Now let h denote the function $F^k(\mathbf{x}_0, \cdot)$ restricted to $U_{g(\mathbf{x}_0)}$. The function h is a C^1 function and $V_{\mathbf{z}}$ is included into the image of h. Both $U_{g(\mathbf{x}_0)}$ to $V_{\mathbf{z}}$ are open sets. Furthermore $\mathbf{x}_0 \in N_1$ and for all $\mathbf{u} \in h^{-1}(V_{\mathbf{z}})$ since $h^{-1}(V_{\mathbf{z}}) \subset U_{g(\mathbf{x}_0)} \subset N_2$ we have $\mathbf{u} \in N_2$ so the function h is a submersion at $\mathbf{u}$. Therefore we can apply Lemma 8 to h, and so if $\Lambda_m(h^{-1}(V_{\mathbf{z}} \cap A)) = 0$ then $\Lambda_n(V_{\mathbf{z}} \cap A) = 0$. Since $h^{-1}(V_{\mathbf{z}}) = U_{g(\mathbf{x}_0)}$, we have $h^{-1}(V_{\mathbf{z}} \cap A) = \tilde{B}$, so we do have $\Lambda_m(h^{-1}(V_{\mathbf{z}} \cap A)) = 0$ which implies $\Lambda_n(V_{\mathbf{z}} \cap A) = 0$.

(vii) The equivalent formulation between (22) and (23) is simply obtained by taking the contrapositive.

□

If the function F is C^∞, then the condition of Lemma 9 on the differential of $F(\mathbf{x}^*, \cdot)$ can be relaxed by asking the control model to be forward accessible using Lemma 6. If the point $\mathbf{x}^*$ used in Lemma 9 is a globally attracting state it follows from Lemma 9 that the chain Φ is irreducible, as stated in the following theorem.

Theorem 2. *Suppose that F is C^1, the function $(\mathbf{x}, \mathbf{w}) \mapsto p_{\mathbf{x}}(\mathbf{w})$ is lower semi-continuous and there exists a globally attracting state $\mathbf{x}^* \in X$, $k \in \mathbb{N}^*$ and $\mathbf{w}^* \in O_{\mathbf{x}^*,k}$ such that the function $\mathbf{w} \in \mathbb{R}^{mk} \mapsto F^k(\mathbf{x}^*, \mathbf{w}) \in \mathbb{R}^n$ is a submersion at $\mathbf{w}^*$. Then Φ is a μ_{B_0}-irreducible Markov chain, where B_0 is a non empty open subset of $A_+^k(\mathbf{x}^*)$ containing $F^k(\mathbf{x}^*, \mathbf{w}^*)$.*

Furthermore if F is C^∞, the function $(\mathbf{x}, \mathbf{w}) \mapsto p_{\mathbf{x}}(\mathbf{w})$ lower semi-continuous, and the control model is forward accessible, then the existence of a globally attracting state is equivalent to the φ-irreducibility of the Markov chain Φ.

Proof. We want to show that for φ a non-trivial measure, Φ is φ-irreducible; i.e. for any $A \in \mathcal{B}(X)$, we need to prove that $\varphi(A) > 0$ implies that $\sum_{t \in \mathbb{N}^*} P^t(\mathbf{x}, A) > 0$ for all $\mathbf{x} \in X$.

According to Lemma 9 there exists a non-empty open set $B_0 \subset A_+^k(\mathbf{x}^*)$ containing $F^k(\mathbf{x}^*, \mathbf{w}^*)$, such that for all $\mathbf{z} \in B_0$ there exists $U_{\mathbf{x}^*}$ a neighbourhood of $\mathbf{x}^*$ that depends on $\mathbf{z}$ having the following property: if for $\mathbf{y} \in X$ there exists a t-steps path from $\mathbf{y}$ to $U_{\mathbf{x}^*}(\mathbf{z})$, and if for all $V_{\mathbf{z}}$ neighbourhood of $\mathbf{z}$, $V_{\mathbf{z}} \cap A$ has positive Lebesgue measure, then $P^{t+k}(\mathbf{y}, A) > 0$. Since B_0 is a non-empty open set, the trace-measure μ_{B_0} is non-trivial. Suppose that $\mu_{B_0}(A) > 0$, then there exists $\mathbf{z}_0 \in B_0 \cap A$ such that for all $V_{\mathbf{z}_0}$ neighbourhood of $\mathbf{z}_0$, $V_{\mathbf{z}_0} \cap A$ has positive Lebesgue measure[2]. And since $\mathbf{x}^*$ is globally attractive, according to Proposition 1 for all $\mathbf{y} \in X$ there exists $t_{\mathbf{y}} \in \mathbb{N}^*$ such that there exists a $t_{\mathbf{y}}$-steps path from $\mathbf{y}$ to the set $U_{\mathbf{x}^*}$ corresponding to $\mathbf{z}_0$. Hence, with Lemma 9, $P^{t_{\mathbf{y}}+k}(\mathbf{y}, A) > 0$ for all $\mathbf{y} \in X$ and so Φ is μ_{B_0}-irreducible.

If F is C^∞, according to Lemma 6, forward accessibility implies that for all $\mathbf{x} \in X$ there exists $k \in \mathbb{N}^*$ and $\mathbf{w} \in O_{\mathbf{x},k}$ such that the function $F^k(\mathbf{x}, \cdot)$ is a submersion at $\mathbf{w}$, which, using the first part of the proof of Theorem 2, shows that the existence of a globally attracting state implies the irreducibility of the Markov chain.

[2] If not, it would mean that for all $\mathbf{z} \in B_0 \cap A$, there exists $V_{\mathbf{z}}$ a neighbourhood of $\mathbf{z}$ such that $B_0 \cap A \cap V_{\mathbf{z}}$ is Lebesgue-negligible, which with Lemma 1 would imply that $B_0 \cap A$ is Lebesgue negligible and bring a contradiction.

Finally, if Φ is φ-irreducible, take $\mathbf{x}^* \in \operatorname{supp}\varphi$. By definition of the support of a measure, for all U neighbourhood of $\mathbf{x}^*$, $\varphi(U) > 0$. This imply through (15) that for all $\mathbf{y} \in X$ there exists $t \in \mathbb{N}^*$ such that $P^t(\mathbf{y}, U) > 0$. Since

$$P^t(\mathbf{y}, U) = \int_{O_{\mathbf{y},t}} 1_U(F^t(\mathbf{y}, \mathbf{w}))p_{\mathbf{y},t}(\mathbf{w})d\mathbf{w} > 0$$

this implies the existence of a t-steps path from $\mathbf{y}$ to U. Then, according to Proposition 1, $\mathbf{x}^*$ is a globally attracting state. □

Let $\mathbf{x}^* \in X$ be the globally attracting state used in Theorem 2. The support of the irreducibility measure used in Theorem 2 is a subset of $\overline{A_+(\mathbf{x}^*)}$. In the following proposition, we expend on this and show that when F is continuous and $p_{\mathbf{x}}$ lower semi-continuous, the support of the maximal irreducibility measure is exactly $\overline{A_+(\mathbf{x}^*)}$ for any globally attractive state $\mathbf{x}^*$.

Proposition 4. *Suppose that the function F is continuous, that the function $(\mathbf{x}, \mathbf{w}) \mapsto p_{\mathbf{x}}(\mathbf{w})$ is lower semi-continuous, and that the Markov chain Φ is φ-irreducible. Take ψ the maximal irreducibility measure of Φ. Then*

$$\operatorname{supp}\psi = \{\mathbf{x}^* \in X | \mathbf{x}^* \text{ is a globally attracting state}\} \ ,$$

and so, for $\mathbf{x}^* \in X$ a globally attracting state,

$$\operatorname{supp}\psi = \overline{A_+(\mathbf{x}^*)} \ .$$

Proof. Take $\mathbf{x}^* \in \operatorname{supp}\psi$, we will show that it is a globally attracting state. By definition of the support of a measure, for all U neighbourhood of $\mathbf{x}^*$, $\psi(U) > 0$. The measure ψ being a irreducibility measure, this imply through (15) that for all $\mathbf{y} \in X$ there exists $t \in \mathbb{N}^*$ such that $P^t(\mathbf{y}, U) > 0$, which in turns imply the existence of a t-steps path from $\mathbf{y}$ to U. Then, according to Proposition 1, $\mathbf{x}^*$ is a globally attracting state, and so $\operatorname{supp}\psi \subset \{\mathbf{x}^* \in X | \mathbf{x}^* \text{ is a globally attracting state}\}$.

Take $\mathbf{x}^* \in X$ a globally attracting state, then according to Proposition 1, for all $\mathbf{y} \in X$ there exists $t_{\mathbf{y}} \in \mathbb{N}^*$ and $\mathbf{w} \in O_{\mathbf{y},t_{\mathbf{y}}}$ such that $F^{t_{\mathbf{y}}}(\mathbf{y}, \mathbf{w}) \in B(\mathbf{x}^*, \epsilon)$. And since according to Lemma 2, $F^{t_{\mathbf{y}}}$ is continuous and that $B(\mathbf{x}^*, \epsilon)$ is an open, there exists $\eta > 0$ such that for all $\mathbf{u} \in B(\mathbf{w}, \eta)$, $F^{t_{\mathbf{y}}}(\mathbf{y}, \mathbf{u}) \in B(\mathbf{x}^*, \epsilon)$. Since p is lower semi-continuous and F continuous, according to Lemma 3 so is the function $(\mathbf{x}, \mathbf{w}) \mapsto p_{\mathbf{x},t_{\mathbf{y}}}(\mathbf{w})$ and so the set $O_{\mathbf{y},t_{\mathbf{y}}}$ is an open set. We can then chose the value of η small enough such that $B(\mathbf{w}, \eta) \subset O_{\mathbf{y},t_{\mathbf{y}}}$. Hence

$$P^{t_{\mathbf{y}}}(\mathbf{y}, B(\mathbf{x}^*, \epsilon)) \geq \int_{B(\mathbf{w},\eta)} \mathbf{1}_{B(\mathbf{x}^*,\epsilon)}\left(F^{t_{\mathbf{y}}}(\mathbf{y}, \mathbf{u})\right) p_{\mathbf{y},t_{\mathbf{y}}}(\mathbf{u})d\mathbf{u} = \int_{B(\mathbf{w},\eta)} p_{\mathbf{y},t_{\mathbf{y}}}(\mathbf{u})d\mathbf{u} > 0 \ .$$

The measure ψ being the maximal irreducibility measure, then

$$\psi(A) > 0 \Leftrightarrow \sum_{t\in\mathbb{N}^*} P^t(\mathbf{y}, A) > 0, \text{ for all } \mathbf{y} \in X \ .^3$$

[3] The implication $\Rightarrow$ is by definition of a irreducibility measure. For the converse suppose that A is a

Since we proved that for all $\mathbf{y} \in X$, $P^{t_\mathbf{y}}(\mathbf{y}, B(\mathbf{x}^*, \epsilon)) > 0$, we have $\psi(B(\mathbf{x}^*, \epsilon)) > 0$. Finally, since we can chose ϵ arbitrarily small, this implies that $\mathbf{x}^* \in \operatorname{supp}\psi$.

Let $(\mathbf{x}^*, \mathbf{y}^*) \in X^2$ be globally attracting states, then $\mathbf{y}^* \in \Omega_+(\mathbf{x}^*) \subset \overline{A_+(\mathbf{x}^*)}$, so $\{\mathbf{y}^* \in X | \mathbf{y}^* \text{ is a globally attracting state}\} \subset \overline{A_+(\mathbf{x}^*)}$.

Conversely, take $\mathbf{y}^* \in \overline{A_+(\mathbf{x}^*)}$, we will show that $\mathbf{y}^*$ is a globally attracting state. Since $\mathbf{y}^* \in \overline{A_+(\mathbf{x}^*)}$, for all $\epsilon > 0$ there exists $k_\epsilon \in \mathbb{N}^*$ and $\mathbf{w}_\epsilon$ a k_ϵ-steps path from $\mathbf{x}^*$ to $B(\mathbf{y}^*, \epsilon)$. Take $\mathbf{x} \in X$. Since $\mathbf{x}^*$ is a globally attracting state, according to Proposition 1 for all $\eta > 0$ there exists $t \in \mathbb{N}^*$ and $\mathbf{u}_\eta$ a t-steps path from $\mathbf{x}$ to $B(\mathbf{x}^*, \eta)$. And since F^{k_ϵ} is continuous, there exists $\eta_0 > 0$ such that for all $\mathbf{z} \in B(\mathbf{x}^*, \eta_0)$, $F^{k_\epsilon}(\mathbf{z}, \mathbf{w}_\epsilon) \in B(\mathbf{y}^*, \epsilon)$. Furthermore, since the set $\{(\mathbf{x}, \mathbf{w}) \in X \times O^{k_\epsilon} | p_{\mathbf{x}, k_\epsilon}(\mathbf{w}) > 0\}$ is an open set we can take η_0 small enough to ensure that $\mathbf{w}_\epsilon \in O_{F^t(\mathbf{x}, \mathbf{u}_{\eta_0}), k_\epsilon}$. Hence for any $\mathbf{x} \in X$, $\epsilon > 0$, $(\mathbf{u}_{\eta_0}, \mathbf{w}_\epsilon)$ is a $t + k_\epsilon$-steps path from $\mathbf{x}$ to $B(\mathbf{y}^*, \epsilon)$, which with Proposition 1 proves that $\mathbf{y}^*$ is a globally attracting state. Hence $\overline{A_+(\mathbf{x}^*)} \subset \{\mathbf{y}^* | \mathbf{y}^* \text{ is a globally attracting state}\}$. □

3.2. Aperiodicity

The results of Lemma 9 give the existence of a non-empty open set B_0 such that for all $\mathbf{z} \in B_0$ there exists $U_{\mathbf{x}^*}$ a neighbourhood of $\mathbf{x}^*$ which depends of $\mathbf{z}$. And if $V_\mathbf{z} \cap A$ has positive Lebesgue measure for all $V_\mathbf{z}$ neighbourhood of $\mathbf{z}$, then for all $\mathbf{y} \in X$ the existence of a t-steps path from $\mathbf{y}$ to $U_{\mathbf{x}^*}$ implies that $P^{t+k}(\mathbf{y}, A) > 0$. Note that $P^t(\mathbf{y}, A) > 0$ holds true for any $t \in \mathbb{N}^*$ such that there exists a t-steps path from $\mathbf{y}$ to $U_{\mathbf{x}^*}$.

The global attractivity of $\mathbf{x}^*$ gives for any $\mathbf{y} \in X$ the existence of one such t for which there exists a t-step path from $\mathbf{y}$ to $U_{\mathbf{x}^*}$; and as seen in Theorem 2 this can be exploited to prove the irreducibility of the Markov chain. However, the strong global attractivity of $\mathbf{x}^*$ gives for all $\mathbf{y} \in X$ the existence of a $t_\mathbf{y}$ such that for all $t \geq t_\mathbf{y}$ there exists a t-step path from $\mathbf{y}$ to $U_{\mathbf{x}^*}$, which implies that $P^t(\mathbf{y}, A) > 0$ for all $t \geq t_\mathbf{y}$ and for all $\mathbf{y} \in X$. We will see in the following theorem that this implies the aperiodicity of the Markov chain.

Theorem 3. *Suppose that*

1. *the function* $(\mathbf{x}, \mathbf{w}) \mapsto F(\mathbf{x}, \mathbf{w})$ *is* C^1,
2. *the function* $(\mathbf{x}, \mathbf{w}) \mapsto p_\mathbf{x}(\mathbf{w})$ *is lower semi-continuous,*
3. *there exists* $\mathbf{x}^* \in X$ *a strongly globally attractive state,* $k \in \mathbb{N}^*$ *and* $\mathbf{w}^* \in O_{\mathbf{x}^*, k}$ *such that* $F^k(\mathbf{x}^*, \cdot)$ *is a submersion at* $\mathbf{w}^*$.

Then there exists B_0 *a non-empty open subset of* $A_+^k(\mathbf{x}^*)$ *containing* $F^k(\mathbf{x}^*, \mathbf{w}^*)$ *such that* Φ *is a* μ_{B_0}*-irreducible aperiodic Markov chain.*

Proof. According to Theorem 2 there exists B_0 an open neighbourhood of $F^k(\mathbf{x}^*, \mathbf{w}^*)$ such that the chain Φ is μ_{B_0}-irreducible. Let ψ be its maximal irreducibility measure

set such that $\sum_{t \in \mathbb{N}^*} P^t(\mathbf{y}, A) > 0$ for all $\mathbf{y} \in X$, so the set $\{\mathbf{y} \in X | \sum_{t \in \mathbb{N}^*} P^t(\mathbf{y}, A) > 0\}$ equals X. If $\psi(A) = 0$, from [8, Theorem 4.0.1] this would imply that the set $\{\mathbf{y} \in X | \sum_{t \in \mathbb{N}^*} P^t(\mathbf{y}, A) > 0\}$, which equals X, is also ψ-null, which is impossible since by definition ψ is a non-trivial measure. Therefore $\sum_{t \in \mathbb{N}^*} P^t(\mathbf{y}, A) > 0$ for all $\mathbf{y} \in X$ implies that $\psi(A) > 0$.

(which exists according to [8, Theorem 4.0.1]). According to [8, Theorem 5.4.4.] there exists $d \in \mathbb{N}^*$ and a sequence $(D_i)_{i\in[0..d-1]} \in \mathcal{B}(X)^d$ of sets such that

1. for $i \neq j$, $D_i \cap D_j = \emptyset$
2. $\mu_{B_0}((\bigcup_{i=0}^{d-1} D_i)^c) = 0$
3. for $i = 0, \ldots, d-1 \pmod d$, for $\mathbf{x} \in D_i$, $P(\mathbf{x}, D_{i+1}) = 1$

Note that 2. is usually stated with the maximal measure ψ but then of course also holds for μ_{B_0}. We will prove that $d = 1$.

From 3. we deduce that for $\mathbf{x} \in D_i$ and $j \in \mathbb{N}^*$, $P^j(\mathbf{x}, D_{i+j \bmod d}) = 1$. And with the first point for $l \neq j \bmod d$, $P^j(\mathbf{x}, D_{i+l \bmod d}) = 0$.

From Lemma 9, there exists $\tilde{B}_0$, an open neighbourhood of $F^k(\mathbf{x}^*, \mathbf{w}^*)$ such that for all $\mathbf{z} \in \tilde{B}_0$ there exists $U_{\mathbf{x}^*}$ an open neighbourhood of $\mathbf{x}^*$ having the following property: for $\mathbf{y} \in X$ if there exists a t-steps path from $\mathbf{y}$ to $U_{\mathbf{x}^*}$ and if given A in $\mathcal{B}(X)$, for all $V_\mathbf{z}$ open neighbourhood of $\mathbf{z}$, $V_\mathbf{z} \cap A$ has positive Lebesgue measure then $P^{t+k}(\mathbf{y}, A) > 0$. We did not show that $B_0 = \tilde{B}_0$, but we can consider the set $B_1 = B_0 \cap \tilde{B}_0$ which is also an open neighbourhood of $F^k(\mathbf{x}^*, \mathbf{w}^*)$.

Then with 2., $\mu_{B_0}((\bigcup_{i=0}^{d-1} D_i)^c) \geq \mu_{B_1}((\bigcup_{i=0}^{d-1} D_i)^c) = 0$, and since B_1 is a non-empty open set, μ_{B_1} is not trivial hence $\mu_{B_1}(\bigcup_{i=0}^{d-1} D_i) > 0$. So there exists $i \in [0..d-1]$ and $\mathbf{z} \in B_1$ such that for all $V_\mathbf{z}$ open neighbourhood of $\mathbf{z}$, $V_\mathbf{z} \cap D_i$ has positive Lebesgue measure (as implied by the contrapositive of Lemma 1).

Since $\mathbf{x}^*$ is a strongly globally attracting state, for all $\mathbf{y} \in X$ there exists $t_\mathbf{y} \in \mathbb{N}^*$ such that for all $t \geq t_\mathbf{y}$ there exists a t-steps path from $\mathbf{y}$ to $U_{\mathbf{x}^*}$. Using the property of $U_{\mathbf{x}^*}$, this implies that $P^{t+k}(\mathbf{y}, D_i) > 0$. Since this holds for any $t \geq t_\mathbf{y}$, it also holds for $t = d(t_\mathbf{y} + k) + 1 - k$, and so $P^{d(t_\mathbf{y}+k)+1-k+k}(\mathbf{y}, D_i) > 0$. As we had deduced that for $l \neq j \pmod d$, $P^j(\mathbf{y}, D_{i+l \bmod d}) = 0$, we can conclude that $d(t_\mathbf{y} + k) + 1 = 0 \pmod d$, hence $1 = 0 \pmod d$ meaning $d = 1$ and so Φ is aperiodic. □

In [8, Proposition 7.3.4 and Theorem 7.3.5] in the context of the function $(\mathbf{x}, \mathbf{w}) \mapsto p_\mathbf{x}(\mathbf{w})$ being independent of $\mathbf{x}$, F being C^∞ and $p_\mathbf{x}$ lower semi-continuous, under the assumption that the control model is forward accessible, that there exists a globally attracting state $\mathbf{x}^* \in X$ and that the set $O_\mathbf{x}$ is connected, aperiodicity is proven equivalent to the connexity of $\overline{A_+(\mathbf{x}^*)}$. Although in most practical cases the set $O_\mathbf{x}$ is connected, it is good to keep in mind that when $O_\mathbf{x}$ is not connected, $\overline{A_+(\mathbf{x}^*)}$ can also not be connected and yet the Markov chain can be aperiodic (e.g. any sequence of i.i.d. random variables with non-connected support is a φ-irreducible aperiodic Markov chain). In such problems our approach still offer conditions to show the aperiodicity of the Markov chain.

3.3. Weak-Feller

Our main result summarized in Theorem 1 uses the fact that the chain is weak Feller. Our experience is that this property can be often easily verified by proving that if f is

continuous and bounded then

$$\mathbf{x} \in X \mapsto \int f(F(\mathbf{x}, \mathbf{w}))p_{\mathbf{x}}(\mathbf{w})d\mathbf{w}$$

is continuous and bounded. This latter property often deriving from the dominated convergence theorem. We however provide below another result to automatically prove the weak Feller property.

Proposition 5. *Suppose that*

- *for all* $\mathbf{w} \in O$ *the function* $\mathbf{x} \in X \mapsto F(\mathbf{x}, \mathbf{w})$ *is continuous,*
- *for all* $\mathbf{x} \in X$ *the function* $\mathbf{w} \in O \mapsto F(\mathbf{x}, \mathbf{w})$ *is measurable,*
- *for all* $\mathbf{w} \in O$*, the function* $\mathbf{x} \in X \mapsto p_{\mathbf{x}}(\mathbf{w})$ *is lower semi-continuous.*
- *for all* $\mathbf{x} \in X$ *the function* $\mathbf{w} \in O \mapsto p_{\mathbf{x}}(\mathbf{w})$ *is measurable,*

Then the Markov chain Φ *is weak-Feller.*

Proof. To be weak-Feller means that for any open set $U \in \mathcal{B}(X)$ the function $\mathbf{x} \in X \mapsto P(\mathbf{x}, U)$ is lower semi-continuous.

Take $\mathbf{x} \in X$ and $\mathbf{w} \in O$. If $F(\mathbf{x}, \mathbf{w}) \notin U$ then $\forall \mathbf{y} \in X$, $\mathbf{1}_U(F(\mathbf{y}, \mathbf{w})) \geq \mathbf{1}_U(F(\mathbf{x}, \mathbf{w})) = 0$. If $F(\mathbf{x}, \mathbf{w}) \in U$ as U is an open set there exists $\epsilon > 0$ such that $B(F(\mathbf{x}, \mathbf{w}), \epsilon) \subset U$, and as the function $\mathbf{y} \mapsto F(\mathbf{y}, \mathbf{w})$ is continuous for all $\epsilon > 0$ there exists $\eta > 0$ such that if $\mathbf{y} \in B(\mathbf{x}, \eta)$ then $F(\mathbf{y}, \mathbf{w}) \in B(F(\mathbf{x}, \mathbf{w}), \epsilon) \subset U$. Therefore for all $\mathbf{y}$ in the neighbourhood $B(\mathbf{x}, \eta)$ we have $\mathbf{1}_U(F(\mathbf{y}, \mathbf{w})) = \mathbf{1}_U(F(\mathbf{x}, \mathbf{w})) \geq \mathbf{1}_U(F(\mathbf{x}, \mathbf{w})) - \epsilon$, meaning the function $\mathbf{x} \in X \mapsto \mathbf{1}_U(F(\mathbf{x}, \mathbf{w}))$ is lower semi-continuous. For $\mathbf{w} \in O$ the function $\mathbf{x} \mapsto p_{\mathbf{x}}(\mathbf{w})$ is assumed lower semi-continuous, hence so is $\mathbf{x} \mapsto \mathbf{1}_U(F(\mathbf{x}, \mathbf{w}))p_{\mathbf{x}}(\mathbf{w})$.

Finally we can apply Fatou's Lemma for all sequence $(\mathbf{x}_t)_{t\in\mathbb{N}} \in X^{\mathbb{N}}$ converging to $\mathbf{x}$:

$$\begin{aligned}
\liminf P(\mathbf{x}_t, U) &= \liminf \int_O \mathbf{1}_U(F(\mathbf{x}_t, \mathbf{w}))p_{\mathbf{x}_t}(\mathbf{w})\mathrm{d}\mathbf{w} \\
&\geq \int_O \liminf \mathbf{1}_U(F(\mathbf{x}_t, \mathbf{w}))p_{\mathbf{x}_t}(\mathbf{w})\mathrm{d}\mathbf{w} \\
&\geq \int_O \mathbf{1}_U(F(\mathbf{x}, \mathbf{w}))p_{\mathbf{x}}(\mathbf{w})\mathrm{d}\mathbf{w} = P(\mathbf{x}, U) \ .
\end{aligned}$$

□

4. Applications

We illustrate now the usefulness of Theorem 1. For this, we present two examples of Markov chains that can be modeled via (2) and detail how to apply Theorem 1 to prove their φ-irreducibility, aperiodicity and the fact that compact sets are small sets. Those Markov chains stem from adaptive stochastic algorithms aiming at optimizing continuous optimization problems. Their stability study implies the linear convergence

(or divergence) of the underlying algorithm. Those examples are not artificial: in both cases, showing the φ-irreducibility, aperiodicity and the fact that compact sets are small sets by hand without the results of the current paper seem to be very difficult. They actually motivated the development of the theory of this paper.

4.1. A step-size adaptive randomized search on scaling-invariant functions

We consider first a step-size adaptive stochastic search algorithm optimizing an objective function $f : \mathbb{R}^n \to \mathbb{R}$ without constraints. The algorithm pertains to the class of so-called *Evolution Strategies* (ES) algorithms [12] that date back to the 70's. The algorithm is however related to information geometry. It was recently derived from taking the natural gradient of a joint objective function defined on the Riemannian manifold formed by the family of Gaussian distributions [4, 9]. More precisely, let $\mathbf{X}_0 \in \mathbb{R}^n$ and let $(\mathbf{U}_t)_{t\in\mathbb{N}^*}$ be an i.i.d. sequence of random vectors where each $\mathbf{U}_t$ is composed of $\lambda \in \mathbb{N}^*$ components $\mathbf{U}_t = (\mathbf{U}_t^1, \ldots, \mathbf{U}_t^\lambda) \in (\mathbb{R}^n)^\lambda$ with $(\mathbf{U}_t^i)_{i\in[1..\lambda]}$ i.i.d. and following each a standard multivariate normal distribution $\mathcal{N}(0, \mathrm{I}_n)$. Given $(\mathbf{X}_t, \sigma_t) \in \mathbb{R}^n \times \mathbb{R}_+^*$, the current state of the algorithm, λ candidate solutions centered on $\mathbf{X}_t$ are sampled using the vector $\mathbf{U}_{t+1}$, i.e. for i in $[1..\lambda]$

$$\mathbf{X}_t + \sigma_t \mathbf{U}_{t+1}^i \ , \tag{24}$$

where σ_t called the step-size of the algorithm corresponds to the overall standard deviation of $\sigma_t \mathbf{U}_{t+1}^i$. Those solutions are ranked according to their f-values. More precisely, let $\mathcal{S}$ be the permutation of λ elements such that

$$f\left(\mathbf{X}_t + \sigma_t \mathbf{U}_{t+1}^{\mathcal{S}(1)}\right) \leq f\left(\mathbf{X}_t + \sigma_t \mathbf{U}_{t+1}^{\mathcal{S}(2)}\right) \leq \ldots \leq f\left(\mathbf{X}_t + \sigma_t \mathbf{U}_{t+1}^{\mathcal{S}(\lambda)}\right) \ . \tag{25}$$

To break the possible ties and have an uniquely defined permutation $\mathcal{S}$, we can simply consider the natural order, i.e. if for instance $\lambda = 2$ and $f\left(\mathbf{X}_t + \sigma_t \mathbf{U}_{t+1}^1\right) = f\left(\mathbf{X}_t + \sigma_t \mathbf{U}_{t+1}^2\right)$, then $\mathcal{S}(1) = 1$ and $\mathcal{S}(2) = 2$. The new estimate of the optimum $\mathbf{X}_{t+1}$ is formed by taking a weighted average of the μ best directions (typically $\mu = \lambda/2$), that is

$$\mathbf{X}_{t+1} = \mathbf{X}_t + \kappa_m \sum_{i=1}^{\mu} w_i \mathbf{U}_{t+1}^{\mathcal{S}(i)} \tag{26}$$

where the sequence of weights $(w_i)_{i\in[1..\mu]}$ sums to 1, and $\kappa_m > 0$ is called a learning rate. The step-size is adapted according to

$$\sigma_{t+1} = \sigma_t \exp\left(\frac{\kappa_\sigma}{2n}\left(\sum_{i=1}^{\mu} w_i \left(\|\mathbf{U}_{t+1}^{\mathcal{S}(i)}\|^2 - n\right)\right)\right) \ , \tag{27}$$

where $\kappa_\sigma > 0$ is a learning rate for the step-size. The equations (26) and (27) correspond to the so-called xNES algorithm with covariance matrix restricted to $\sigma_t^2 \mathrm{I}_n$ [4]. One crucial question in optimization is related to the convergence of an algorithm.

On the class of so-called scaling-invariant functions (see below for the definition) with optimum in $\mathbf{x}^* \in \mathbb{R}^n$, a proof of the *linear convergence* of the aforementioned algorithm can be obtained if the normalized chain $\mathbf{Z}_t = (\mathbf{X}_t - \mathbf{x}^*)/\sigma_t$—which turns out to be an homogeneous Markov chain—is stable enough to satisfy a Law of Large Numbers. This result is explained in details in [1] but in what follows we remind for the sake of completeness the definition of a scaling invariant function and detail the expression of the chain $\mathbf{Z}_t$.

A function is scaling-invariant with respect to $\mathbf{x}^*$ if for all $\rho > 0$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$

$$f(\mathbf{x}) \leq f(\mathbf{y}) \Leftrightarrow f(\mathbf{x}^* + \rho(\mathbf{x} - \mathbf{x}^*)) \leq f(\mathbf{x}^* + \rho(\mathbf{y} - \mathbf{x}^*)) \ . \tag{28}$$

Examples of scaling-invariant functions include $f(\mathbf{x}) = \|\mathbf{x} - \mathbf{x}^*\|$ for any arbitrary norm on $\mathbb{R}^n$. It also includes functions with non-convex sublevel sets, i.e. non-quasi-convex functions.

As mentioned above, on this class of functions, $\mathbf{Z}_t = (\mathbf{X}_t - \mathbf{x}^*)/\sigma_t$ is an homogeneous Markov chain that can be defined independently of the Markov chain $(\mathbf{X}_t, \sigma_t)$ in the following manner. Given $\mathbf{Z}_t \in \mathbb{R}^n$ sample λ candidate solutions centered on $\mathbf{Z}_t$ using a vector $\mathbf{U}_{t+1}$, i.e. for i in $[1..\lambda]$

$$\mathbf{Z}_t + \mathbf{U}_{t+1}^i \ , \tag{29}$$

where similarly as for the chain $(\mathbf{X}_t, \sigma_t)$, $(\mathbf{U}_t)_{t\in\mathbb{N}}$ are i.i.d. and each $\mathbf{U}_t$ is a vectors of λ i.i.d. components following a standard multivariate normal distribution. Those λ solutions are evaluated and ranked according to their f-values. Similarly to (25), the permutation $\mathcal{S}$ containing the order of the solutions is extracted. This permutation can be uniquely defined if we break the ties as explained below (25). The update of $\mathbf{Z}_t$ then reads

$$\mathbf{Z}_{t+1} = \frac{\mathbf{Z}_t + \kappa_m \sum_{i=1}^{\mu} w_i \mathbf{U}_{t+1}^{\mathcal{S}(i)}}{\exp\left(\frac{\kappa_\sigma}{2n}\left(\sum_{i=1}^{\mu} w_i (\|\mathbf{U}_{t+1}^{\mathcal{S}(i)}\|^2 - n)\right)\right)} \ . \tag{30}$$

We refer to [1, Proposition 1] for the details. Let us now define $\mathbf{W}_{t+1} = (\mathbf{U}_{t+1}^{\mathcal{S}(1)}, \ldots, \mathbf{U}_{t+1}^{\mathcal{S}(\mu)}) \in \mathbb{R}^{n\times\mu}$ and for $\mathbf{z} \in \mathbb{R}^n$, $\mathbf{y} \in (\mathbb{R}^n)^\mu$ (with $\mathbf{y} = (\mathbf{y}^1, \ldots, \mathbf{y}^\mu)$)

$$F_{\text{xNES}}(\mathbf{z}, \mathbf{y}) = \frac{\mathbf{z} + \kappa_m \sum_{i=1}^{\mu} w_i \mathbf{y}^i}{\exp\left(\frac{\kappa_\sigma}{2n}\left(\sum_{i=1}^{\mu} w_i (\|\mathbf{y}^i\|^2 - n)\right)\right)} \ , \tag{31}$$

such that

$$\mathbf{Z}_{t+1} = F_{\text{xNES}}(\mathbf{Z}_t, \mathbf{W}_{t+1}) \ .$$

Also there exists a function $\alpha : (\mathbb{R}^n, \mathbb{R}^{n\times\lambda}) \to \mathbb{R}^{n\times\mu}$ such that $\mathbf{W}_{t+1} = \alpha(\mathbf{Z}_t, \mathbf{U}_{t+1})$. Indeed, given $\mathbf{z}$ and $\mathbf{u}$ in $\mathbb{R}^{n\times\lambda}$ we have explained how the permutation giving the ranking of the candidate solutions $\mathbf{z} + \mathbf{u}^i$ on f can be uniquely defined. Then $\alpha(\mathbf{z}, \mathbf{u}) = (\mathbf{u}^{\mathcal{S}(1)}, \ldots, \mathbf{u}^{\mathcal{S}(\lambda)})$. Hence we have just explained why the Markov chain defined via (30) fits the Markov chain model underlying this paper, that is (2). If we assume that the level sets of the function f are Lebesgue negligible, then a density $p : (\mathbf{z}, \mathbf{w}) \in \mathbb{R}^n \times \mathbb{R}^{n\times\mu} \mapsto \mathbb{R}_+$

associated to $\mathbf{W}_{t+1}$ writes

$$p(\mathbf{z},\mathbf{w}) = \frac{\lambda!}{(\lambda-\mu)!}\mathbf{1}_{\{f(\mathbf{z}+\mathbf{w}^1)<\ldots<f(\mathbf{z}+\mathbf{w}^\mu)\}}(1-Q^f_{\mathbf{z}}(\mathbf{w}^\mu))^{\lambda-\mu}p_{\mathcal{N}}(\mathbf{w}^1)\ldots p_{\mathcal{N}}(\mathbf{w}^\mu) \quad (32)$$

with each $\mathbf{w}^i \in \mathbb{R}^n$ and $\mathbf{w} = (\mathbf{w}^1,\ldots,\mathbf{w}^\mu)$, where $Q^f_{\mathbf{z}}(\mathbf{w}^\mu) = \Pr\left(f(\mathbf{z}+\mathcal{N}) \leq f(\mathbf{z}+\mathbf{w}^\mu)\right)$ with $\mathcal{N}$ following a standard multivariate normal distribution and

$$p_{\mathcal{N}}(\mathbf{y}) = \frac{1}{(\sqrt{2\pi})^n}\exp(-\mathbf{y}^T\mathbf{y}/2)$$

the density of a standard multivariate normal distribution in dimension n. If the objective function f is continuous, then the density $p(\mathbf{z},\mathbf{w})$ is lower semi-continuous.

We now prove by applying Theorem 1 that the Markov chain $\mathbf{Z}_t$ is a φ-irreducible aperiodic T-chain and compact sets are small sets for the chain. Those properties together with a drift for positivity will imply the linear convergence of the xNES algorithm [1].

Proposition 6. *Suppose that the scaling invariant function f is continuous, and that its level sets are Lebesgue negligible. Then the Markov chain $(\mathbf{Z}_t)_{t\in\mathbb{N}}$ defined in (30) is a φ-irreducible aperiodic T-chain and compact sets are small sets for the chain.*

Proof. It is not difficult to see that $p(\mathbf{z},\mathbf{w})$ is lower-semi continuous since f is continuous and that F_{xNES} is a C^1 function. We remind that $O_{\mathbf{z}} = \{\mathbf{w}\in\mathbb{R}^{n\times\mu} | p(\mathbf{z},\mathbf{w}) > 0\}$ hence with (32) $O_{\mathbf{z}} = \{\mathbf{w}\in\mathbb{R}^{n\times\mu} | \mathbf{1}_{\{f(\mathbf{z}+\mathbf{w}^1)<\ldots<f(\mathbf{z}+\mathbf{w}^\mu)\}>0}\}$.

We will now prove that the point $\mathbf{z}^* := \mathbf{0}$ is a strongly globally attracting state. For $\mathbf{y}\in\mathbb{R}^n$ and $\epsilon\in\mathbb{R}^*_+$, this means there exists a $t_{\mathbf{y},\epsilon}\in\mathbb{N}^*$ such that for all $t\geq t_{\mathbf{y},\epsilon}$, there exists a t-steps path from $\mathbf{y}$ to $B(\mathbf{0},\epsilon)$. Note that $\lim_{\|\mathbf{w}\|\to+\infty} F_{\text{xNES}}(\mathbf{y},\mathbf{w}) = \mathbf{0}$, meaning that there exists a $r\in\mathbb{R}^*_+$ such that if $\|\mathbf{w}\|\geq r$ then $F_{\text{xNES}}(\mathbf{z},\mathbf{w})\in B(\mathbf{0},\epsilon)$. Therefore, and since $O_{\mathbf{y}}\cap\{\mathbf{w}\in\mathbb{R}^{n\times\mu} | \|\mathbf{w}\|\geq r\}$ is non empty, there exists a $\mathbf{w}_{\mathbf{y},\epsilon}\in O_{\mathbf{y}}$ which is a 1-step path from $\mathbf{y}$ to $B(\mathbf{0},\epsilon)$. Now, showing that there is such a path from $\tilde{\mathbf{y}}\in\mathbb{R}^n$ for all $t\geq 1$ is trivial: take $\mathbf{w}\in O_{\mathbf{y},t-1}$, and denote $\mathbf{y} = F^{t-1}(\tilde{\mathbf{y}},\mathbf{w})$; $(\mathbf{w},\mathbf{w}_{\mathbf{y},\epsilon})$ is a t-steps path from $\tilde{\mathbf{y}}$ to $B(\mathbf{0},\epsilon)$.

We now prove that there exists $\mathbf{w}^*\in O_{\mathbf{0}}$ such that $F(\mathbf{0},\cdot)$ is a submersion at $\mathbf{w}^*$, by proving that the differential of $F(\mathbf{0},\cdot)$ at $\mathbf{w}^*$ is surjective. Take $\mathbf{w}_0 = (\mathbf{0},\ldots,\mathbf{0})\in\mathbb{R}^{n\times\mu}$ and $\mathbf{h} = (\mathbf{h}_i)_{i\in[1..\mu]}\in\mathbb{R}^{n\times\mu}$, then

$$\begin{aligned}
F_{\text{xNES}}(\mathbf{0},\mathbf{0}+\mathbf{h}) &= \frac{\kappa_m\sum_{i=1}^{\mu} w_i\mathbf{h}_i}{\exp\left(\frac{\kappa_\sigma}{2n}\left(\sum_{i=1}^{\mu} w_i(\|\mathbf{h}_i\|^2-n)\right)\right)} \\
&= \mathbf{0} + \kappa_m\left(\sum_{i=1}^{\mu} w_i\mathbf{h}_i\right)\exp\left(\frac{\kappa_\sigma}{2}\right)\exp\left(-\frac{\kappa_m}{2n}\sum_{i=1}^{\mu} w_i\|\mathbf{h}_i\|^2\right) \\
&= F_{\text{xNES}}(\mathbf{0},\mathbf{0}) + \kappa_m\left(\sum_{i=1}^{\mu} w_i\mathbf{h}_i\right)\exp\left(\frac{\kappa_\sigma}{2}\right)(1+o(\|\mathbf{h}\|)) \quad .
\end{aligned}$$

Hence $D_{\mathbf{w}_0} F_{\text{xNES}}(\mathbf{0}, \cdot)(\mathbf{h}) = \kappa_m \exp(\kappa_\sigma/2) \sum_{i=1}^{\mu} w_i \mathbf{h}_i$, and is therefore a surjective linear map. The point $\mathbf{w}_0$ is not in $O_{\mathbf{0}}$, but according to Lemma 5 since $F_{\text{xNES}}(\mathbf{0}, \cdot)$ is a submersion at $\mathbf{w}_0$ there exists $V_{\mathbf{w}_0}$ an open neighbourhood of $\mathbf{w}_0$ such that for all $\mathbf{v} \in V_{\mathbf{w}_0}$, $F_{\text{xNES}}(\mathbf{0}, \cdot)$ is a submersion at $\mathbf{v}$. Finally since $V_{\mathbf{w}_0} \cap O_{\mathbf{0}}$ is not empty, there exists $\mathbf{w}^* \in O_{\mathbf{0}}$ such that $F_{\text{xNES}}(\mathbf{0}, \cdot)$ is a submersion at $\mathbf{w}^*$.

We can then apply Theorem 1 which shows that $(\mathbf{Z}_t)_{t \in \mathbb{N}}$ is a ψ-irreducible aperiodic T-chain, and that compact sets are small sets for the chain. □

4.2. A step-size adaptive randomized search on a simple constraint optimization problem

We now consider a similar algorithm belonging to the class of evolution strategies optimizing a linear function under a linear constraint. The goal for the algorithm is to diverge as fast as possible as the optimum of the problem is at infinity. More precisely let $f, g : \mathbb{R}^n \to \mathbb{R}$ be two linear functions (w.lo.g. we take $f(\mathbf{x}) = [\mathbf{x}]_1$, and $g(\mathbf{x}) = -\cos\theta[\mathbf{x}]_1 - \sin\theta[\mathbf{x}]_2$) with $\theta \in (0, \pi/2)$. The goal is to maximize f while respecting the constraint $g(\mathbf{x}) > 0$.

As for the previous algorithm, the state of the algorithm is reduced to $(\mathbf{X}_t, \sigma_t) \in \mathbb{R}^n \times \mathbb{R}_+^*$ where $\mathbf{X}_t$ represents the favorite solution and σ_t is the step-size controlling the standard deviation of the sampling distribution used to generate new solutions. From $\mathbf{X}_t$, λ new solutions are sampled

$$\mathbf{Y}_{t+1}^i = \mathbf{X}_t + \sigma_t \mathbf{V}_{t+1}^i \ , \tag{33}$$

where each $\mathbf{V}_t = (\mathbf{V}_t^1, \ldots, \mathbf{V}_t^\lambda)$ with $(\mathbf{V}_t^i)_i$ i.i.d. following a standard multivariate normal distribution in dimension n. Those solutions may lie in the infeasible domain, that is they might violate the constraint, i.e. $g(\mathbf{Y}_{t+1}^i) \leq 0$. Hence a specific mechanism is added to ensure that we have λ solutions within the feasible domain. Here this mechanism is very simple, it consists in resampling a solution till it lies in the feasible domain. We denote $\tilde{\mathbf{Y}}_{t+1}^i$ the candidate solution i that satisfies the constraint. While the resampling of a candidate solution can possibly call for an infinite numbers of multivariate normal distribution, it can be shown in our specific case that this candidate solution can be generated using a single random vector $\mathbf{U}_{t+1}^i$ and is a function of the normalized distance to the constraint $\delta_t = g(\mathbf{X}_t)/\sigma_t$. This is due to the fact that the distribution of the feasible candidate solution orthogonal to the constraint direction follows a truncated Gaussian distribution and orthogonal to the constraint a Gaussian distribution (we refer to [2, Lemma 2] for the details). Hence overall,

$$\tilde{\mathbf{Y}}_{t+1}^i = \mathbf{X}_t + \sigma_t \tilde{\mathcal{G}}(\delta_t, \mathbf{U}_{t+1}^i)$$

where $[\mathbf{U}_t = (\mathbf{U}_t^1, \ldots, \mathbf{U}_t^\lambda)]_t$ are i.i.d. (see [2, Lemma 2]) and the function $\tilde{\mathcal{G}}$ is defined in [2, equation (15)]. Those λ feasible candidate solutions are ranked on the objective function f and as before, the permutation $\mathcal{S}$ containing the ranking of the solutions is

extracted. The update of $\mathbf{X}_{t+1}$ then reads

$$\mathbf{X}_{t+1} = \mathbf{X}_t + \sigma_t \tilde{\mathcal{G}}(\delta_t, \mathbf{U}_{t+1}^{\mathcal{S}(1)}) \ , \tag{34}$$

that is the best solution is kept. The update of the step-size satisfies

$$\sigma_{t+1} = \sigma_t \exp\left(\frac{1}{2d_\sigma}\left(\frac{\|\tilde{\mathcal{G}}(\delta_t, \mathbf{U}_{t+1}^{\mathcal{S}(1)})\|^2}{n} - 1\right)\right) \ , \ d_\sigma \in \mathbb{R}_+^* \ . \tag{35}$$

This algorithm corresponds to a so-called $(1, \lambda)$-ES with resampling using the cumulative step-size adaptation mechanism of the covariance matrix adaptation ES (CMA-ES) algorithm [5].

It is not difficult to show that $(\delta_t)_{t\in\mathbb{N}}$ is an homogeneous Markov chain (see [2, Proposition 5]) whose update reads

$$\delta_{t+1} = \left(\delta_t + g(\tilde{\mathcal{G}}(\delta_t, \mathbf{U}_{t+1}^{\mathcal{S}(1)}))\right) \exp\left(-\frac{1}{2d_\sigma}\left(\frac{\|\tilde{\mathcal{G}}(\delta_t, \mathbf{U}_{t+1}^{\mathcal{S}(1)})\|^2}{n} - 1\right)\right) \ . \tag{36}$$

and that the divergence of the algorithm can be proven if $(\delta_t)_{t\in\mathbb{N}}$ satisfies a Law of Large Numbers. Given that typical conditions to prove that an homogeneous Markov chain satisfies a LLN is φ-irreducibility, aperiodicity, Harris-recurrence and positivity and that those latter two conditions are practical to verify with drift conditions that hold outside a small set, we see the interest to be able to prove the irreducibility aperiodicity and identify that compact sets are small sets for $(\delta_t)_{t\in\mathbb{N}}$.

With respect to the modeling of the paper, let $\mathbf{W}_t = \tilde{\mathcal{G}}(\delta_t, \mathbf{U}_{t+1}^{\mathcal{S}(1)})$, then there is a well-defined function α such that $\mathbf{W}_t = \alpha(\delta_t, \mathbf{U}_{t+1})$ and according to [2, Lemma 3] the density $p(\delta, \mathbf{w})$ of $\mathbf{W}_t$ knowing that $\delta_t = \delta$ equals

$$p(\delta, \mathbf{w}) = \lambda \frac{p_{\mathcal{N}}(\mathbf{w}) \mathbf{1}_{\mathbb{R}_+^*}(\delta + g(\mathbf{w}))}{\mathcal{F}_{p_{\mathcal{N}}}(\delta)} \left(\int_{-\infty}^{[\mathbf{w}]_1} p_1(u) \frac{\mathcal{F}_{p_{\mathcal{N}}}(\frac{\delta - u\cos\theta}{\sin\theta})}{\mathcal{F}_{p_{\mathcal{N}}}(\delta)} \mathrm{d}u\right)^{\lambda-1} \ , \tag{37}$$

where p_1 is the density of a one dimensional normal distribution, $p_{\mathcal{N}}$ the density of a n-dimensional multivariate normal distribution and $\mathcal{F}_{p_{\mathcal{N}}}$ its associated cumulative distribution function.

The state space X for the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is $\mathbb{R}_+^*$, the set O equals to $\mathbb{R}^n$ and the function F implicitly given in (36):

$$F(\delta, \mathbf{w}) = (\delta + g(\mathbf{w})) \exp\left(-\frac{1}{2d_\sigma}\left(\frac{\|\mathbf{w}\|^2}{n} - 1\right)\right) \ . \tag{38}$$

The control set $O_{x,t}$ equals

$$O_{x,t} := \{(\mathbf{w}_1, \ldots, \mathbf{w}_t) \in \mathbb{R}^{nt} | x > -g(\mathbf{w}_1), \ldots, F^{t-1}(x, \mathbf{w}_1, \ldots, \mathbf{w}_{t-1}) > -g(\mathbf{w}_t)\} \ .$$

We are now ready to apply the results develop within the paper to prove that the chain $(\delta_t)_{t\in\mathbb{N}}$ is a φ-irreducible aperiodic T-chain and that compact sets are small sets.

Proposition 7. *The Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is a φ-irreducible aperiodic T-chain and compact sets of $\mathbb{R}_+^*$ are small sets.*

Proof. The function $p(\delta, \mathbf{w})$ defined in (37) is lower semi-continuous, and the function F defined in (38) is C^1.

We now prove that any point $\delta^* \in \mathbb{R}_+^*$ is a strongly globally attracting state, i.e. for all $\delta_0 \in \mathbb{R}_+^*$ and $\epsilon \in \mathbb{R}_+^*$ small enough there exists $t_0 \in \mathbb{N}^*$ such that for all $t \geq t_0$ there exists $\mathbf{w} \in O_{\delta_0,t}$ such that $F^t(\delta_0, \mathbf{w}) \in B(\delta^*, \epsilon)$. Let $\delta_0 \in \mathbb{R}_+^*$. Let $k \in \mathbb{N}^*$ be such that $\delta_0 \exp(k/(2d_\sigma)) > \delta^*$. We take $\mathbf{w}_i = \mathbf{0}$ for all $i \in [1..k]$ and define $\delta_k := F^k(\delta_0, \mathbf{w}_1, \ldots, \mathbf{w}_k)$. By construction of k, we have $\delta_k = \delta_0 \exp(-k/(2d_\sigma)(-1)) > \delta^*$. Now, take $\mathbf{u} = (-1, \ldots, -1)$ and note that the limit $\lim_{\alpha\to+\infty} F(\delta_k, \alpha\mathbf{u}) = 0$. Since the function F is continuous and that $F(\delta_k, 0\mathbf{u}) > \delta_k$, this means that the set $(0, \delta_k)$ is included into the image of the function $\alpha \mapsto F(\delta_k, \alpha\mathbf{u})$. And since $\delta^* < \delta_k$, there exists $\alpha_0 \in \mathbb{R}_+$ such that $F(\delta_k, \alpha_0\mathbf{u}) = \delta^*$. Now let $\bar{\mathbf{w}} = (\mathbf{w}_1, \ldots, \mathbf{w}_k, \alpha_0\mathbf{u})$, and note that since $g(\mathbf{u}) \geq 0$ and g is linear, $\alpha\mathbf{u} \in O_\delta = \{\mathbf{v} \in \mathbb{R}^n | \delta + g(\mathbf{v}) > 0\}$ for all $\alpha \in \mathbb{R}_+$ and all $\delta \in \mathbb{R}_+^*$; hence $\alpha_0\mathbf{u} \in O_{\delta_k}$ and $\mathbf{w}_i = 0\mathbf{u} \in O_\delta$ for all $\delta \in \mathbb{R}_+^*$. Therefore $\bar{\mathbf{w}} \in O_{\delta_0,k+1}$ and $F^{k+1}(\delta_0, \bar{\mathbf{w}}) = \delta^*$, so $\bar{\mathbf{w}}$ is a $k+1$-steps path from δ_0 to $B(\delta^*, \epsilon)$. As the proof stand for all k large enough, δ^* is a strongly globally attractive state.

We will now show that $F(0, \cdot)$ is a submersion at some point $\mathbf{w} \in \mathbb{R}^n$. To do so we compute the differential $D_\mathbf{w}F(0, \cdot)$ of $F(0, \cdot)$ at $\mathbf{w}$:

$$
\begin{aligned}
F(0, \mathbf{w}+\mathbf{h}) &= g(\mathbf{w}+\mathbf{h}) \exp\left(-\frac{1}{2d_\sigma}\left(\frac{\|\mathbf{w}+\mathbf{h}\|^2}{n} - 1\right)\right) \\
&= g(\mathbf{w}+\mathbf{h}) \exp\left(-\frac{1}{2d_\sigma}\left(\frac{\|\mathbf{w}\|^2 + 2\mathbf{w}.\mathbf{h} + \|\mathbf{h}\|^2}{n} - 1\right)\right) \\
&= g(\mathbf{w}+\mathbf{h}) \exp\left(-\frac{1}{2d_\sigma}\left(\frac{\|\mathbf{w}\|^2}{n} - 1\right)\right) \exp\left(-\frac{1}{2d_\sigma}\left(\frac{2\mathbf{w}.\mathbf{h} + \|\mathbf{h}\|^2}{n}\right)\right) \\
&= g(\mathbf{w}+\mathbf{h}) \exp\left(-\frac{1}{2d_\sigma}\left(\frac{\|\mathbf{w}\|^2}{n} - 1\right)\right)\left(1 - \frac{2\mathbf{w}.\mathbf{h}}{2d_\sigma n} + o(\|\mathbf{h}\|)\right) \\
&= F(0, \mathbf{w}) - F(0, \mathbf{w})\frac{\mathbf{w}.\mathbf{h}}{d_\sigma n} + g(\mathbf{h}) \exp\left(-\frac{1}{2d_\sigma}\left(\frac{\|\mathbf{w}\|^2}{n} - 1\right)\right) + o(\|\mathbf{h}\|) \ .
\end{aligned}
$$

Hence for $\mathbf{w} = (-\sqrt{n}, 0, \ldots, 0)$ and $\mathbf{h} = (0, \alpha, 0, \ldots, 0)$, $D_\mathbf{w}F(0, \cdot)(\mathbf{h}) = -\alpha \sin\theta \exp(0)$. Hence for α spanning $\mathbb{R}$, $D_\mathbf{w}F(0, \cdot)(\mathbf{h})$ spans $\mathbb{R}$ such that the image of $D_\mathbf{w}F(0, \cdot)$ equals $\mathbb{R}$, i.e. $D_\mathbf{w}F(0, \cdot)$ is surjective meaning $F(0, \cdot)$ is a submersion at $\mathbf{w}$. According to Lemma 5 this means there exists N an open neighbourhood of $(0, \mathbf{w})$ such that for all $(\delta, \mathbf{u}) \in N$, $F(\delta, \cdot)$ is a submersion at $\mathbf{u}$. So for $\delta^* \in \mathbb{R}_+^*$ small enough, $F(\delta^*, \cdot)$ is a submersion at $\mathbf{w} = (-\sqrt{n}, 0, \ldots, 0) \in O_{\delta^*}$.

Adding this with the fact that δ^* is a strongly globally attracting state, we can then apply Theorem 1 which concludes the proof. □

References

[1] Anne Auger and Nikolaus Hansen. On Proving Linear Convergence of Comparison-based Step-size Adaptive Randomized Search on Scaling-Invariant Functions via Stability of Markov Chains, 2013. ArXiv eprint.

[2] A. Chotard, A. Auger, and N. Hansen. Markov chain analysis of cumulative step-size adaptation on a linear constraint problem. *Evol. Comput.*, 2015.

[3] Alexandre Chotard, Anne Auger, and Nikolaus Hansen. Cumulative step-size adaptation on linear functions. In *Parallel Problem Solving from Nature - PPSN XII*, pages 72–81. Springer, september 2012.

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

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

[6] Bronislaw Jakubczyk and Eduardo D. Sontag. Controllability of nonlinear discrete time systems: A lie-algebraic approach. *SIAM J. Control Optim.*, 28:1–33, 1990.

[7] Francois Laudenbach. *Calcul différentiel et intégral*. Ecole Polytechnique, 2000.

[8] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, second edition, 1993.

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

[10] Frédéric Pham. Géométrie et calcul différentiel sur les variétés. *InterEditions, Paris*, 2002.

[11] SP Ponomarev. Submersions and preimages of sets of measure zero. *Siberian Mathematical Journal*, 28(1):153–163, 1987.

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

[13] Shlomo Sternberg. Lectures on differential geometry. Prentice-Hall Mathematics Series. Englewood Cliffs: Prentice-Hall, Inc. xi, 390 pp. (1964)., 1964.

Chapter 4

Analysis of Evolution Strategies

In this chapter we present analyses of different ESs optimizing a linear function with or without linear constraints. The aim of these analyses is to fully prove whether a given ES successfully optimizes these problems or not (which on a linear function translates into log-linear divergence), and to get a better understanding of how the different parameters of the ES or of the problem affect the behaviour of the ES on these problems.

Linear functions constitute an important class of problems which justify the focus of this work on them. Indeed, linear functions model when the distance between the mean of the sampling distribution $\boldsymbol{X}_t$ and the optimum is large compared to the step-size σ_t, as on a C^1 function the sets of equal values can then generally be approximated by hyperplanes, which correspond to the sets of equal values of linear functions. Hence, intuitively, linear functions need to be solved by diverging log-linearly in order for an ES to converge on other functions log-linearly independently of the initialization. Indeed, in [24] the log-linear convergence of the $(1+1)$-ES with 1/5 success rule [117] is proven on C^1 positively homogeneous functions (see (2.34) for a definition of positively homogeneous functions), under the condition that the step-size diverges on the linear function (more precisely that the expected inverse of the step-size change, $\mathbf{E}(\sigma_t/\sigma_{t+1})$, is strictly smaller than 1 on linear functions). In [2] the ES-IGO-flow (which can be linked to a continuous-time $(\mu/\mu_W, \lambda)$-ES when μ is proportional to λ and $\lambda \to \infty$, see 2.5.3) is shown to locally converge[1] on C^2 functions under two conditions. One of these conditions is that a variable which corresponds to the step-size of a standard ES diverges log-linearly on linear functions. Hence, log-linear divergence on linear functions appears to be a key to the log-linear convergence of ESs on a very wide range of problems.

In Section 4.1 we explain the methodology that we use to analyse ESs using Markov chain theory. In Section 4.2 we analyse the $(1, \lambda)$-CSA-ES on a linear function without constraints. In Section 4.3 we analyse a $(1, \lambda)$-ES on a linear function with a linear constraint; in 4.3.1 we both study a $(1, \lambda)$-ES with constant step-size and the $(1, \lambda)$-CSA-ES, and in 4.3.2 we study a $(1, \lambda)$-ES with constant step-size and with a general sampling distribution that can be non-Gaussian.

[1] According to private communications, log-linear convergence has been proven and is about to be published.

4.1 Markov chain Modelling of Evolution Strategies

Following [25] we present here our methodology and reasoning when analysing ESs using Markov chains on scaling invariant functions (see (2.33) for a definition of scaling invariant functions). We remind that linear functions, which will be the main object of study in this chapter, are scaling-invariant functions. Moreover many more functions are scaling-invariant, for instance all functions $g \circ f$ where $f : \mathbb{R}^n \to \mathbb{R}$ is a norm and $g : \mathbb{R} \to \mathbb{R}$ is strictly increasing are scaling invariant.

For a given ES optimizing a function $f : X \subset \mathbb{R}^n \to \mathbb{R}$ with optimum $\boldsymbol{x}^* \in X$, we would like to prove the almost sure log-linear convergence or divergence of the step-size σ_t to 0 or of the mean of the sampling distribution $\boldsymbol{X}_t$ to the optimum $\boldsymbol{x}^*$. This corresponds to the almost sure convergence of respectively

$$\frac{1}{t}\ln\left(\frac{\sigma_t}{\sigma_0}\right) = \frac{1}{t}\sum_{k=0}^{t-1}\ln\left(\frac{\sigma_{k+1}}{\sigma_k}\right) \xrightarrow[t\to+\infty]{a.s.} r \in \mathbb{R}^* \tag{4.1}$$

and

$$\frac{1}{t}\ln\left(\frac{\|\boldsymbol{X}_t - \boldsymbol{x}^*\|}{\|\boldsymbol{X}_0 - \boldsymbol{x}^*\|}\right) = \frac{1}{t}\sum_{k=0}^{t-1}\ln\left(\frac{\|\boldsymbol{X}_{k+1} - \boldsymbol{x}^*\|}{\|\boldsymbol{X}_k - \boldsymbol{x}^*\|}\right) \xrightarrow[t\to+\infty]{a.s.} r \in \mathbb{R}^* \tag{4.2}$$

to a rate $r \in \mathbb{R}^*$ ($r > 0$ corresponds to divergence, $r < 0$ to convergence). Expressing $1/t\ln(\sigma_t/\sigma_0)$ as the average of the terms $\ln(\sigma_{k+1}/\sigma_k)$ allows to apply a law of large numbers provided that each term $\ln(\sigma_{k+1}/\sigma_k)$ can be expressed as a function h of a positive Harris recurrent Markov chain $(\Phi_t)_{t\in\mathbb{N}}$ with invariant measure π, that is

$$\ln\left(\frac{\sigma_{k+1}}{\sigma_k}\right) = h(\Phi_{k+1}) \ . \tag{4.3}$$

Indeed, if $\pi(h)$ (which is defined in (1.18)) if finite, then according to 1.2.11 we have that

$$\frac{1}{t}\ln\left(\frac{\sigma_t}{\sigma_0}\right) = = \frac{1}{t}\sum_{k=1}^{t} h(\Phi_k) \xrightarrow[t\to+\infty]{a.s.} \pi(h) \ . \tag{4.4}$$

The same holds for the mean of the sampling distribution, provided that the terms $\ln(\|\boldsymbol{X}_{k+1} - \boldsymbol{x}^*\|/\|\boldsymbol{X}_k - \boldsymbol{x}^*\|)$ can be expressed as a function $\tilde{h}$ of a positive Harris recurrent Markov chain $(\tilde{\Phi}_t)_{t\in\mathbb{N}}$.

For example, for the step-size of the $(1,\lambda)$-CSA-ES, inductively from the update rule of σ_t (2.13), composing by ln and dividing by t

$$\frac{1}{t}\ln\left(\frac{\sigma_t}{\sigma_0}\right) = \frac{c_\sigma}{d_\sigma}\left(\frac{\frac{1}{t}\sum_{k=1}^{t}\|\boldsymbol{p}_k^\sigma\|}{\mathbf{E}\left(\|\mathcal{N}(\boldsymbol{0}, \boldsymbol{Id}_n)\|\right)} - 1\right) \ , \tag{4.5}$$

where $\boldsymbol{p}_t^\sigma$ is the evolution path defined in (2.12). By showing that $\|\boldsymbol{p}_t^\sigma\|$ is a function of a positive Harris Markov chain which is integrable under its invariant measure π (i.e. $\mathbf{E}_\pi(\|\boldsymbol{p}_t^\sigma\|) < \infty$), a law of large numbers could be applied to deduce the log-linear convergence or divergence

of the step-size. Note that we cannot always express the term $\ln(\sigma_{k+1}/\sigma_k)$ or $\ln(\|\boldsymbol{X}_{k+1} - \boldsymbol{x}^*\|/\|\boldsymbol{X}_k - \boldsymbol{x}^*\|)$ as a function of a positive Harris recurrent Markov chain. It will however typically be true on the linear functions studied in this chapter, and more generally on scaling-invariant functions [24, 25]. Following [25] we will give the expression of a suitable Markov-chain for scaling-invariant functions.

Let us define the state of an algorithm as the parameters of its sampling distribution (e.g. in the case of a multivariate Gaussian distribution, its mean $\boldsymbol{X}_t$, the step-size σ_t and the covariance matrix $\boldsymbol{C}_t$) combined with any other variable that is updated at each iteration (e.g. the evolution path $\boldsymbol{p}_t^\sigma$ for the $(1,\lambda)$-CSA-ES), and denote Θ the state space. The sequence of states $(\theta_t)_{t\in\mathbb{N}} \in \Theta^{\mathbb{N}}$ of an ES is naturally a Markov chain, as θ_{t+1} is a function of the previous state θ_t and of the new samples $(\boldsymbol{N}_t^i)_{i\in[1..\lambda]}$ (defined in (2.7)) through the selection and update rules. However, the convergence of the algorithm implies that the distribution of the step-size and the distribution of the mean of the sampling distribution converge to a Dirac distribution, which typically renders the Markov chain non-positive or non φ-irreducible. To apply a law of large numbers on our Markov chain as in (4.4), we need the Markov chain to be Harris recurrent which requires the Markov chain to be φ-irreducible to be properly defined.

Hence on scaling invariant functions instead of considering σ_t and $\boldsymbol{X}_t$ separately, as in [25] we usually consider them combined through the random vector $\boldsymbol{Z}_t := (\boldsymbol{X}_t - \boldsymbol{x}^*)/\sigma_t$, where $\boldsymbol{x}^*$ is the optimum of the function f optimized by the ES (usually assumed to be $\mathbf{0}$). The sequence $(\boldsymbol{Z}_t)_{t\in\mathbb{N}}$ is not always a Markov chain for a given objective function f or ES. It has however been shown in [25, Proposition 4.1] that on scaling invariant functions, given some conditions on the ES, $(\boldsymbol{Z}_t)_{t\in\mathbb{N}}$ is a time-homogeneous Markov chain which can be defined independently of $\boldsymbol{X}_t$ and σ_t. For a $(1,\lambda)$-CSA-ES with cumulation parameter $c_\sigma = 1$ (defined in 2.3.8 through (2.13) and (2.12)) we have the following update for $\boldsymbol{Z}_t$:

$$\boldsymbol{Z}_{t+1} = \frac{\boldsymbol{Z}_t + \boldsymbol{N}_t^{1:\lambda}}{\exp\left(\frac{1}{2d_\sigma}\left(\frac{\|\boldsymbol{N}_t^{1:\lambda}\|}{\mathbf{E}(\|\mathcal{N}(\mathbf{0},\boldsymbol{Id}_n)\|)} - 1\right)\right)}\ , \tag{4.6}$$

where $\boldsymbol{N}_t^{1:\lambda}$ is the step associated to the best sample $\boldsymbol{Y}_t^{1:\lambda}$, i.e. $\boldsymbol{N}_t^{1:\lambda} = (\boldsymbol{X}_t - \boldsymbol{Y}_t^{1:\lambda})/\sigma_t$. On scaling invariant functions, the step $\boldsymbol{N}_t^{1:\lambda}$ can be redefined only from $\boldsymbol{Z}_t$, independently of $\boldsymbol{X}_t$ and σ_t and hence $\boldsymbol{Z}_t$ can also be defined independently from $\boldsymbol{X}_t$ and σ_t and can be shown to be a Markov chain [25, Proposition 4.1]. This can be easily generalized to the $(1,\lambda)$-CSA-ES with $c_\sigma \in (0,1]$ on scaling-invariant functions to the sequence $(\boldsymbol{Z}_t, \boldsymbol{p}_t^\sigma)_{t\in\mathbb{N}}$ which is a Markov chain (although it is not proven here). The update for $\boldsymbol{Z}_t$ then writes

$$\boldsymbol{Z}_{t+1} = \frac{\boldsymbol{Z}_t + \boldsymbol{N}_t^{1:\lambda}}{\exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|\boldsymbol{p}_{t+1}^\sigma\|}{\mathbf{E}(\|\mathcal{N}(\mathbf{0},\boldsymbol{Id}_n)\|)} - 1\right)\right)}\ . \tag{4.7}$$

To show that the Markov chain $(\boldsymbol{Z}_t, \boldsymbol{p}_t^\sigma)_{t\in\mathbb{N}}$ is positive Harris recurrent, we first show the irreducibility of the Markov chain, and identify its small sets. This can be done by studying the

transition kernel P which here writes

$$P((\boldsymbol{z},\boldsymbol{p}),A\times B)=\int_{\mathbb{R}^n}\mathbf{1}_A\left(F_z(\boldsymbol{z},\boldsymbol{p},\boldsymbol{w})\right)\mathbf{1}_B\left(F_p(\boldsymbol{z},\boldsymbol{p},\boldsymbol{w})\right)p_{\boldsymbol{z},\boldsymbol{p}}(\boldsymbol{w})d\boldsymbol{w}\ , \tag{4.8}$$

where F_p is the function associated to (2.12)

$$F_p(\boldsymbol{z},\boldsymbol{p},\boldsymbol{w}):=(1-c_\sigma)\boldsymbol{p}+\sqrt{c_\sigma(2-c_\sigma)}\boldsymbol{w} \tag{4.9}$$

and where F_z is the function associated to (4.7)

$$F_z(\boldsymbol{z},\boldsymbol{p},\boldsymbol{w}):=\frac{\boldsymbol{z}+\boldsymbol{w}}{\exp\left(\frac{c_\sigma}{d_\sigma}\left(\frac{\|F_p(\boldsymbol{z},\boldsymbol{p},\boldsymbol{w})\|}{\mathbf{E}(\|\mathcal{N}(\mathbf{0},\boldsymbol{Id}_n)\|)}-1\right)\right)}\ , \tag{4.10}$$

and $p_{\boldsymbol{z},\boldsymbol{p}}(\boldsymbol{w})$ the conditional probability density function of $\boldsymbol{N}_t^{1:\lambda}$ knowing that $\boldsymbol{Z}_t=\boldsymbol{z}$ and $\boldsymbol{p}_t^\sigma=\boldsymbol{p}$ (note that $\boldsymbol{N}_t^{1:\lambda}$ is valued in $\mathbb{R}^n$, as indicated in (4.8)). In simple situations (no cumulation, optimizing a linear function), $\boldsymbol{z}$ and $\boldsymbol{p}$ can be moved out of the indicator functions by a change of variables, and the resulting expression can be used to build a non-trivial measure to show irreducibility, aperiodicity or small sets property (see Lemma 12 in 4.2.1, Proposition 3 in 4.3.1 and Proposition 2 in 4.3.2). However, this is a ad-hoc and tedious technique, and when the step-size is strongly coupled to the mean update it is very difficult[2]. The techniques developed in Chapter 3 could instead be used. In the example of a $(1,\lambda)$-CSA-ES, we define the transition function F as the combination of F_z and F_p, and we inductively define $F^1:=F$, and $F^{t+1}((\boldsymbol{z},\boldsymbol{p}),\boldsymbol{w}_1,\dots,\boldsymbol{w}_{t+1}):=F^t(F((\boldsymbol{z},\boldsymbol{p}),\boldsymbol{w}_1),\boldsymbol{w}_2,\dots,\boldsymbol{w}_{t+1})$, and $O_{(\boldsymbol{z},\boldsymbol{p}),t}$ is defined as the support of the distribution of $(\boldsymbol{N}_k^{1:\lambda})_{k\in[1..t]}$ conditionally to $(\boldsymbol{Z}_0,\boldsymbol{p}_0^\sigma)=(\boldsymbol{z},\boldsymbol{p})$. Provided that the transition function F is C^1, that the function $(\boldsymbol{z},\boldsymbol{p},\boldsymbol{w})\mapsto p_{\boldsymbol{z},\boldsymbol{p}}(\boldsymbol{w})$) is lower semi-continuous, and that there exists a strongly globally attracting state $(\boldsymbol{z}^*,\boldsymbol{p}^*)$ (see (3.3) for a definition) for which there exists $k\in\mathbb{N}^*$ and $\boldsymbol{w}^*\in O_{(\boldsymbol{z}^*,\boldsymbol{p}^*),k}$ such that $F^k((\boldsymbol{z}^*,\boldsymbol{p}^*),\cdot)$ is a submersion at $\boldsymbol{w}^*$, then the Markov chain $(\boldsymbol{Z}_t,\boldsymbol{p}_t^\sigma)_{t\in\mathbb{N}}$ is φ-irreducible, aperiodic, and compact sets are small sets for the Markov chain.

Once the irreducibility measure and the small sets are identified, the positivity and Harris recurrence of the Markov chain are proved using drift conditions defined in 1.2.10. In the case of the $(1,\lambda)$-CSA-ES with cumulation parameter c_σ equal to 1, we saw that on scaling invariant functions $(\boldsymbol{Z}_t)_{t\in\mathbb{N}}$ is a Markov chain. In this case the drift function V can be taken as $V(\boldsymbol{z})=\|\boldsymbol{z}\|^\alpha$, and the desired properties can be obtained by studying the limit of $\Delta V(\boldsymbol{z})/V(\boldsymbol{z})$ when $\|\boldsymbol{z}\|$ tends to infinity, as done in Proposition 6 in 4.3.1 on the linear function with a linear constraint. A negative limit shows the drift condition for geometric ergodicity, which not only allows us to apply a law of large numbers and ensures the convergence of Monte Carlo simulations to the value measured, but also ensures a fast convergence of these simulations. In the case of the $(1,\lambda)$-CSA-ES for any cumulation parameter $c_\sigma\in(0,1]$, on scaling invariant functions $(\boldsymbol{Z}_t,\boldsymbol{p}_t^\sigma)_{t\in\mathbb{N}}$ is a Markov chain and the natural extension of the drift function for $c_\sigma=1$ would be to consider $V(\boldsymbol{z},\boldsymbol{p})=\|\boldsymbol{z}\|^\alpha+\|\boldsymbol{p}\|^\beta$; however for large values of $\|\boldsymbol{z}\|$ and low values of $\|\boldsymbol{p}\|$, from (4.7) we see that $\|\boldsymbol{z}\|$ would be basically multiplied by $\exp(c_\sigma/d_\sigma)$ which

[2] Anne Auger, private communication, 2013.

makes $\|\boldsymbol{z}\|$ increase, and results in a positive drift, and so this drift function fails. The evolution path induces an inertia in the algorithm, and it may take several iterations for the algorithm to recover from a ill-initialized evolution path. Our intuition is that a drift function for the evolution path would therefore need to measure several steps into the future to see a negative drift.

4.2 Linear Function

In this section we present an analysis of the $(1,\lambda)$-CSA-ES on a linear function. This analysis is presented through a technical report [43] which contains the paper [46] which was published at the conference Parallel Problem Solving from Nature in 2012 and which includes the full proofs of every proposition of [46], and a proof of the log-linear divergence of $(|f(\boldsymbol{X}_t)|)_{t\in\mathbb{N}}$. This analysis investigates a slightly different step-size adaptation rule than the one defined in (2.13), and instead as proposed in [16] the step-size is adapted following

$$\sigma_{t+1} = \sigma_t \exp\left(\frac{c_\sigma}{2d_\sigma}\left(\frac{\|\boldsymbol{p}^\sigma_{t+1}\|^2}{n} - 1\right)\right) \ , \tag{4.11}$$

where as introduced in 2.3.8, $c_\sigma \in (0,1]$ is the cumulation parameter, $d_\sigma \in \mathbb{R}^*_+$ is the damping parameter, and $\boldsymbol{p}^\sigma_{t+1}$ as defined in (2.12) is the evolution path. This step-size adaptation rule is selected as it is easier to analyse, and similar to the original one.

An important point of this analysis is that on the linear function, the sequence of random vectors $(\boldsymbol{\xi}^\star_{t+1})_{t\in\mathbb{N}} := ((\boldsymbol{X}_{t+1} - \boldsymbol{X}_t)/\sigma_t)_{t\in\mathbb{N}}$ is i.i.d.. This implies that the evolution path $(\boldsymbol{p}^\sigma_t)_{t\in\mathbb{N}}$ is a time-homogeneous Markov chain. Since, as can be seen in (4.11), the distribution of the step-size is entirely determined by σ_0 and the evolution path $(\boldsymbol{p}^\sigma_t)_{t\in\mathbb{N}}$, this has the consequence that to prove the log-linear divergence of the step-size, we do not need to study the full Markov chain $(\boldsymbol{Z}_t, \boldsymbol{p}^\sigma_t)_{t\in\mathbb{N}}$, where $\boldsymbol{Z}_t$ is defined through (4.7), as proposed in Section 4.1. Instead, studying $(\boldsymbol{p}^\sigma_t)_{t\in\mathbb{N}}$ suffice. In the following article we establish that when $c_\sigma = 1$ and $\lambda \geq 3$ or when $c_\sigma < 1$ and $\lambda \geq 2$, the Markov chain $(\boldsymbol{p}^\sigma_t)_{t\in\mathbb{N}}$ is geometrically ergodic, from which we deduce that the step-size of the $(1,\lambda)$-CSA-ES diverges log-linearly almost surely at a rate that that we specify.

However to establish the log-linear divergence of $(|f(\boldsymbol{X}_t)|)_{t\in\mathbb{N}^*}$, we need to study the full Markov chain $(\boldsymbol{Z}_t, \boldsymbol{p}_t)_{t\in\mathbb{N}}$. While, for reasons suggested in Section 4.1, the analysis of this Markov is a difficult problem, we consider in the following technical report a simpler case where $c_\sigma = 1$ and so $\boldsymbol{p}_{t+1} = \boldsymbol{\xi}^\star_t$, and $(\boldsymbol{Z}_t)_{t\in\mathbb{N}}$ is a time-homogeneous Markov chain. We establish from studying $(\boldsymbol{Z}_t)_{t\in\mathbb{N}}$ that when c_σ equals 1, $(\boldsymbol{Z}_t)_{t\in\mathbb{N}}$ is a geometrically ergodic Markov chain and we derive almost sure log-linear divergence when $\lambda \geq 3$ of $(|f(\boldsymbol{X}_t)|)_{t\in\mathbb{N}}$ at the same rate than for the log-linear divergence of the step-size.

Furthermore a study of the variance of the logarithm of the step-size is conducted, and the scaling of this variance with the dimension gives elements regarding how to adapt the cumulation parameter c_σ with the dimension of the problem.

4.2.1 Paper: Cumulative Step-size Adaptation on Linear Functions

Cumulative Step-size Adaptation on Linear Functions: Technical Report

Alexandre Chotard[1], Anne Auger[1] and Nikolaus Hansen[1]

TAO team, INRIA Saclay-Ile-de-France, LRI, Paris-Sud University, France
firstname.lastname@lri.fr

Abstract. The CSA-ES is an Evolution Strategy with Cumulative Step size Adaptation, where the step size is adapted measuring the length of a so-called cumulative path. The cumulative path is a combination of the previous steps realized by the algorithm, where the importance of each step decreases with time. This article studies the CSA-ES on composites of strictly increasing functions with affine linear functions through the investigation of its underlying Markov chains. Rigorous results on the change and the variation of the step size are derived with and without cumulation. The step-size diverges geometrically fast in most cases. Furthermore, the influence of the cumulation parameter is studied.

Keywords: CSA, cumulative path, evolution path, evolution strategies, step-size adaptation

1 Introduction

Evolution strategies (ESs) are continuous stochastic optimization algorithms searching for the minimum of a real valued function $f : \mathbb{R}^n \to \mathbb{R}$. In the $(1, \lambda)$-ES, in each iteration, λ new children are generated from a single parent point $\boldsymbol{X}_t \in \mathbb{R}^n$ by adding a random Gaussian vector to the parent,

$$\boldsymbol{X} \in \mathbb{R}^n \mapsto \boldsymbol{X} + \sigma \boldsymbol{\mathcal{N}}(\mathbf{0}, \boldsymbol{C}) \ .$$

Here, $\sigma \in \mathbb{R}_+^*$ is called step-size and $\boldsymbol{C}$ is a covariance matrix. The best of the λ children, i.e. the one with the lowest f-value, becomes the parent of the next iteration. To achieve reasonably fast convergence, step size and covariance matrix have to be adapted throughout the iterations of the algorithm. In this paper, $\boldsymbol{C}$ is the identity and we investigate the so-called Cumulative Step-size Adaptation (CSA), which is used to adapt the step-size in the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) [13,10]. In CSA, a cumulative path is introduced, which is a combination of all steps the algorithm has made, where the importance of a step decreases exponentially with time. Arnold and Beyer studied the behavior of CSA on sphere, cigar and ridge functions [2,3,1,7] and on dynamical optimization problems where the optimum moves randomly [5] or linearly [6]. Arnold also studied the behaviour of a $(1, \lambda)$-ES on linear functions with linear constraint [4].

In this paper, we study the behaviour of the $(1, \lambda)$-CSA-ES on composites of strictly increasing functions with affine linear functions, e.g. $f : \boldsymbol{x} \mapsto \exp(x_2 - 2)$. Because

the CSA-ES is invariant under translation, under change of an orthonormal basis (rotation and reflection), and under strictly increasing transformations of the f-value, we investigate, w.l.o.g., $f : \boldsymbol{x} \mapsto x_1$. Linear functions model the situation when the current parent is far (here infinitely far) from the optimum of a smooth function. To be far from the optimum means that the distance to the optimum is large, *relative to the step-size* σ. This situation is undesirable and threatens premature convergence. The situation should be handled well, by increasing step widths, by any search algorithm (and is not handled well by the $(1, 2)$-σSA-ES [9]). Solving linear functions is also very useful to prove convergence independently of the initial state on more general function classes.

In Section 2 we introduce the $(1, \lambda)$-CSA-ES, and some of its characteristics on linear functions. In Sections 3 and 4 we study $\ln(\sigma_t)$ without and with cumulation, respectively. Section 5 presents an analysis of the variance of the logarithm of the step-size and in Section 6 we summarize our results.

Notations In this paper, we denote t the iteration or time index, n the search space dimension, $\mathcal{N}(0, 1)$ a standard normal distribution, i.e. a normal distribution with mean zero and standard deviation 1. The multivariate normal distribution with mean vector zero and covariance matrix identity will be denoted $\boldsymbol{\mathcal{N}}(\mathbf{0}, I_n)$, the i^{th} order statistic of λ standard normal distributions $\mathcal{N}_{i:\lambda}$, and $\Psi_{i:\lambda}$ its distribution. If $\boldsymbol{x} = (x_1, \cdots, x_n) \in \mathbb{R}^n$ is a vector, then $[x]_i$ will be its value on the i^{th} dimension, that is $[x]_i = x_i$. A random variable $\boldsymbol{X}$ distributed according to a law $\mathcal{L}$ will be denoted $\boldsymbol{X} \sim \mathcal{L}$. If A is a subset of $\mathcal{X}$, we will denote A^c its complement in $\mathcal{X}$.

2 The $(1, \lambda)$-CSA-ES

We denote with $\boldsymbol{X}_t$ the parent at the t^{th} iteration. From the parent point $\boldsymbol{X}_t$, λ children are generated: $\boldsymbol{Y}_{t,i} = \boldsymbol{X}_t + \sigma_t \boldsymbol{\xi}_{t,i}$ with $i \in [[1, \lambda]]$, and $\boldsymbol{\xi}_{t,i} \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, I_n)$, $(\boldsymbol{\xi}_{t,i})_{i \in [[1,\lambda]]}$ i.i.d. Due to the $(1, \lambda)$ selection scheme, from these children, the one minimizing the function f is selected: $\boldsymbol{X}_{t+1} = \operatorname{argmin}\{f(\boldsymbol{Y}), \boldsymbol{Y} \in \{\boldsymbol{Y}_{t,1}, ..., \boldsymbol{Y}_{t,\lambda}\}\}$. This latter equation implicitly defines the random variable $\boldsymbol{\xi}_t^\star$ as

$$\boldsymbol{X}_{t+1} = \boldsymbol{X}_t + \sigma_t \boldsymbol{\xi}_t^\star \ . \tag{1}$$

In order to adapt the step-size, the cumulative path is defined as

$$\boldsymbol{p}_{t+1} = (1 - c)\boldsymbol{p}_t + \sqrt{c(2 - c)}\, \boldsymbol{\xi}_t^\star \tag{2}$$

with $0 < c \leq 1$. The constant $1/c$ represents the life span of the information contained in $\boldsymbol{p}_t$, as after $1/c$ generations $\boldsymbol{p}_t$ is multiplied by a factor that approaches $1/e \approx 0.37$ for $c \to 0$ from below (indeed $(1-c)^{1/c} \leq \exp(-1)$). The typical value for c is between $1/\sqrt{n}$ and $1/n$. We will consider that $\boldsymbol{p}_0 \sim \boldsymbol{\mathcal{N}}(\mathbf{0}, I_n)$ as it makes the algorithm easier to analyze.

The normalization constant $\sqrt{c(2 - c)}$ in front of $\boldsymbol{\xi}_t^\star$ in Eq. (2) is chosen so that under random selection and if $\boldsymbol{p}_t$ is distributed according to $\boldsymbol{\mathcal{N}}(\mathbf{0}, I_n)$ then also $\boldsymbol{p}_{t+1}$ follows $\boldsymbol{\mathcal{N}}(\mathbf{0}, I_n)$. Hence the length of the path can be compared to the expected length of $\|\boldsymbol{\mathcal{N}}(\mathbf{0}, I_n)\|$ representing the expected length under random selection.

The step-size update rule increases the step-size if the length of the path is larger than the length under random selection and decreases it if the length is shorter than under random selection:

$$\sigma_{t+1} = \sigma_t \exp\left(\frac{c}{d_\sigma}\left(\frac{\|\boldsymbol{p}_{t+1}\|}{E(\|\boldsymbol{\mathcal{N}}(\mathbf{0}, I_n)\|)} - 1\right)\right)$$

where the damping parameter d_σ determines how much the step-size can change and is set to $d_\sigma = 1$. A simplification of the update considers the squared length of the path [5]:

$$\sigma_{t+1} = \sigma_t \exp\left(\frac{c}{2d_\sigma}\left(\frac{\|\boldsymbol{p}_{t+1}\|^2}{n} - 1\right)\right). \tag{3}$$

This rule is easier to analyse and we will use it throughout the paper. We will denote $\eta_t^\star$ the random variable for the step-size change, i.e. $\eta_t^\star = \exp(c/(2d_\sigma)(\|\boldsymbol{p}_{t+1}\|^2/n - 1))$, and for $\boldsymbol{u} \in \mathbb{R}^n$, $\eta^\star(\boldsymbol{u}) = \exp(c/(2d_\sigma)(\|u\|^2/n - 1))$.

Preliminary results on linear functions. Selection on the linear function, $f(\boldsymbol{x}) = [\boldsymbol{x}]_1$, is determined by $[\boldsymbol{X}_t]_1 + \sigma_t [\boldsymbol{\xi}_t^\star]_1 \leq [\boldsymbol{X}_t]_1 + \sigma_t \left[\boldsymbol{\xi}_{t,i}\right]_1$ for all i which is equivalent to $[\boldsymbol{\xi}_t^\star]_1 \leq \left[\boldsymbol{\xi}_{t,i}\right]_1$ for all i where by definition $\left[\boldsymbol{\xi}_{t,i}\right]_1$ is distributed according to $\mathcal{N}(0,1)$. Therefore the first coordinate of the selected step is distributed according to $\mathcal{N}_{1:\lambda}$ and all others coordinates are distributed according to $\mathcal{N}(0,1)$, i.e. selection does not bias the distribution along the coordinates $2, \ldots, n$. Overall we have the following result.

Lemma 1. *On the linear function $f(\boldsymbol{x}) = x_1$, the selected steps $(\boldsymbol{\xi}_t^\star)_{t\in\mathbb{N}}$ of the $(1,\lambda)$-ES are i.i.d. and distributed according to the vector $\boldsymbol{\xi} := (\mathcal{N}_{1:\lambda}, \mathcal{N}_2, \ldots, \mathcal{N}_n)$ where $\mathcal{N}_i \sim \mathcal{N}(0,1)$ for $i \geq 2$.*

Because the selected steps $\boldsymbol{\xi}_t^\star$ are i.i.d. the path defined in Eq. 2 is an autonomous Markov chain, that we will denote $\boldsymbol{\mathcal{P}} = (\boldsymbol{p}_t)_{t\in\mathbb{N}}$. Note that if the distribution of the selected step depended on $(\boldsymbol{X}_t, \sigma_t)$ as it is generally the case on non-linear functions, then the path alone would not be a Markov Chain, however $(\boldsymbol{X}_t, \sigma_t, \boldsymbol{p}_t)$ would be an autonomous Markov Chain. In order to study whether the $(1,\lambda)$-CSA-ES diverges geometrically, we investigate the log of the step-size change, whose formula can be immediately deduced from Eq. 3:

$$\ln\left(\frac{\sigma_{t+1}}{\sigma_t}\right) = \frac{c}{2d_\sigma}\left(\frac{\|\boldsymbol{p}_{t+1}\|^2}{n} - 1\right) \tag{4}$$

By summing up this equation from 0 to $t-1$ we obtain

$$\frac{1}{t}\ln\left(\frac{\sigma_t}{\sigma_0}\right) = \frac{c}{2d_\sigma}\left(\frac{1}{t}\sum_{k=1}^{t}\frac{\|\boldsymbol{p}_k\|^2}{n} - 1\right) \quad . \tag{5}$$

We are interested to know whether $\frac{1}{t}\ln(\sigma_t/\sigma_0)$ converges to a constant. In case this constant is positive this will prove that the $(1,\lambda)$-CSA-ES diverges geometrically. We recognize thanks to (5) that this quantity is equal to the sum of t terms divided by t that suggests the use of the law of large numbers to prove convergence of (5). We will start by investigating the case without cumulation $c = 1$ (Section 3) and then the case with cumulation (Section 4).

3 Divergence rate of $(1, \lambda)$-CSA-ES without cumulation

In this section we study the $(1, \lambda)$-CSA-ES without cumulation, i.e. $c = 1$. In this case, the path always equals to the selected step, i.e. for all t, we have $\boldsymbol{p}_{t+1} = \boldsymbol{\xi}_t^\star$. We have proven in Lemma 1 that $\boldsymbol{\xi}_t^\star$ are i.i.d. according to $\boldsymbol{\xi}$. This allows us to use the standard law of large numbers to find the limit of $\frac{1}{t}\ln(\sigma_t/\sigma_0)$ as well as compute the expected log-step-size change.

Proposition 1. *Let* $\Delta_\sigma := \frac{1}{2d_\sigma n}\left(\mathbb{E}\left(\mathcal{N}_{1:\lambda}^2\right) - 1\right)$. *On linear functions, the* $(1, \lambda)$*-CSA-ES without cumulation satisfies (i) almost surely* $\lim_{t\to\infty} \frac{1}{t}\ln\left(\sigma_t/\sigma_0\right) = \Delta_\sigma$, *and (ii) for all* $t \in \mathbb{N}$, $\mathbb{E}(\ln(\sigma_{t+1}/\sigma_t)) = \Delta_\sigma$.

Proof. We have identified in Lemma 1 that the first coordinate of $\boldsymbol{\xi}_t^\star$ is distributed according to $\mathcal{N}_{1:\lambda}$ and the other coordinates according to $\mathcal{N}(0,1)$, hence $\mathbb{E}(\|\boldsymbol{\xi}_t^\star\|^2) = \mathbb{E}([\boldsymbol{\xi}_t^\star]_1{}^2) + \sum_{i=2}^n \mathbb{E}([\boldsymbol{\xi}_t^\star]_i^2) = \mathbb{E}(\mathcal{N}_{1:\lambda}^2) + n - 1$. Therefore $\mathbb{E}(\|\boldsymbol{\xi}_t^\star\|^2)/n - 1 = (\mathbb{E}(\mathcal{N}_{1:\lambda}^2) - 1)/n$. By applying this to Eq. (4), we deduce that $\mathbb{E}(\ln(\sigma_{t+1}/\sigma_t) = 1/(2d_\sigma n)(\mathbb{E}(\mathcal{N}_{1:\lambda}^2) - 1)$. Furthermore, as $\mathbb{E}(\mathcal{N}_{1:\lambda}^2) \leq \mathbb{E}((\lambda\mathcal{N}(0,1))^2) = \lambda^2 < \infty$, we have $\mathbb{E}(\|\boldsymbol{\xi}_t^\star\|^2) < \infty$. The sequence $(\|\boldsymbol{\xi}_t^\star\|^2)_{t\in\mathbb{N}}$ being i.i.d according to Lemma 1, and being integrable as we just showed, we can apply the strong law of large numbers on Eq. (5). We obtain

$$\begin{aligned}\frac{1}{t}\ln\left(\frac{\sigma_t}{\sigma_0}\right) &= \frac{1}{2d_\sigma}\left(\frac{1}{t}\sum_{k=0}^{t-1}\frac{\|\boldsymbol{\xi}_k^\star\|^2}{n} - 1\right) \\ &\xrightarrow[t\to\infty]{a.s.} \frac{1}{2d_\sigma}\left(\frac{\mathbb{E}\left(\|\boldsymbol{\xi}^\star\|^2\right)}{n} - 1\right) = \frac{1}{2d_\sigma n}\left(\mathbb{E}\left(\mathcal{N}_{1:\lambda}^2\right) - 1\right)\end{aligned}$$

□

The proposition reveals that the sign of $\mathbb{E}(\mathcal{N}_{1:\lambda}^2) - 1)$ determines whether the step-size diverges to infinity or converges to 0. In the following, we show that $\mathbb{E}(\mathcal{N}_{1:\lambda}^2)$ increases in λ for $\lambda \geq 2$ and that the $(1, \lambda)$-ES diverges for $\lambda \geq 3$. For $\lambda = 1$ and $\lambda = 2$, the step-size follows a random walk on the log-scale. To prove this we need the following lemma:

Lemma 2 ([11]). *Let* $g : \mathbb{R} \to \mathbb{R}$ *be a function, and* $(N_i)_{i\in[1..\lambda]}$ *be a sequence of i.i.d. random variables, and let* $N_{i:\lambda}$ *denote the* i^{th} *order statistic of the sequence* $(N_i)_{i\in[1..\lambda]}$. *For* $\lambda \in \mathbb{N}^*$,

$$(\lambda + 1)\,\mathbb{E}\left(g\left(N_{1:\lambda}\right)\right) = \mathbb{E}\left(g\left(N_{2:\lambda+1}\right)\right) + \lambda\mathbb{E}\left(g\left(N_{1:\lambda+1}\right)\right) \quad . \tag{6}$$

Proof. This method can be found with more details in [11].

Let $\chi_i = g(N_i)$, and $\chi_{i:\lambda} = g(N_{i:\lambda})$. Note that in general $\chi_{1:\lambda} \neq \min_{i\in[[1,\lambda]]}\chi_i$. The sorting is made on $(N_i)_{i\in[1..\lambda]}$, not on $(\chi_i)_{i\in[1..\lambda]}$. We will also note $\chi_{i:\lambda}^{\{j\}}$ the i^{th} order statistic after that the variable χ_j has been taken away, and $\chi_{i:\lambda}^{[j]}$ the i^{th} order statistic after $\chi_{j:\lambda}$ has been taken away. If $i \neq 1$ then we have $\chi_{1:\lambda}^{[i]} = \chi_{1:\lambda}$, and for $i = 1$ and $\lambda \geq 2$, $\chi_{1:\lambda}^{[i]} = \chi_{2:\lambda}$.

We have that (i) $\mathbb{E}(\chi_{1:\lambda}^{\{i\}}) = \mathbb{E}(\chi_{1:\lambda-1})$, and (ii) $\sum_{i=1}^\lambda \chi_{1:\lambda}^{\{i\}} = \sum_{i=1}^\lambda \chi_{1:\lambda}^{[i]}$. From (i) we deduce that $\lambda\mathbb{E}\left(\chi_{1:\lambda-1}\right) = \lambda\mathbb{E}(\chi_{1:\lambda}^{\{i\}}) = \sum_{i=1}^\lambda \mathbb{E}(\chi_{1:\lambda}^{\{i\}}) = \mathbb{E}(\sum_{i=1}^\lambda \chi_{1:\lambda}^{\{i\}})$. With (ii),

we get that $\mathbb{E}(\sum_{i=1}^{\lambda} \chi_{1:\lambda}^{\{i\}}) = \mathbb{E}(\sum_{i=1}^{\lambda} \chi_{1\lambda}^{[i]}) = \mathbb{E}(\chi_{2:\lambda}) + (\lambda-1)\mathbb{E}(\chi_{1:\lambda})$. By combining both, we get

$$\lambda\mathbb{E}(\chi_{1:\lambda-1}) = \mathbb{E}(\chi_{2:\lambda}) + (\lambda-1)\mathbb{E}(\chi_{1:\lambda}) \ .$$

□

We are now ready to prove the following result.

Lemma 3. *Let $(\mathcal{N}_i)_{i\in[[1,\lambda]]}$ be independent random variables, distributed according to $\mathcal{N}(0,1)$, and $\mathcal{N}_{i:\lambda}$ the i^{th} order statistic of $(\mathcal{N}_i)_{i\in[[1,\lambda]]}$. Then $\mathbb{E}\left(\mathcal{N}_{1:1}^2\right) = \mathbb{E}\left(\mathcal{N}_{1:2}^2\right) = 1$. In addition, for all $\lambda \geq 2$, $\mathbb{E}\left(\mathcal{N}_{1:\lambda+1}^2\right) > \mathbb{E}\left(\mathcal{N}_{1:\lambda}^2\right)$.*

Proof. For $\lambda = 1$, $\mathcal{N}_{1:1} = \mathcal{N}_1$ and so $\mathbb{E}(\mathcal{N}_{1:1}^2) = 1$. So using Lemma 2 and taking g as the square function, $\mathbb{E}(\mathcal{N}_{1:2}^2) + \mathbb{E}(\mathcal{N}_{1:1}^2) = 2\mathbb{E}(\mathcal{N}_{1:1}^2) = 2$. By symmetry of the standard normal distribution, $\mathbb{E}(\mathcal{N}_{1:2}^2) = \mathbb{E}(\mathcal{N}_{1:2}^2)$, and so $\mathbb{E}(\mathcal{N}_{1:2}^2) = 1$.

Now for $\lambda \geq 2$, using Lemma 2 and taking g as the square function, $(\lambda+1)\mathbb{E}(\mathcal{N}_{1:\lambda}^2) = \mathbb{E}(\mathcal{N}_{2:\lambda+1}^2) + \lambda\mathbb{E}(\mathcal{N}_{1:\lambda+1}^2)$, and so

$$(\lambda+1)(\mathbb{E}(\mathcal{N}_{1:\lambda}^2) - \mathbb{E}(\mathcal{N}_{1:\lambda+1}^2)) = \mathbb{E}(\mathcal{N}_{2:\lambda+1}^2) - \mathbb{E}(\mathcal{N}_{1:\lambda+1}^2) \ . \tag{7}$$

Hence $\mathbb{E}(\mathcal{N}_{1:\lambda+1}^2) > \mathbb{E}(\mathcal{N}_{1:\lambda}^2)$ for $\lambda \geq 2$ is equivalent to $\mathbb{E}(\mathcal{N}_{1:\lambda}^2) > \mathbb{E}\left(\mathcal{N}_{2:\lambda}^2\right)$ for $\lambda \geq 3$.

For $\boldsymbol{\omega} \in \mathbb{R}^\lambda$ let $\boldsymbol{\omega}_{i:\lambda} \in \mathbb{R}$ denote the i^{th} order statistic of the sequence $([\boldsymbol{\omega}]_j)_{j\in[1..\lambda]}$. Let p be the density of the sequence $(\mathcal{N}_i)_{i\in[1..\lambda]}$, i.e. $p(\boldsymbol{\omega}) := \exp(-\|\boldsymbol{\omega}\|^2/2)/\sqrt{2\pi}$ and let E_1 be the set $\{\boldsymbol{\omega} \in \mathbb{R}^\lambda | \boldsymbol{\omega}_{1:\lambda}^2 < \boldsymbol{\omega}_{2:\lambda}^2\}$. For $\boldsymbol{\omega} \in E_1$, $|\boldsymbol{\omega}_{1:\lambda}| < |\boldsymbol{\omega}_{2:\lambda}|$, and since $\boldsymbol{\omega}_{1:\lambda} < \boldsymbol{\omega}_{2:\lambda}$, we have $-\boldsymbol{\omega}_{2:\lambda} < \boldsymbol{\omega}_{1:\lambda} < \boldsymbol{\omega}_{2:\lambda}$. For $\boldsymbol{\omega} \in E_1$, take $\bar{\boldsymbol{\omega}} \in \mathbb{R}^\lambda$ such that $\bar{\boldsymbol{\omega}}_{1:\lambda} = -\boldsymbol{\omega}_{2:\lambda}$, $\bar{\boldsymbol{\omega}}_{2:\lambda} = \boldsymbol{\omega}_{1:\lambda}$, and $[\bar{\boldsymbol{\omega}}]_i = [\boldsymbol{\omega}]_i$ for $i \in [1..\lambda]$ such that $\boldsymbol{\omega}_{2:\lambda} < [\boldsymbol{\omega}]_i$. The function $g : E_1 \to \mathbb{R}^\lambda$ that maps $\boldsymbol{\omega}$ to $\bar{\boldsymbol{\omega}}$ is a diffeomorphism from E_1 to its image by g, that we denote E_2, and the Jacobian determinant of g is 1. Note that by symmetry of p, $p(\bar{\boldsymbol{\omega}}) = p(\boldsymbol{\omega})$, hence with the change of variables $\bar{\boldsymbol{\omega}} = g(\boldsymbol{\omega})$,

$$\int_{E_2} (\bar{\boldsymbol{\omega}}_{2:\lambda}^2 - \bar{\boldsymbol{\omega}}_{1:\lambda}^2) p(\bar{\boldsymbol{\omega}}_{i:\lambda}) d\bar{\boldsymbol{\omega}} = \int_{E_1} (\boldsymbol{\omega}_{1:\lambda}^2 - \boldsymbol{\omega}_{2:\lambda}^2) p(\boldsymbol{\omega}) d\boldsymbol{\omega} \ . \tag{8}$$

Since $\mathbb{E}(\mathcal{N}_{1:\lambda}^2 - \mathcal{N}_{2:\lambda}^2) = \int_{\mathbb{R}^\lambda} (\boldsymbol{\omega}_{1:\lambda}^2 - \boldsymbol{\omega}_{2:\lambda}^2) p(\boldsymbol{\omega}) d\boldsymbol{\omega}$ and E_1 and E_2 being disjoint sets, with (8)

$$\mathbb{E}(\mathcal{N}_{1:\lambda}^2 - \mathcal{N}_{2:\lambda}^2) = \int_{\mathbb{R}^\lambda \setminus (E_1 \cup E_2)} (\boldsymbol{\omega}_{1:\lambda}^2 - \boldsymbol{\omega}_{2:\lambda}^2) p(\boldsymbol{\omega}) d\boldsymbol{\omega} \ . \tag{9}$$

Since $p(\boldsymbol{\omega}) > 0$ for all $\boldsymbol{\omega} \in \mathbb{R}^\lambda$ and that $\boldsymbol{\omega}_{1:\lambda}^2 - \boldsymbol{\omega}_{2:\lambda}^2 \leq 0$ if and only if $\boldsymbol{\omega} \in E_1$ or $\boldsymbol{\omega}_{1:\lambda}^2 = \boldsymbol{\omega}_{2:\lambda}^2$, Eq. (9) shows that $\mathbb{E}(\mathcal{N}_{1:\lambda}^2 - \mathcal{N}_{2:\lambda}^2) > 0$ if and only if there exists a subset of $\mathbb{R}^\lambda \setminus (E_1 \cup E_2 \cup \{\boldsymbol{\omega} \in \mathbb{R}^\lambda | \boldsymbol{\omega}_{1:\lambda}^2 = \boldsymbol{\omega}_{2:\lambda}^2\})$ with positive Lebesgue-measure. For $\lambda \geq 3$, the set $E_3 := \{\boldsymbol{\omega} \in \mathbb{R}^\lambda | \boldsymbol{\omega}_{1:\lambda} < \boldsymbol{\omega}_{2:\lambda} < \boldsymbol{\omega}_{3:\lambda} < 0\}$ has positive Lebesgue measure. For all $\boldsymbol{\omega} \in E_3$, $\boldsymbol{\omega}_{1:\lambda}^2 > \boldsymbol{\omega}_{2:\lambda}^2$ so $E_3 \cap E_1 = \emptyset$. Furthermore, $\boldsymbol{\omega}_{1:\lambda} \neq \boldsymbol{\omega}_{2:\lambda}$ so $E_3 \cap \{\boldsymbol{\omega} \in \mathbb{R}^\lambda | \boldsymbol{\omega}_{1:\lambda}^2 = \boldsymbol{\omega}_{2:\lambda}^2\} = \emptyset$. Finally for $\boldsymbol{\omega} \in E_3$, denoting $\bar{\boldsymbol{\omega}} = g^{-1}(\boldsymbol{\omega})$, since $\bar{\boldsymbol{\omega}}_{1:\lambda} = \boldsymbol{\omega}_{2:\lambda}$ and $\bar{\boldsymbol{\omega}}_{2:\lambda} = \boldsymbol{\omega}_{3:\lambda}$, $\bar{\boldsymbol{\omega}}_{1:\lambda}^2 > \bar{\boldsymbol{\omega}}_{2:\lambda}^2$ and so $\bar{\boldsymbol{\omega}}_{i:\lambda} \notin E_1$, that is $\boldsymbol{\omega} \notin E_2$. So E_3 is a subset of $\mathbb{R}^\lambda \setminus (E_1 \cup E_2 \cup \{\boldsymbol{\omega} \in \mathbb{R}^\lambda | \boldsymbol{\omega}_{1:\lambda}^2 = \boldsymbol{\omega}_{2:\lambda}^2\}$ with positive Lebesgue measure, which proves the lemma. □

We can now link Proposition 1 and Lemma 3 into the following theorem:

Theorem 1. *On linear functions, for $\lambda \geq 3$, the step-size of the $(1, \lambda)$-CSA-ES without cumulation ($c = 1$) diverges geometrically almost surely and in expectation at the rate $1/(2d_\sigma n)(\mathbb{E}(\mathcal{N}_{1:\lambda}^2) - 1)$, i.e.*

$$\frac{1}{t} \ln\left(\frac{\sigma_t}{\sigma_0}\right) \xrightarrow[t\to\infty]{a.s.} \mathbb{E}\left(\ln\left(\frac{\sigma_{t+1}}{\sigma_t}\right)\right) = \frac{1}{2d_\sigma n}\left(\mathbb{E}\left(\mathcal{N}_{1:\lambda}^2\right) - 1\right) \quad . \tag{10}$$

For $\lambda = 1$ and $\lambda = 2$, without cumulation, the logarithm of the step-size does an additive unbiased random walk i.e. $\ln \sigma_{t+1} = \ln \sigma_t + W_t$ where $E[W_t] = 0$. More precisely $W_t \sim 1/(2d_\sigma)(\chi_n^2/n - 1)$ for $\lambda = 1$, and $W_t \sim 1/(2d_\sigma)((\mathcal{N}_{1:2}^2 + \chi_{n-1}^2)/n - 1)$ for $\lambda = 2$, where χ_k^2 stands for the chi-squared distribution with k degree of freedom.

Proof. For $\lambda > 2$, from Lemma 3 we know that $\mathbb{E}(\mathcal{N}_{1:\lambda}^2) > \mathbb{E}(\mathcal{N}_{1:2}^2) = 1$. Therefore $\mathbb{E}(\mathcal{N}_{1:\lambda}^2) - 1 > 0$, hence Eq. (10) is strictly positive, and with Proposition 1 we get that the step-size diverges geometrically almost surely at the rate $1/(2d_\sigma)(\mathbb{E}(\mathcal{N}_{1:\lambda}^2) - 1)$.

With Eq. 4 we have $\ln(\sigma_{t+1}) = \ln(\sigma_t) + W_t$, with $W_t = 1/(2d_\sigma)(\|\boldsymbol{\xi}_t^\star\|^2/n - 1)$. For $\lambda = 1$ and $\lambda = 2$, according to Lemma 3, $\mathbb{E}(W_t) = 0$. Hence $\ln(\sigma_t)$ does an additive unbiased random walk. Furthermore $\|\boldsymbol{\xi}\|^2 = \mathcal{N}_{1:\lambda}^2 + \chi_{n-1}^2$, so for $\lambda = 1$, since $\mathcal{N}_{1:1} = \mathcal{N}(0, 1)$, $\|\boldsymbol{\xi}\|^2 = \chi_n^2$. □

3.1 Geometric divergence of $([\boldsymbol{X}_t]_1)_{t\in\mathbb{N}}$

We now establish a result similar to Theorem 1 for the sequence $([X_t]_1)_{t\in\mathbb{N}}$. Using Eq (1)

$$\ln\left|\frac{[\boldsymbol{X}_{t+1}]_1}{[\boldsymbol{X}_t]_1}\right| = \ln\left|1 + \frac{\sigma_t}{[\boldsymbol{X}_t]_1}[\boldsymbol{\xi}_t^\star]_1\right| \quad .$$

Summing the previous equation from 0 till $t - 1$ and dividing by t gives that

$$\frac{1}{t}\ln\left|\frac{[\boldsymbol{X}_t]_1}{[\boldsymbol{X}_0]_1}\right| = \frac{1}{t}\sum_{k=0}^{t-1}\ln\left|1 + \frac{\sigma_k}{[\boldsymbol{X}_k]_1}[\boldsymbol{\xi}_t^\star]_1\right| \quad . \tag{11}$$

Let $Z_t = \frac{[\boldsymbol{X}_{t+1}]_1}{\sigma_t}$ for $t \in \mathbb{N}$, then

$$Z_{t+1} = \frac{[\boldsymbol{X}_{t+2}]_1}{\sigma_{t+1}} = \frac{[\boldsymbol{X}_{t+1}]_1 + \sigma_{t+1}\left[\boldsymbol{\xi}_{t+1}^\star\right]_1}{\sigma_{t+1}}$$
$$Z_{t+1} = \frac{Z_t}{\eta_t^\star} + \left[\boldsymbol{\xi}_{t+1}^\star\right]_1$$

using that $\sigma_{t+1} = \sigma_t \eta_t^\star$. According to Lemma 1, $(\boldsymbol{\xi}_t^\star)_{t\in\mathbb{N}}$ is a i.i.d. sequence. As $\eta_t^\star = \exp((\|\boldsymbol{\xi}_t^\star\|^2/n - 1)/(2d_\sigma))$, $(\eta_t^\star)_{t\in\mathbb{N}}$ is also independent over time. Therefore, $\mathcal{Z} := (Z_t)_{t\in\mathbb{N}}$, is a Markov chain.

By introducing $\mathcal{Z}$ in Eq (11), we obtain:

$$\begin{aligned}\frac{1}{t}\ln\left|\frac{[\boldsymbol{X}_t]_1}{[\boldsymbol{X}_0]_1}\right| &= \frac{1}{t}\sum_{k=0}^{t-1}\ln\left|1+\frac{\sigma_{k-1}\eta^\star_{k-1}}{[\boldsymbol{X}_k]_1}[\boldsymbol{\xi}^\star_k]_1\right| \\ &= \frac{1}{t}\sum_{k=0}^{t-1}\ln\left|1+\frac{\eta^\star_{k-1}}{Z_{k-1}}[\boldsymbol{\xi}^\star_k]_1\right| \\ &= \frac{1}{t}\sum_{k=0}^{t-1}\ln\left|\frac{\frac{Z_{k-1}}{\eta^\star_{k-1}}+\boldsymbol{\xi}^\star_k}{\frac{Z_{k-1}}{\eta^\star_{k-1}}}\right| \\ &= \frac{1}{t}\sum_{k=0}^{t-1}\left(\ln|Z_k|-\ln|Z_{k-1}|+\ln\left|\eta^\star_{k-1}\right|\right) \end{aligned} \tag{12}$$

The right hand side of this equation reminds us again of the law of large numbers. The sequence $(Z_t)_{t\in\mathbb{N}}$ is not independent over time, but $\mathcal{Z}$ being a Markov chain, if it follows some specific stability properties of Markov chains, then a law of large numbers may apply.

Study of the Markov chain $\mathcal{Z}$ To apply a law of large numbers to a Markov chain $\boldsymbol{\Phi} = (\Phi_t)_{t\in\mathbb{N}}$ taking values in X a subset of $\mathbb{R}^n$, it has to satisfies some stability properties: in particular, the Markov chain $\boldsymbol{\Phi}$ has to be φ-irreducible, that is, there exists a measure φ such that every Borel set A of X with $\varphi(A) > 0$ has a positive probability to be reached in a finite number of steps by $\boldsymbol{\Phi}$ starting from any $\boldsymbol{x} \in \mathbb{R}^n$, i.e. $\Pr(\Phi_t \in A|\Phi_0 = \boldsymbol{x}) > 0$ for all $\boldsymbol{x} \in X$. In addition, the chain $\boldsymbol{\Phi}$ needs to be (i) positive, that is the chain admits an invariant probability measure π, i.e., for any Borel set A, $\pi(A) = \int_X P(\boldsymbol{x}, A)\pi(d\boldsymbol{x})$ with $P(\boldsymbol{x}, A) := \Pr(\Phi_1 \in A|\Phi_0 = \boldsymbol{x})$, and (ii) Harris recurrent which means for any Borel set A such that $\varphi(A) > 0$, the chain $\boldsymbol{\Phi}$ visits A an infinite number of times with probability one. Under those conditions, $\boldsymbol{\Phi}$ satisfies a law of large numbers as written in the following lemma.

Lemma 4. *[12, 17.0.1] Suppose that $\boldsymbol{\Phi}$ is a positive Harris chain defined on a set X with stationary measure π, and let $g : X \to \mathbb{R}$ be a π-integrable function, i.e. such that $\pi(|g|) := \int_X |g(\boldsymbol{x})|\pi(d\boldsymbol{x})$ is finite. Then*

$$1/t\sum_{k=1}^{t} g(\boldsymbol{\Phi}_k) \xrightarrow[t\to\infty]{a.s} \pi(g) \ . \tag{13}$$

To show that a φ-irreducible Markov defined on a set $X \subset \mathbb{R}^n$ equipped with its Borel σ-algebra $\mathcal{B}(X)$ is positive Harris recurrent, we generally show that the chain follows a so-called drift condition over a small set, that is for a function V, an inequality over the drift operator $\Delta V : \boldsymbol{x} \in X \mapsto \int_X V(\boldsymbol{y})P(\boldsymbol{x}, d\boldsymbol{y}) - V(\boldsymbol{x})$. A small set $C \in \mathcal{B}(X)$ is a set such that there exists a $m \in \mathbb{N}^*$ and a non-trivial measure ν_m on $\mathcal{B}(X)$ such that for all $\boldsymbol{x} \in C$, $B \in \mathcal{B}(X)$, $P^m(\boldsymbol{x}, B) \geq \nu_m(B)$. The set C is then called a ν_m-small set. The chain also needs to be aperiodic, meaning for all sequence $(D_i)_{i\in[0..d-1]} \in \mathcal{B}(X)^d$ of disjoint sets such that for $\boldsymbol{x} \in D_i$,

$P(\boldsymbol{x}, D_{i+1 \mod d}) = 1$, and $[\cup_{i=1}^{d}]^c$ is φ-negligible, d equals 1. If there exists a ν_1-small-set C such that $\nu_1(C) > 0$, then the chain is strongly aperiodic (and therefore aperiodic). We then have the following lemma.

Lemma 5. *[12, 14.0.1] Suppose that the chain $\boldsymbol{\Phi}$ is φ-irreductible and aperiodic, and $f \geq 1$ a function on X. Let us assume that there exists V some extended-valued non-negative function finite for some $x_0 \in X$, a small set C and $b \in \mathbb{R}$ such that*

$$\Delta V(x) \leq -f(x) + b\mathbf{1}_C(x) \ , x \in X. \tag{14}$$

Then the chain $\boldsymbol{\Phi}$ is positive Harris recurrent with invariant probability measure π and

$$\pi(f) = \int_X \pi(dx) f(x) < \infty \ . \tag{15}$$

Proving the irreducibility, aperiodicity and exhibiting the small sets of a Markov chain $\boldsymbol{\Phi}$ can be done by showing some properties of its underlying control model. In our case, the model associated to $\mathcal{Z}$ is called a non-linear state space model. We will, in the following, define this non-linear state space model and some of its properties.

Suppose there exists $O \in \mathbb{R}^m$ an open set and $F : X \times O \to X$ a smooth function (C^∞) such that $\Phi_{t+1} = F(\Phi_t, \boldsymbol{W}_{t+1})$ with $(\boldsymbol{W}_t)_{t\in\mathbb{N}}$ being a sequence of i.i.d. random variables, whose distribution Γ possesses a semi lower-continuous density γ_w which is supported on an open set O_w; then $\boldsymbol{\Phi}$ follows a non-linear state space model driven by F or NSS(F) model, with control set O_w. We define its associated control model CM(F) as the deterministic system $\boldsymbol{x}_t = F_t(\boldsymbol{x}_0, \boldsymbol{u}_1, \cdots, \boldsymbol{u}_t)$, where F_t is inductively defined by $F_1 := F$ and

$$F_t(\boldsymbol{x}_0, \boldsymbol{u}_1, \cdots, \boldsymbol{u}_k) := F(F_{t-1}(\boldsymbol{x}_0, \boldsymbol{u}_1, \cdots, \boldsymbol{u}_{t-1}), \boldsymbol{u}_t) \ ,$$

provided that $(\boldsymbol{u}_t)_{t\in\mathbb{N}}$ lies in the control set O_w. For a point $\boldsymbol{x} \in X$, and $k \in \mathbb{N}^*$ we define

$$A_+^k(\boldsymbol{x}) := \{F_k(\boldsymbol{x}, u_1, \cdots, u_k) | u_i \in O_w, \ \ \forall i \in [1..k]\} \ ,$$

the set of points reachable from $\boldsymbol{x}$ after k steps of time, for $k = 0$, $A_+^k(\boldsymbol{x}) := \{\boldsymbol{x}\}$, and the set of points reachable from $\boldsymbol{x}$

$$A_+(\boldsymbol{x}) = \bigcup_{k\in\mathbb{N}} A_+^k(\boldsymbol{x}) \ .$$

The associated control model CM(F) is called forward accessible if for each $\boldsymbol{x} \in X$, the set $A_+(\boldsymbol{x})$ has non empty-interior.

Let E be a subset of X. We note $A_+(E) = \bigcup_{\boldsymbol{x}\in E} A_+(\boldsymbol{x})$, and we say that E is invariant if $A_+(E) \subset E$. We call a set minimal if it is closed, invariant, and does not strictly contain any closed and invariant subset. Restricted to a minimal set, a Markov chain has strong properties, as stated in the following lemma.

Lemma 6. *[12, 7.2.4, 7.2.6 and 7.3.5] Let $M \subset X$ be a minimal set for CM(F). If CM(F) is forward accessible then the NSS(F) model restricted to M is an open set irreducible T-chain.*

Furthermore, if the control set O_w and M are connected, and that M is the unique minimal set of the CM(F), then the NSS(F) model is a ψ-irreducible aperiodic T-chain for which every compact set is a small set.

We can now prove the following lemma:

Lemma 7. *The Markov chain $\mathcal{Z}$ is open-set irreducible, ψ-irreducible, aperiodic, and compacts of $\mathbb{R}$ are small-sets.*

Proof. This is deduced from Lemma 6 when all its conditions are fulfilled. We then have to show the right properties of the underlying control model.

Let $F : (z, \boldsymbol{u}) \mapsto z \exp\left(-1/(2d_\sigma)\left(\|\boldsymbol{u}\|^2/n - 1\right)\right) + [\boldsymbol{u}]_1$, then we do have $Z_{t+1} = F(Z_t, \boldsymbol{\xi}_t^\star)$. The function F is smooth (it is not smooth along the instances $\boldsymbol{\xi}_{t,i}$, but along the chosen step $\boldsymbol{\xi}_t^\star$). Furthermore, with Lemma 1 the distribution of $\boldsymbol{\xi}_t^\star$ admits a continuous density, whose support is $\mathbb{R}^n$. Therefore the process $\mathcal{Z}$ is a NSS(F) model of control set $\mathbb{R}^n$.

We now have to show that the associated control model is forward accessible. Let $z \in \mathbb{R}$. When $[\boldsymbol{\xi}_t^\star]_1 \to \pm\infty$, $F(z, \boldsymbol{\xi}_t^\star) \to \pm\infty$. As F is continuous, for the right value of $[\boldsymbol{\xi}_t^\star]_1$ any point of $\mathbb{R}$ can be reach. Therefore for any $z \in \mathbb{R}$, $A_+(z) = \mathbb{R}$. The set $\mathbb{R}$ has a non-empty interior, so the CM(F) is forward accessible.

As from any point of $\mathbb{R}$, all of $\mathbb{R}$ can be reached, so the only invariant set is $\mathbb{R}$ itself. It is therefore the only minimal set. Finally, the control set $O_w = \mathbb{R}^n$ is connected, and so is the only minimal set, so all the conditions of Lemma 6 are met. So the Markov chain $\mathcal{Z}$ is ψ-irreducible, aperiodic, and compacts of $\mathbb{R}$ are small-sets. □

We may now show Foster-Lyapunov drift conditions to ensure the Harris positive recurrence of the chain $\mathcal{Z}$. In order to do so, we will need the following lemma.

Lemma 8. *Let* $\exp(-\frac{1}{2d_\sigma}(\frac{\|\boldsymbol{\xi}^\star\|^2}{n} - 1))$ *be denoted* $\eta^\star$. *For all* $\lambda > 2$ *there exists* $\alpha > 0$ *such that*

$$\mathbb{E}\left(\eta^{\star -\alpha}\right) - 1 < 0 \ . \tag{16}$$

Proof. Let φ denote the density of the standard normal law. According to Lemma 1 the density of $\boldsymbol{\xi}_t^\star$ is the function $(u_i)_{i\in[1..n]} \mapsto \Psi_{1:\lambda}(u_1)\varphi(u_2)\cdots\varphi(u_n)$. Using the Taylor series of the exponential function we have

$$\begin{aligned}
\mathbb{E}\left(\eta^{\star -\alpha}\right) &= \mathbb{E}\left(\exp\left(-\frac{\alpha}{2d_\sigma}\left(\frac{\|\boldsymbol{\xi}^\star\|^2}{n} - 1\right)\right)\right) \\
&= \mathbb{E}\left(\sum_{i=0}^{\infty} \frac{\left(-\frac{\alpha}{2d_\sigma}\left(\frac{\|\boldsymbol{\xi}^\star\|^2}{n} - 1\right)\right)^i}{i!}\right) \ .
\end{aligned}$$

Since $\Psi_{1:\lambda}(u_1) \leq \lambda\varphi(u_1)$,

$$\exp\left|\frac{\alpha}{2d_\sigma}\left(\frac{\|\boldsymbol{\xi}^\star\|^2}{n} - 1\right)\right| \Psi_{1:\lambda}(u_1)\prod_{i=2}^{n}\varphi(u_i) \leq \lambda \exp\left|\frac{\alpha}{2d_\sigma}\left(\frac{\|\boldsymbol{\xi}^\star\|^2}{n} - 1\right)\right| \prod_{i=1}^{n}\varphi(u_i)$$

which is integrable, and so $\mathbb{E}|\eta^{\star -\alpha}| < \infty$. Hence we can apply Fubini's theorem and invert integral with series (which are integrals for the counting measure). Therefore

$$\begin{aligned}
\mathbb{E}\left(\eta^{\star -\alpha}\right) &= \sum_{i=0}^{\infty} \frac{1}{i!} \mathbb{E}\left(\left(-\frac{\alpha}{2d_\sigma}\left(\frac{\|\boldsymbol{\xi}^\star\|^2}{n} - 1\right)\right)^i\right) \\
&= 1 - \frac{\alpha}{2d_\sigma n}\left(\frac{\mathbb{E}\left(\|\boldsymbol{\xi}^\star\|^2\right)}{n} - 1\right) - o\left(\alpha^2\right) \\
&= 1 - \frac{\alpha}{2d_\sigma n}\left(\mathbb{E}\left(\mathcal{N}_{1:\lambda}^2\right) - 1\right) - o\left(\alpha^2\right) \quad .
\end{aligned}$$

According to Lemma 3 $\mathbb{E}\left(\mathcal{N}_{1:\lambda}^2\right) > 1$ for $\lambda > 2$, so when $\alpha > 0$ goes to 0 we have $\mathbb{E}\left(\eta^{\star -\alpha}\right) < 1$. □

We are now ready to prove the following lemma:

Lemma 9. *The Markov chain $\mathcal{Z}$ is Harris recurrent positive, and admits a unique invariant measure μ. Furthermore, for $f\colon x \mapsto |x|^\alpha \in \mathbb{R}$, $\mu(f) = \int_{\mathbb{R}} \mu(dx) f(x) < \infty$, with α such that Eq. (16) holds true.*

Proof. By using Lemma 7 and Lemma 5, we just need the drift condition (14) to prove Lemma 9. Let V be such that for $x \in \mathbb{R}$, $V(x) = |x|^\alpha + 1$.

$$\begin{aligned}
\Delta V(x) &= \int_{\mathbb{R}} P(x, dy) V(y) - V(x) \\
&= \int_{\mathbb{R}} P\left(\frac{x}{\eta^\star} + [\boldsymbol{\xi}^\star]_1 \in dy\right)(1 + |y|^\alpha) - (1 + |x|^\alpha) \\
&= \mathbb{E}\left(\left|\frac{x}{\eta^\star} + [\boldsymbol{\xi}^\star]_1\right|^\alpha\right) - |x|^\alpha \\
&\leq |x|^\alpha \mathbb{E}\left(\eta^{\star -\alpha} - 1\right) + \mathbb{E}\left([\boldsymbol{\xi}^\star]_1^\alpha\right) \\
\frac{\Delta V(x)}{V(x)} &= \frac{|x|^\alpha}{1 + |x|^\alpha} \mathbb{E}\left(\eta^{\star -\alpha} - 1\right) + \frac{1}{1 + |x|^\alpha} \mathbb{E}\left(\mathcal{N}_{1:\lambda}^\alpha\right) \\
\lim_{|x| \longrightarrow \infty} \frac{\Delta V(x)}{V(x)} &= \mathbb{E}\left(\eta^{\star -\alpha} - 1\right)
\end{aligned}$$

We take α such that Eq. (16) holds true (as according to Lemma 8, there exists such a α). As $\mathbb{E}(\eta^{\star -\alpha} - 1) < 0$, there exists $\epsilon > 0$ and $M > 0$ such that for all $|x| \geq M$, $\Delta V / V(x) \leq -\epsilon$. Let b be equal to $\mathbb{E}(\mathcal{N}_{1:\lambda}^2) + \epsilon V(M)$. Then for all $|x| \leq M$, $\Delta V(x) \leq -\epsilon V(x) + b$. Therefore, if we note $C = [-M, M]$, which is according to Lemma 7 a small-set, we do have $\Delta V(x) \leq -\epsilon V(x) + b \mathbf{1}_C(x)$ which is Eq. (14) with $f = \epsilon V$. Therefore from Lemma 5 the chain $\mathcal{Z}$ is positive Harris recurrent with invariant probability measure μ, and ϵV is μ-integrable. And since μ is a probability measure, $\mu(\mathbb{R}) = 1$. Since $\mu(f) = \int_R |x|^\alpha \mu(dx) + \mu(\mathbb{R}) < \infty$, the function $x \mapsto |x|^\alpha$ is also μ-integrable. □

Since the sequence $(\boldsymbol{\xi}_t^\star)_{t\in\mathbb{N}}$ is i.i.d. with a distribution that we denote μ_ξ, and that $\mathcal{Z}$ is a Harris positive Markov chain with invariant measure μ, the sequence $(Z_t, \boldsymbol{\xi}_t^\star)_{t\in\mathbb{N}}$ is also a Harris positive Markov chain, with invariant measure $\mu \times \mu_\xi$. In order to use Lemma 4 on the right hand side of (12), we need to show that $\mathbb{E}_{\mu\times\mu_\xi}|\ln|Z_k| - \ln|Z_{k-1}| + \ln|\eta_{k-1}^\star|| < \infty$, which would be implied by $\mathbb{E}_\mu|\ln|Z|| < \infty$ and $\mathbb{E}_{\mu_\xi}|\ln|\eta^\star|| < \infty$. To show that $\mathbb{E}_\mu|\ln|Z|| < \infty$ we will use the following lemma on the existence of moments for stationary Markov chains.

Lemma 10. *Let $\mathcal{Z}$ be a Harris-recurrent Markov chain with stationary measure μ, on a state space $(S, \mathcal{F})$, with $\mathcal{F}$ is σ-field of subsets of S. Let f be a positive measurable function on S.*

In order that $\int_S f(z)\mu(dz) < \infty$, it suffices that for some set $A \in \mathcal{F}$ such that $0 < \mu(A)$ and $\int_A f(z)\mu(dz) < \infty$, and some measurable function g with $g(z) \geq f(z)$ for $z \in A^c$,

1. $$\int_{A^c} P(z, dy)g(y) \leq g(z) - f(z) \quad , \forall x \in A^c$$
2. $$\sup_{z\in A} \int_{A^c} P(z, dy)g(y) < \infty$$

.

We may now prove the following theorem on the geometric divergence of $([\boldsymbol{X}_t]_1)_{t\in\mathbb{N}}$.

Theorem 2. *On linear functions, for $\lambda \geq 3$, the absolute value of the first dimension of the parent point in the $(1, \lambda)$-CSA-ES without cumulation ($c = 1$) diverges geometrically almost surely at the rate of $1/(2d_\sigma n)\mathbb{E}(\mathcal{N}_{1:\lambda}^2 - 1)$, i.e.*

$$\frac{1}{t}\ln\left|\frac{[\boldsymbol{X}_t]_1}{[\boldsymbol{X}_0]_1}\right| \xrightarrow[t\to\infty]{a.s} \frac{1}{2d_\sigma n}\left(\mathbb{E}\left(\mathcal{N}_{1:\lambda}^2\right) - 1\right) > 0 \ . \tag{17}$$

Proof. According to Lemma 9 the Markov chain $\mathcal{Z}$ is Harris positive with invariant measure μ. According to Lemma 1 the sequence $(\boldsymbol{\xi}_t^\star)_{t\in\mathbb{N}}$ is i.i.d. with a distribution that we denote μ_ξ, so the sequence $(Z_t, \boldsymbol{\xi}_t^\star)_{t\in\mathbb{N}}$ is a Harris positive Markov chain with invariant measure $\mu \times \mu_\xi$. In order to apply Lemma 4 to the right hand side of (12), we need to prove that $\mathbb{E}_{\mu\times\mu_\xi}|\ln|Z_t| - \ln|Z_{t-1}| + \ln|\eta_{t-1}^\star||$ is finite. With the triangular inequality, this is implied if $\mathbb{E}_\mu|\ln|Z_t||$ is finite and $\mathbb{E}_{\mu_\xi}|\ln|\eta^\star||$ is finite. We have $\ln|\eta^\star| = (\|\xi^\star\|^2/n - 1)/(2d_\sigma)$. Since the density of $\xi^\star$ is the function $\boldsymbol{u} \in \mathbb{R}^n \mapsto \Psi_{1:\lambda}([\boldsymbol{u}]_1)\prod_{i=2}^n \varphi([\boldsymbol{u}]_i)$, with φ the density of the standard normal law, and that $\Psi_{1:\lambda}([\boldsymbol{u}]_1) \leq \lambda\varphi([\boldsymbol{u}]_1)$, the function $\boldsymbol{u} \in \mathbb{R}^n \mapsto |\|\boldsymbol{u}\|^2/n - 1|/(2d_\sigma)\Psi_{1:\lambda}([\boldsymbol{u}]_1)\prod_{i=2}^n \varphi([\boldsymbol{u}]_i)$ is integrable and so $\mathbb{E}_{\mu_\xi}|\ln|\eta^\star||$ is finite.

We now prove that the function $g : x \mapsto \ln|x|$ is μ-integrable. From Lemma 9 we know that the function $f : x \mapsto |x|^\alpha$ is μ-integrable, and as for any $M > 0$, and any $x \in A := [-M, M]^c$ there exists $K > 0$ such that $K|x|^\alpha > |\ln|x||$, then $\int_A |g(x)|\mu(dx) < \int_1 K|x|^\alpha\mu(dx) < \infty$. So what is left is to prove that $\int_{A^c} |g(x)|\mu(dx)$ is also finite. We now check the conditions to use Lemma 10.

According to Lemma 7 the chain $\mathcal{Z}$ is open-set irreducible, so $\mu(A) > 0$. For $C > 0$, if we take $h : z \mapsto C/\sqrt{|z|}$, with M small enough we do have for all $z \in A^c$, $h(z) \geq |g(z)|$. Furthermore, denoting η the function that maps $\boldsymbol{u} \in \mathbb{R}^n$ to $\exp((\|\boldsymbol{u}\|^2/n - 1)/(2d_\sigma))$,

$$\begin{aligned}\int_{A^c} P(z, dy)h(y) &= \int_S P\left(\frac{z}{\eta^\star} + [\boldsymbol{\xi}^\star]_1 \in dy\right) \frac{C}{\sqrt{|y|}} \mathbf{1}_{A^c}(y) \\ &= \mathbb{E}\left(\frac{C}{\sqrt{\left|\frac{z}{\eta^\star} + [\boldsymbol{\xi}^\star]_1\right|}} \mathbf{1}_{[-M,M]}\left(\frac{z}{\eta^\star} + [\boldsymbol{\xi}^\star]_1\right)\right) \\ &= \int_{\mathbb{R}^n} C \frac{\mathbf{1}_{[-M,M]}\left(\frac{z}{\eta(\boldsymbol{u})} + [\boldsymbol{u}]_1\right)}{\sqrt{\left|\frac{z}{\eta(\boldsymbol{u})} + [\boldsymbol{u}]_1\right|}} \Psi_{1:\lambda}([\boldsymbol{u}]_1) \prod_{i=2}^n \varphi([\boldsymbol{u}_i]) d\boldsymbol{u} \ .\end{aligned}$$

Using that $\Psi_{1:\lambda}(x) \leq \lambda\varphi(\boldsymbol{x})$ and that characteristic functions are upper bounded by 1,

$$C \frac{\mathbf{1}_{[-M,M]}\left(\frac{z}{\eta(\boldsymbol{u})} + [\boldsymbol{u}]_1\right)}{\sqrt{\left|\frac{z}{\eta(\boldsymbol{u})} + [\boldsymbol{u}]_1\right|}} \Psi_{1:\lambda}([\boldsymbol{u}]_1) \prod_{i=2}^n \varphi([\boldsymbol{u}_i]) \leq \frac{\lambda C}{\sqrt{\left|\frac{z}{\eta(\boldsymbol{u})} + [\boldsymbol{u}]_1\right|}} \prod_{i=1}^n \varphi([\boldsymbol{u}_i]) \ . \tag{18}$$

The right hand side of (18) is integrable for high values of $\|\boldsymbol{u}\|$, and as the function $x \mapsto |z + x|^{-1/2}$ is integrable around 0, the right hand side of (18) is integrable. Also, for $a > 0$ since $\int_{-a}^{a} |z + x|^{-1/2} dx$ is maximal for $z = 0$,

$$\sup_{z \in \mathbb{R}} \int_{A^c} P(z, dy)h(y) \leq \int_{\mathbb{R}^n} \frac{\lambda C}{\sqrt{|[\boldsymbol{u}]_1|}} \prod_{i=1}^n \varphi([\boldsymbol{u}_i]) d\boldsymbol{u} < \infty \ ,$$

which satisfies the second condition of Lemma 10. Furthermore, we can apply Lebesgue's dominated convergence theorem using the fact that the left hand side of (18) converges to 0 almost everywhere when $M \to 0$, and so $\int_{A^c} P(z, dy)h(y)$ converges to 0 when $M \to 0$. Combining this with the fact that $\int_{A^c} P(z, dy)h(y)$ is bounded for all $z \in \mathbb{R}$, there exists M small enough such that $\int_{A^c} P(z, dy)h(y) \leq h(z)$ for all $z \in [-M, M]$. Finally, for C large enough $|g(z)|$ is negligible compared to $C/\sqrt{|z|}$, hence for M small enough and C large enough the first condition of Lemma 10 is satisfied. Hence, according to Lemma 10 the function $|g|$ is μ-integrable.

This allows us to apply Lemma 4 to the right hand side of (12) and obtain that

$$\frac{1}{t} \sum_{k=0}^{t-1} \left(\ln|Z_k| - \ln|Z_{k-1}| + \ln\left|\eta^\star_{k-1}\right|\right) \xrightarrow[t \to +\infty]{a.s.} \mathbb{E}_\mu(\ln(Z)) - \mathbb{E}_\mu(\ln(Z)) + \mathbb{E}_{\mu_\xi}(\ln(\eta^\star)) \ .$$

Since $\mathbb{E}_{\mu_\xi}(\ln(\eta^\star)) = \mathbb{E}((\|\boldsymbol{\xi}^\star\|^2/n - 1)/(2d_\sigma))$ and that $\mathbb{E}(\|\boldsymbol{\xi}^\star\|^2) = \mathbb{E}(\mathcal{N}^2_{1:\lambda}) + (n-1)$, we have $\mathbb{E}_{\mu_\xi}(\ln(\eta^\star)) = (\mathbb{E}(\mathcal{N}^2_{1:\lambda}) - 1)/(2d_\sigma n)$, which with (12) gives (17), and with Lemma 3 is strictly positive for $\lambda \geq 3$. □

4 Divergence rate of CSA-ES with cumulation

We are now investigating the $(1, \lambda)$-CSA-ES with cumulation, i.e. $0 < c < 1$. According to Lemma 1, the random variables $(\boldsymbol{\xi}_t^\star)_{t\in\mathbb{N}}$ are i.i.d., hence the path $\boldsymbol{\mathcal{P}} := (\boldsymbol{p}_t)_{t\in\mathbb{N}}$ is a Markov chain. By a recurrence on Eq. (2) we see that the path follows

$$\boldsymbol{p}_t = (1-c)^t \boldsymbol{p}_0 + \sqrt{c(2-c)} \sum_{k=0}^{t-1} (1-c)^k \underbrace{\boldsymbol{\xi}_{t-1-k}^\star}_{\text{i.i.d.}} \quad . \tag{19}$$

For $i \neq 1$, $[\boldsymbol{\xi}_t^\star]_i \sim \mathcal{N}(0,1)$ and, as also $[\boldsymbol{p}_0]_i \sim \mathcal{N}(0,1)$, by recurrence $[\boldsymbol{p_t}]_i \sim \mathcal{N}(0,1)$ for all $t \in \mathbb{N}$. For $i = 1$ with cumulation ($c < 1$), the influence of $[\boldsymbol{p}_0]_1$ vanishes with $(1-c)^t$. Furthermore, as from Lemma 1 the sequence $([\boldsymbol{\xi}_t^\star]_1])_{t\in\mathbb{N}}$ is independent, we get by applying the Kolgomorov's three series theorem that the series $\sum_{k=0}^{t-1}(1-c)^k \left[\boldsymbol{\xi}_{t-1-k}^\star\right]_1$ converges almost surely. Therefore, the first component of the path becomes distributed as the random variable $[\boldsymbol{p}_\infty]_1 = \sqrt{c(2-c)}\sum_{k=0}^{\infty}(1-c)^k[\boldsymbol{\xi}_k^\star]_1$ (by re-indexing the variable $\boldsymbol{\xi}_{t-1-k}^\star$ in $\boldsymbol{\xi}_k^\star$, as the sequence $(\boldsymbol{\xi}_t^\star)_{t\in\mathbb{N}}$ is i.i.d.). We will specify the series $\sqrt{c(2-c)}\sum_{k=0}^{\infty}(1-c)^k[\boldsymbol{\xi}_k^\star]_1$ by applying a law of large numbers to the right hand side of (5), after showing that the Markov chain $[\boldsymbol{\mathcal{P}}]_1 := ([\boldsymbol{p}_t]_1)_{t\in\mathbb{N}}$ has the right stability properties to apply a law of large numbers to it.

Lemma 11. *The Markov chain $[\boldsymbol{\mathcal{P}}]_1$ is φ-irreducible, aperiodic, and compacts of $\mathbb{R}$ are small-sets.*

Proof. Using (19) and Lemma 1, $[\boldsymbol{p}_{t+1}]_1 = (1-c)[\boldsymbol{p}_t]_1 + \sqrt{c(2-c)}[\boldsymbol{\xi}_t^\star]_1$ with $[\boldsymbol{\xi}_t^\star]_1 \sim \mathcal{N}_{1:\lambda}$. Hence the transition kernel for $[\boldsymbol{\mathcal{P}}]_1$ writes

$$P(p, A) = \int_{\mathbb{R}} \mathbf{1}_A \left((1-c)p + \sqrt{c(2-c)}u\right) \Psi_{1:\lambda}(u)du \quad .$$

With a change of variables $\tilde{u} = (1-c)p + \sqrt{c(2-c)}u$, we get that

$$P(p, A) = \frac{1}{\sqrt{c(2-c)}} \int_{\mathbb{R}} \mathbf{1}_A(\tilde{u}) \Psi_{1:\lambda}\left(\frac{\tilde{u} - (1-c)p}{\sqrt{(2-c)c}}\right) d\tilde{u} \quad .$$

As $\Psi_{1:\lambda}(u) > 0$ for all $u \in \mathbb{R}$, for all A with positive Lebesgue measure we have $P(p, A) > 0$, thus the chain $[\boldsymbol{\mathcal{P}}]_1$ is μ_{Leb}-irreducible with μ_{Leb} denoting the Lebesgue measure.

Furthermore, if we take C a compact of $\mathbb{R}$, and ν_C a measure such that for A a Borel set of $\mathbb{R}$

$$\nu_C(A) = \frac{1}{\sqrt{(2-c)c}} \int_{\mathbb{R}} \mathbf{1}_A(\tilde{u}) \min_{p\in C} \Psi_{1:\lambda}\left(\frac{\tilde{u} - (1-c)p}{\sqrt{(2-c)c}}\right) d\tilde{u} \quad , \tag{20}$$

we see that $P(p, A) \geq \nu_C(A)$ for all $p \in C$. Furthermore C being a compact for all $\tilde{u} \in \mathbb{R}$ there exists $\delta_{\tilde{u}} > 0$ such that $\Psi_{1:\lambda}((\tilde{u} - (1-c)p)/\sqrt{(2-c)c}) > \delta_{\tilde{u}}$ for all $p \in C$. Hence ν_C is not a trivial measure. And therefore compact sets of $\mathbb{R}$ are small sets for $[\boldsymbol{\mathcal{P}}]_1$. Finally, if C has positive Lebesgue measure $\nu_C(C) > 0$, so the chain $[\boldsymbol{\mathcal{P}}]_1$ is strongly aperiodic. □

We now prove that the Markov chain $[\boldsymbol{\mathcal{P}}]_1$ is Harris positive.

Lemma 12. *The chain $[\boldsymbol{\mathcal{P}}]_1$ is Harris recurrent positive with invariant measure μ_{path}, and the function $x \mapsto x^2$ is μ_{path}-integrable.*

Proof. Let $V : x \mapsto x^2 + 1$. Then

$$\begin{aligned}
\Delta V(x) &= \int_{\mathbb{R}} V(y)P(x, dy) - V(x) \\
\Delta V(x) &= \mathbb{E}\left(\left((1-c)x + \sqrt{c(2-c)}\left[\boldsymbol{\xi}^\star\right]_1\right)^2 + 1\right) - x^2 - 1 \\
\Delta V(x) &\leq ((1-c)^2 - 1)x^2 + 2|x|\sqrt{c(2-c)}\mathbb{E}\left(\left[\boldsymbol{\xi}^\star\right]_1\right) + c(2-c)\mathbb{E}\left(\left[\boldsymbol{\xi}^\star\right]_1{}^2\right) \\
\frac{\Delta V(x)}{V(x)} &\leq -c(2-c)\frac{x^2}{1+x^2} + \frac{2|x|\sqrt{c(2-c)}}{1+x^2}\mathbb{E}\left(\left|\left[\boldsymbol{\xi}^\star\right]_1\right|\right) + \frac{c(2-c)}{1+x^2}\mathbb{E}\left(\left[\boldsymbol{\xi}^\star\right]_1{}^2\right) \\
\lim_{|x|\to\infty} \frac{\Delta V(x)}{V(x)} &\leq -c(2-c)
\end{aligned}$$

As $0 < c \leq 1$, $c(2-c)$ is strictly positive and therefore, for $\epsilon > 0$ there exists $C = [-M, M]$ with $M > 0$ such that for all $x \in C^c$, $\Delta V(x)/V(x) \leq -\epsilon$. And since $\mathbb{E}([\boldsymbol{\xi}^\star]_1)$ and $\mathbb{E}([\boldsymbol{\xi}^\star]_1^2)$ are finite, ΔV is bounded on the compact C and so there exists $b \in \mathbb{R}$ such that $\Delta V(x) \leq -\epsilon V(x) + b\mathbf{1}_C$ for all $x \in \mathbb{R}$.

According to Lemma 11 the chain $[\boldsymbol{\mathcal{P}}]_1$ is φ-irreducible and aperiodic, so with Lemma 5 it is positive Harris recurrent, with invariant measure μ_{path}, and V is μ_{path}-integrable. Therefore the function $x \mapsto x^2$ is also μ_{path}-integrable.

For $g \geq 1$ a function and ν a signed measure, the g-norm of ν is defined through $\|\nu\|_g = \sup_{h:|h|\leq g} |\nu(h)|$ Lemma 12 allow us to show the convergence of the transition kernel P to the stationary measure μ_{path} in g-norm through the following lemma. .

Lemma 13. *[12, 14.3.5] Suppose $\boldsymbol{\Phi}$ is an aperiodic positive Harris chain on a space $\mathcal{X}$ with stationary measure π, and that there exists some non-negative function V, a function $f \geq 1$, a small-set C and $b \in \mathbb{R}$ such that for all $x \in \mathcal{X}$, $\Delta V(x) \leq -f(x) + b\mathbf{1}_C(x)$. Then for all initial probability distribution ν, $\|\nu P^n - \pi\|_f \underset{t\to\infty}{\longrightarrow} 0$.*

We now obtain geometric divergence of the step-size and get an explicit estimate of the expression of the divergence rate.

Theorem 3. *The step-size of the $(1, \lambda)$-CSA-ES with $\lambda \geq 2$ diverges geometrically fast if $c < 1$ or $\lambda \geq 3$. Almost surely and in expectation we have for $0 < c \leq 1$,*

$$\frac{1}{t}\ln\left(\frac{\sigma_t}{\sigma_0}\right) \underset{t\to\infty}{\longrightarrow} \frac{1}{2d_\sigma n}\underbrace{\left(2(1-c)\,\mathbb{E}\left(\mathcal{N}_{1:\lambda}\right)^2 + c\left(\mathbb{E}\left(\mathcal{N}_{1:\lambda}^2\right) - 1\right)\right)}_{>0 \text{ for } \lambda\geq 3 \text{ and for } \lambda=2 \text{ and } c<1} . \tag{21}$$

Proof. We will start by the convergence in expectation. With Lemma 1, $[\boldsymbol{\xi}^\star]_1 \sim \mathcal{N}_{1:\lambda}$, and $[\boldsymbol{\xi}^\star]_i \sim \mathcal{N}(0, 1)$ for all $i \in [2..n]$. Hence, using that $[\boldsymbol{p}_0]_i \sim \mathcal{N}(0, 1)$, $[\boldsymbol{p}_t]_i \sim$

$\mathcal{N}(0,1)$ for all $i \in [2..n]$ too. Therefore $\mathbb{E}(\|\boldsymbol{p}_{t+1}\|^2) = \mathbb{E}([\boldsymbol{p}_{t+1}]_1^2) + n - 1$. By recurrence $\left[\boldsymbol{p}_{t+1}\right]_1 = (1-c)^{t+1}[\boldsymbol{p}_0]_1 + \sqrt{c(2-c)}\sum_{i=0}^{t}(1-c)^i \left[\boldsymbol{\xi}^\star_{t-i}\right]_1$. When t goes to infinity, the influence of $[\boldsymbol{p}_0]_1$ in this equation goes to 0 with $(1-c)^{t+1}$, so we can remove it when taking the limit:

$$\lim_{t\to\infty} \mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^2\right) = \lim_{t\to\infty} \mathbb{E}\left(\left(\sqrt{c(2-c)}\sum_{i=0}^{t}(1-c)^i \left[\boldsymbol{\xi}^\star_{t-i}\right]_1\right)^2\right) \tag{22}$$

We will now develop the sum with the square, such that we have either a product $\left[\boldsymbol{\xi}^\star_{t-i}\right]_1 \left[\boldsymbol{\xi}^\star_{t-j}\right]_1$ with $i \neq j$, or $\left[\boldsymbol{\xi}^\star_{t-j}\right]_1^2$. This way, we can separate the variables by using Lemma 1 with the independence of $\boldsymbol{\xi}^\star_i$ over time. To do so, we use the development formula $(\sum_{i=1}^n a_n)^2 = 2\sum_{i=1}^n \sum_{j=i+1}^n a_i a_j + \sum_{i=1}^n a_i^2$. We take the limit of $\mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^2)$ and find that it is equal to

$$\lim_{t\to\infty} c(2-c)\left(2\sum_{i=0}^{t}\sum_{j=i+1}^{t}(1-c)^{i+j}\underbrace{\mathbb{E}\left(\left[\boldsymbol{\xi}^\star_{t-i}\right]_1\left[\boldsymbol{\xi}^\star_{t-j}\right]_1\right)}_{=\mathbb{E}\left[\boldsymbol{\xi}^\star_{t-i}\right]_1\mathbb{E}\left[\boldsymbol{\xi}^\star_{t-j}\right]_1=\mathbb{E}[\mathcal{N}_{1:\lambda}]^2} + \sum_{i=0}^{t}(1-c)^{2i}\underbrace{\mathbb{E}\left(\left[\boldsymbol{\xi}^\star_{t-i}\right]_1^2\right)}_{=\mathbb{E}[\mathcal{N}_{1:\lambda}^2]}\right) \tag{23}$$

Now the expected value does not depend on i or j, so what is left is to calculate $\sum_{i=0}^t\sum_{j=i+1}^t(1-c)^{i+j}$ and $\sum_{i=0}^t(1-c)^{2i}$. We have $\sum_{i=0}^t\sum_{j=i+1}^t(1-c)^{i+j} = \sum_{i=0}^t(1-c)^{2i+1}\frac{1-(1-c)^{t-i}}{1-(1-c)}$ and when we separates this sum in two, the right hand side goes to 0 for $t \to \infty$. Therefore, the left hand side converges to $\lim_{t\to\infty}\sum_{i=0}^t(1-c)^{2i+1}/c$, which is equal to $\lim_{t\to\infty}(1-c)/c\sum_{i=0}^t(1-c)^{2i}$. And $\sum_{i=0}^t(1-c)^{2i}$ is equal to $(1-(1-c)^{2t+2})/(1-(1-c)^2)$, which converges to $1/(c(2-c))$. So, by inserting this in Eq. (23) we get that $\mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^2) \underset{t\to\infty}{\longrightarrow} 2\frac{1-c}{c}\mathbb{E}(\mathcal{N}_{1:\lambda})^2 + \mathbb{E}(\mathcal{N}_{1:\lambda}^2)$, which gives us the right hand side of Eq. (21).

By summing $\mathbb{E}(\ln(\sigma_{i+1}/\sigma_i))$ for $i = 0,\dots,t-1$ and dividing by t we have the Cesaro mean $1/t\mathbb{E}(\ln(\sigma_t/\sigma_0))$ that converges to the same value that $\mathbb{E}(\ln(\sigma_{t+1}/\sigma_t))$ converges to when t goes to infinity. Therefore we have in expectation Eq. (21).

We will now focus on the almost sure convergence. From Lemma 12, we see that we have the right conditions to apply Lemma 4 to the chain $[\boldsymbol{\mathcal{P}}]_1$ with the μ_{path}-integrable function $g : x \mapsto x^2$. So

$$\frac{1}{t}\sum_{k=1}^{t}[\boldsymbol{p}_k]_1^2 \underset{t\to\infty}{\overset{a.s}{\longrightarrow}} \mu_{path}(g) \ .$$

With Eq. (5) and using that $\mathbb{E}(\|\boldsymbol{p}_{t+1}\|\|^2) = \mathbb{E}([\boldsymbol{p}_{t+1}]_1^2) + n - 1$, we obtain that

$$\frac{1}{t}\ln\left(\frac{\sigma_t}{\sigma_0}\right) \underset{t\to\infty}{\overset{a.s}{\longrightarrow}} \frac{c}{2d_\sigma n}(\mu_{path}(g) - 1) \ .$$

We now prove that $\mu_{path}(g) = \lim_{t\to\infty}\mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^2)$. Let ν be the initial distribution of $[\boldsymbol{p}_0]_1$, so we have $|\mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^2) - \mu_{path}(g)| \leq \|\nu P^{t+1} - \mu_{path}\|_h$, with $h : x \mapsto$

$1 + x^2$. From the proof of Lemma 12 and from Lemma 11 we have all conditions for Lemma 13. Therefore $\|\nu P^{t+1} - \mu_{path}\|_h \xrightarrow[t\to\infty]{} 0$, which shows that $\mu_{path}(g) = \lim_{t\to\infty} \mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^2) = (2-2c)/c\mathbb{E}(\mathcal{N}_{1:\lambda})^2 + \mathbb{E}(\mathcal{N}_{1:\lambda}^2)$.

According to Lemma 3, for $\lambda = 2$, $\mathbb{E}(\mathcal{N}_{1:2}^2) = 1$, so the RHS of Eq. (21) is equal to $(1-c)/(d_\sigma n)\mathbb{E}(\mathcal{N}_{1:2})^2$. The expected value of $\mathcal{N}_{1:2}$ is strictly negative, so the previous expression is strictly positive. Furthermore, according to Lemma 3, $\mathbb{E}(\mathcal{N}_{1:\lambda}^2)$ increases strictly with λ, as does $\mathbb{E}(\mathcal{N}_{1:2})^2$. Therefore we have geometric divergence for $\lambda \geq 2$ if $c < 1$, and for $\lambda \geq 3$.

□

From Eq. (1) we see that the behaviour of the step-size and of $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ are directly related. Geometric divergence of the step-size, as shown in Theorem 3, means that also the movements in search space and the improvements on affine linear functions f increase geometrically fast. Analyzing $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ with cumulation would require to study a double Markov chain, which is left to possible future research.

5 Study of the variations of $\ln(\sigma_{t+1}/\sigma_t)$

The proof of Theorem 3 shows that the step size increase converges to the right hand side of Eq. (21), for $t \to \infty$. When the dimension increases this increment goes to zero, which also suggests that it becomes more likely that σ_{t+1} is smaller than σ_t. To analyze this behavior, we study the variance of $\ln(\sigma_{t+1}/\sigma_t)$ as a function of c and the dimension.

Theorem 4. *The variance of* $\ln(\sigma_{t+1}/\sigma_t)$ *equals to*

$$\mathrm{Var}\left(\ln\left(\frac{\sigma_{t+1}}{\sigma_t}\right)\right) = \frac{c^2}{4d_\sigma^2 n^2}\left(\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^4\right) - \mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^2\right)^2 + 2(n-1)\right) \quad . \tag{24}$$

Furthermore, $\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^2\right) \xrightarrow[t\to\infty]{} \mathbb{E}\left(\mathcal{N}_{1:\lambda}^2\right) + \frac{2-2c}{c}\mathbb{E}\left(\mathcal{N}_{1:\lambda}\right)^2$ *and with* $a = 1-c$

$$\lim_{t\to\infty} \mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^4\right) = \frac{(1-a^2)^2}{1-a^4}\left(k_4 + k_{31} + k_{22} + k_{211} + k_{1111}\right) \quad , \tag{25}$$

where $k_4 = \mathbb{E}(\mathcal{N}_{1:\lambda}^4)$, $k_{31} = 4\frac{a(1+a+2a^2)}{1-a^3}\mathbb{E}\left(\mathcal{N}_{1:\lambda}^3\right)\mathbb{E}\left(\mathcal{N}_{1:\lambda}\right)$, $k_{22} = 6\frac{a^2}{1-a^2}\mathbb{E}\left(\mathcal{N}_{1:\lambda}^2\right)^2$, $k_{211} = 12\frac{a^3(1+2a+3a^2)}{(1-a^2)(1-a^3)}\mathbb{E}(\mathcal{N}_{1:\lambda}^2)\mathbb{E}(\mathcal{N}_{1:\lambda})^2$ *and* $k_{1111} = 24\frac{a^6}{(1-a)(1-a^2)(1-a^3)}\mathbb{E}\left(\mathcal{N}_{1:\lambda}\right)^4$.

Proof.

$$\mathrm{Var}\left(\ln\left(\frac{\sigma_{t+1}}{\sigma_t}\right)\right) = \mathrm{Var}\left(\frac{c}{2d_\sigma}\left(\frac{\|\boldsymbol{p}_{t+1}\|^2}{n} - 1\right)\right) = \frac{c^2}{4d_\sigma^2 n^2}\underbrace{\mathrm{Var}\left(\|\boldsymbol{p}_{t+1}\|^2\right)}_{\mathbb{E}\left(\|\boldsymbol{p}_{t+1}\|^4\right) - \mathbb{E}\left(\|\boldsymbol{p}_{t+1}\|^2\right)^2} \tag{26}$$

The first part of $\mathrm{Var}(\|\boldsymbol{p}_{t+1}\|^2)$, $\mathbb{E}(\|\boldsymbol{p}_{t+1}\|^4)$, is equal to $\mathbb{E}((\sum_{i=1}^n \left[\boldsymbol{p}_{t+1}\right]_i^2)^2)$. We develop it along the dimensions such that we can use the independence of $\left[\boldsymbol{p}_{t+1}\right]_i$ with

$[\boldsymbol{p}_{t+1}]_j$ for $i \neq j$, to get $\mathbb{E}(2\sum_{i=1}^{n}\sum_{j=i+1}^{n}\left[\boldsymbol{p}_{t+1}\right]_i^2\left[\boldsymbol{p}_{t+1}\right]_j^2+\sum_{i=1}^{n}\left[\boldsymbol{p}_{t+1}\right]_i^4)$. For $i \neq 1$ $\left[\boldsymbol{p}_{t+1}\right]_i$ is distributed according to a standard normal distribution, so $\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_i^2\right) = 1$ and $\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_i^4\right) = 3$.

$$\begin{aligned}
\mathbb{E}\left(\|\boldsymbol{p}_{t+1}\|^4\right) &= 2\sum_{i=1}^{n}\sum_{j=i+1}^{n}\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_i^2\right)\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_j^2\right)+\sum_{i=1}^{n}\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_i^4\right)\\
&= \left(2\sum_{i=2}^{n}\sum_{j=i+1}^{n}1\right)+2\sum_{j=2}^{n}\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^2\right)+\left(\sum_{i=2}^{n}3\right)+\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^4\right)\\
&= \left(2\sum_{i=2}^{n}(n-i)\right)+2(n-1)\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^2\right)+3(n-1)+\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^4\right)\\
&= \mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^4\right)+2(n-1)\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^2\right)+(n-1)(n+1)
\end{aligned}$$

The other part left is $\mathbb{E}(\|\boldsymbol{p}_{t+1}\|^2)^2$, which we develop along the dimensions to get $\mathbb{E}(\sum_{i=1}^{n}\left[\boldsymbol{p}_{t+1}\right]_i^2)^2 = (\mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^2)+(n-1))^2$, which equals to $\mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^2)^2+2(n-1)\mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^2)+(n-1)^2$. So by subtracting both parts we get $\mathbb{E}(\|\boldsymbol{p}_{t+1}\|^4)-\mathbb{E}(\|\boldsymbol{p}_{t+1}\|^2)^2 = \mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^4)-\mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^2)^2+2(n-1)$, which we insert into Eq. (26) to get Eq. (24).

The development of $\mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^2)$ is the same than the one done in the proof of Theorem 3, that is $\mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^2) = (2-2c)/c\mathbb{E}(\mathcal{N}_{1:\lambda})^2+\mathbb{E}(\mathcal{N}_{1:\lambda}^2)$. We now develop $\mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^4)$. We have $\mathbb{E}(\left[\boldsymbol{p}_{t+1}\right]_1^4) = \mathbb{E}(((1-c)^t[\boldsymbol{p}_0]_1+\sqrt{c(2-c)}\sum_{i=0}^{t}(1-c)^i\left[\boldsymbol{\xi}_{t-i}^{\star}\right]_1)^4)$. We neglect in the limit when t goes to ∞ the part with $(1-c)^t[\boldsymbol{p}_0]_1$, as it converges fast to 0. So

$$\lim_{t\to\infty}\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^4\right) = \lim_{t\to\infty}\mathbb{E}\left(c^2(2-c)^2\left(\sum_{i=0}^{t}(1-c)^i\left[\boldsymbol{\xi}_{t-i}^{\star}\right]_1\right)^4\right) . \tag{27}$$

To develop the RHS of Eq.(27) we use the following formula: for $(a_i)_{i\in[[1,m]]}$

$$\begin{aligned}
\left(\sum_{i=1}^{m}a_i\right)^4 =& \sum_{i=1}^{m}a_i^4+4\sum_{i=1}^{m}\sum_{\substack{j=1\\j\neq i}}^{m}a_i^3a_j+6\sum_{i=1}^{m}\sum_{j=i+1}^{m}a_i^2a_j^2\\
&+12\sum_{i=1}^{m}\sum_{\substack{j=1\\j\neq i}}^{m}\sum_{\substack{k=j+1\\k\neq i}}^{m}a_i^2a_ja_k+24\sum_{i=1}^{m}\sum_{j=i+1}^{m}\sum_{k=j+1}^{m}\sum_{l=k+1}^{m}a_ia_ja_ka_l \ .
\end{aligned} \tag{28}$$

This formula will allow us to use the independence over time of $[\boldsymbol{\xi}_t^{\star}]_1$ from Lemma 1, so that $\mathbb{E}([\boldsymbol{\xi}_i^{\star}]_1^3\left[\boldsymbol{\xi}_j^{\star}\right]_1) = \mathbb{E}([\boldsymbol{\xi}_i^{\star}]_1^3)\mathbb{E}(\left[\boldsymbol{\xi}_j^{\star}\right]_1) = \mathbb{E}(\mathcal{N}_{1:\lambda}^3)\mathbb{E}(\mathcal{N}_{1:\lambda})$ for $i \neq j$, and so on.

We apply Eq (28) on Eq (25), with $a = 1 - c$.

$$\begin{aligned}\lim_{t\to\infty} \frac{\mathbb{E}\left(\left[\boldsymbol{p}_{t+1}\right]_1^4\right)}{c^2(2-c)^2} = \lim_{t\to\infty} &\sum_{i=0}^{t} a^{4i}\mathbb{E}\left(\mathcal{N}_{1:\lambda}^4\right) + 4\sum_{i=0}^{t}\sum_{\substack{j=0\\j\neq i}}^{t} a^{3i+j}\mathbb{E}\left(\mathcal{N}_{1:\lambda}^3\right)\mathbb{E}\left(\mathcal{N}_{1:\lambda}\right) \\ &+ 6\sum_{i=0}^{t}\sum_{j=i+1}^{t} a^{2i+2j}\mathbb{E}\left(\mathcal{N}_{1:\lambda}^2\right)^2 \\ &+ 12\sum_{i=0}^{t}\sum_{\substack{j=0\\j\neq i}}^{t}\sum_{\substack{k=j+1\\k\neq i}}^{t} a^{2i+j+k}\mathbb{E}\left(\mathcal{N}_{1:\lambda}^2\right)\mathbb{E}\left(\mathcal{N}_{1:\lambda}\right)^2 \\ &+ 24\sum_{i=0}^{t}\sum_{j=i+1}^{t}\sum_{k=j+1}^{t}\sum_{l=k+1}^{t} a^{i+j+k+l}\mathbb{E}\left(\mathcal{N}_{1:\lambda}\right)^4 \end{aligned} \tag{29}$$

We now have to develop each term of Eq. (29).

$$\begin{aligned}\sum_{i=0}^{t} a^{4i} &= \frac{1-a^{4(t+1)}}{1-a^4} \\ \lim_{t\to\infty}\sum_{i=0}^{t} a^{4i} &= \frac{1}{1-a^4}\end{aligned} \tag{30}$$

$$\sum_{i=0}^{t}\sum_{\substack{j=0\\j\neq i}}^{t} a^{3i+j} = \sum_{i=0}^{t-1}\sum_{j=i+1}^{t} a^{3i+j} + \sum_{i=1}^{t}\sum_{j=0}^{i-1} a^{3i+j} \tag{31}$$

$$\begin{aligned}\sum_{i=0}^{t-1}\sum_{j=i+1}^{t} a^{3i+j} &= \sum_{i=0}^{t-1} a^{4i+1}\frac{1-a^{t-i}}{1-a} \\ \lim_{t\to\infty}\sum_{i=0}^{t-1}\sum_{j=i+1}^{t} a^{3i+j} &= \lim_{t\to\infty}\frac{a}{1-a}\sum_{i=0}^{t-1} a^{4i} \\ &= \frac{a}{(1-a)(1-a^4)}\end{aligned} \tag{32}$$

$$\begin{aligned}
\sum_{i=1}^{t}\sum_{j=0}^{i-1} a^{3i+j} &= \sum_{i=1}^{t} a^{3i}\frac{1-a^{i}}{1-a} \\
&= \frac{1}{1-a}\left(a^{3}\frac{1-a^{3t}}{1-a^{3}} - a^{4}\frac{1-a^{4t}}{1-a^{4}}\right) \\
\lim_{t\to\infty}\sum_{i=1}^{t}\sum_{j=0}^{i-1} a^{3i+j} &= \frac{1}{1-a}\left(\frac{a^{3}}{1-a^{3}} - \frac{a^{4}}{1-a^{4}}\right) \\
&= \frac{a^{3}(1-a^{4}) - a^{4}(1-a^{3})}{(1-a)(1-a^{3})(1-a^{4})} \\
&= \frac{a^{3}-a^{4}}{(1-a)(1-a^{3})(1-a^{4})} \qquad (33)
\end{aligned}$$

By combining Eq (32) with Eq (33) to Eq (31) we get

$$\begin{aligned}
\lim_{t\to\infty}\sum_{i=0}^{t}\sum_{\substack{j=0\\ j\neq i}}^{t} a^{3i+j} &= \frac{a(1-a^{3})+a^{3}-a^{4}}{(1-a)(1-a^{3})(1-a^{4})} = \frac{a(1+a^{2}-2a^{3})}{(1-a)(1-a^{3})(1-a^{4})} \\
&= \frac{a(1-a)(1+a+2a^{2}))}{(1-a)(1-a^{3})(1-a^{4})} = \frac{a(1+a+2a^{2}))}{(1-a^{3})(1-a^{4})} \qquad (34)
\end{aligned}$$

$$\begin{aligned}
\sum_{i=0}^{t-1}\sum_{j=i+1}^{t} a^{2i+2j} &= \sum_{i=0}^{t-1} a^{4i+2}\frac{1-a^{2(t-i)}}{1-a^{2}} \\
\lim_{t\to\infty}\sum_{i=0}^{t-1}\sum_{j=i+1}^{t} a^{2i+2j} &= \frac{a^{2}}{1-a^{2}}\sum_{i=0}^{t-1} a^{4i} \\
&= \frac{a^{2}}{(1-a^{2})(1-a^{4})} \qquad (35)
\end{aligned}$$

$$\begin{aligned}
\sum_{i=0}^{t}\sum_{\substack{j=0\\ j\neq i}}^{t-1}\sum_{\substack{k=j+1\\ k\neq i}}^{t} a^{2i+j+k} &= \sum_{i=2}^{t}\sum_{j=0}^{i-2}\sum_{k=j+1}^{i-1} a^{2i+j+k} + \sum_{i=1}^{t-1}\sum_{j=0}^{i-1}\sum_{k=i+1}^{t} a^{2i+j+k} \\
&\quad + \sum_{i=0}^{t-2}\sum_{j=i+1}^{t-1}\sum_{k=j+1}^{t} a^{2i+j+k} \qquad (36)
\end{aligned}$$

$$
\begin{aligned}
\sum_{i=2}^{t}\sum_{j=0}^{i-2}\sum_{k=j+1}^{i-1} a^{2i+j+k} &= \sum_{i=2}^{t}\sum_{j=0}^{i-2} a^{2i+2j+1}\frac{1-a^{i-j-1}}{1-a} \\
&= \frac{1}{1-a}\sum_{i=2}^{t} a^{2i+1}\frac{1-a^{2(i-1)}}{1-a^2} - a^{3i}\frac{1-a^{i-1}}{1-a} \\
&= \frac{1}{1-a}\left(\frac{a^5}{1-a^2}\frac{1-a^{2(t-1)}}{1-a^2} - \frac{a^7}{(1-a^2)}\frac{1-a^{4(t-1)}}{1-a^4}\right. \\
&\quad \left. - \frac{a^6}{1-a}\frac{1-a^{3(t+1)}}{1-a^3} + \frac{a^7}{1-a}\frac{1-a^{4(t+1)}}{1-a^4}\right) \\
&\xrightarrow[t\to\infty]{} \frac{a^5}{1-a}\left(\frac{1}{(1-a^2)^2} - \frac{a^2}{(1-a^2)(1-a^4)}\right. \\
&\quad \left. - \frac{a}{(1-a)(1-a^3)} + \frac{a^2(1+a)}{(1+a)(1-a)(1-a^4)}\right) \\
&\xrightarrow[t\to\infty]{} \frac{a^5}{1-a}\left(\frac{(1+a^2)}{(1-a^2)^2(1+a^2)} + \frac{a^3}{(1-a^2)(1-a^4)}\right. \\
&\quad \left. - \frac{a}{(1-a)(1-a^3)}\right) \\
&\xrightarrow[t\to\infty]{} \frac{a^5}{1-a}\left(\frac{1+a^2+a^3}{(1+a)(1-a)(1-a^4)} - \frac{a}{(1-a)(1-a^3)}\right) \\
&\xrightarrow[t\to\infty]{} \frac{a^5}{(1-a)^2}\frac{(1+a^2+a^3)(1-a^3) - a(1+a)(1-a^4)))}{(1+a)(1-a^3)(1-a^4)} \\
&\xrightarrow[t\to\infty]{} a^5\frac{1+a^2-a^5-a^6-(a+a^2-a^5-a^6)}{(1-a)(1-a^2)(1-a^3)(1-a^4)} \\
&\xrightarrow[t\to\infty]{} \frac{a^5}{(1-a^2)(1-a^3)(1-a^4)}
\end{aligned} \tag{37}
$$

$$
\begin{aligned}
\sum_{i=1}^{t-1}\sum_{j=0}^{i-1}\sum_{k=i+1}^{t} a^{2i+j+k} &= \sum_{i=1}^{t-1}\sum_{j=0}^{i-1} a^{3i+j+1}\frac{1-a^{t-i}}{1-a} \\
&\xrightarrow[t\to\infty]{} \lim_{t\to\infty}\frac{a}{1-a}\sum_{i=1}^{t-1} a^{3i}\frac{1-a^i}{1-a} \\
&\xrightarrow[t\to\infty]{} \lim_{t\to\infty}\frac{a}{(1-a)^2}\left(a^3\frac{1-a^{3t}}{1-a^3} - a^4\frac{1-a^{4t}}{1-a^4}\right) \\
&\xrightarrow[t\to\infty]{} \frac{a}{(1-a)^2}\left(\frac{a^3(1-a^4)-a^4(1-a^3)}{(1-a^3)(1-a^4)}\right) \\
&\xrightarrow[t\to\infty]{} \frac{a^4-a^5}{(1-a)^2(1-a^3)(1-a^4)} = \frac{a^4}{(1-a)(1-a^3)(1-a^4)}
\end{aligned} \tag{38}
$$

$$\begin{aligned}\sum_{i=0}^{t-2}\sum_{j=i+1}^{t-1}\sum_{k=j+1}^{t} a^{2i+j+k} &= \sum_{i=0}^{t-2}\sum_{j=i+1}^{t-1} a^{2i+2j+1}\frac{1-a^{t-j}}{1-a}\\ &\xrightarrow[t\to\infty]{} \lim_{t\to\infty}\frac{a}{1-a}\sum_{i=0}^{t-2} a^{4i+2}\frac{1-a^{2(t-i-1)}}{1-a^2}\\ &\xrightarrow[t\to\infty]{} \lim_{t\to\infty}\frac{a^3}{(1-a)(1-a^2)}\frac{1-a^{4(t-1)}}{1-a^4}\\ &\xrightarrow[t\to\infty]{} \frac{a^3}{(1-a)(1-a^2)(1-a^4)}\end{aligned} \tag{39}$$

We now combine Eq (37), Eq. (38) and Eq. (37) in Eq. (36).

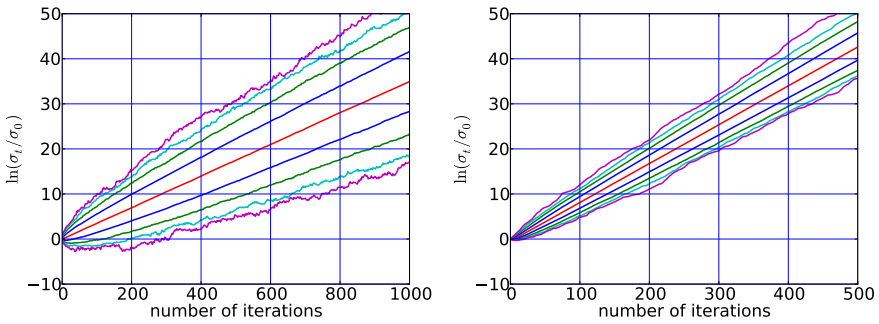
$$\begin{aligned}\sum_{i=0}^{t}\sum_{\substack{j=0\\ j\neq i}}^{t-1}\sum_{\substack{k=j+1\\ k\neq i}}^{t} a^{2i+j+k} &\xrightarrow[t\to\infty]{} \frac{a^5(1-a)+a^4(1-a^2)+a^3(1-a^3)}{(1-a)(1-a^2)(1-a^3)(1-a^4)}\\ &\xrightarrow[t\to\infty]{} \frac{a^3+a^4+a^5-3a^6}{(1-a)(1-a^2)(1-a^3)(1-a^4)}\\ &\xrightarrow[t\to\infty]{} \frac{a^3(1+2a+3a^2)}{((1-a^2)(1-a^3)(1-a^4)}\end{aligned} \tag{40}$$

$$\begin{aligned}\sum_{i=0}^{t-3}\sum_{j=i+1}^{t-2}\sum_{k=j+1}^{t-1}\sum_{l=k+1}^{t} a^{i+j+k+l} &= \sum_{i=0}^{t-3}\sum_{j=i+1}^{t-2}\sum_{k=j+1}^{t-1} a^{i+j+2k+1}\frac{1-a^{t-k}}{1-a}\\ &\xrightarrow[t\to\infty]{} \lim_{t\to\infty}\frac{a}{1-a}\sum_{i=0}^{t-3}\sum_{j=i+1}^{t-2} a^{i+3j+2}\frac{1-a^{2(t-1-j)}}{1-a^2}\\ &\xrightarrow[t\to\infty]{} \lim_{t\to\infty}\frac{a^3}{(1-a)(1-a^2)}\sum_{i=0}^{t-3} a^{4i+3}\frac{1-a^{3(t-2-i)}}{1-a^3}\\ &\xrightarrow[t\to\infty]{} \lim_{t\to\infty}\frac{a^6}{(1-a)(1-a^2)(1-a^3)}\frac{1-a^{4(t-2)}}{1-a^4}\\ &\xrightarrow[t\to\infty]{} \frac{a^6}{(1-a)(1-a^2)(1-a^3)(1-a^4)}\end{aligned} \tag{41}$$

By factorising Eq. (30), Eq. (34), Eq. (35), Eq. (40) and Eq. (41) by $\frac{1}{1-a^4}$ we get the coefficients of Theorem 4. □

Figure 1 shows the time evolution of $\ln(\sigma_t/\sigma_0)$ for 5001 runs and $c = 1$ (left) and $c = 1/\sqrt{n}$ (right). By comparing Figure 1a and Figure 1b we observe smaller variations of $\ln(\sigma_t/\sigma_0)$ with the smaller value of c.

Figure 2 shows the relative standard deviation of $\ln(\sigma_{t+1}/\sigma_t)$ (i.e. the standard deviation divided by its expected value). Lowering c, as shown in the left, decreases



(a) Without cumulation ($c = 1$) (b) With cumulation ($c = 1/\sqrt{20}$)

Fig. 1: $\ln(\sigma_t/\sigma_0)$ against t. The different curves represent the quantiles of a set of $5.10^3 + 1$ samples, more precisely the 10^i-quantile and the $1 - 10^{-i}$-quantile for i from 1 to 4; and the median. We have $n = 20$ and $\lambda = 8$.

the relative standard deviation. To get a value below one, c must be smaller for larger dimension. In agreement with Theorem 4, In Figure 2, right, the relative standard deviation increases like $\sqrt{n}$ with the dimension for constant c (three increasing curves). A careful study [8] of the variance equation of Theorem 4 shows that for the choice of $c = 1/(1 + n^\alpha)$, if $\alpha > 1/3$ the relative standard deviation converges to 0 with $\sqrt{(n^{2\alpha} + n)/n^{3\alpha}}$. Taking $\alpha = 1/3$ is a critical value where the relative standard deviation converges to $1/(\sqrt{2}\mathbb{E}(\mathcal{N}_{1:\lambda})^2)$. On the other hand, lower values of α makes the relative standard deviation diverge with $n^{(1-3\alpha)/2}$.

6 Summary

We investigate throughout this paper the $(1, \lambda)$-CSA-ES on affine linear functions composed with strictly increasing transformations. We find, in Theorem 3, the limit distribution for $\ln(\sigma_t/\sigma_0)/t$ and rigorously prove the desired behaviour of σ with $\lambda \geq 3$ for any c, and with $\lambda = 2$ and cumulation ($0 < c < 1$): the step-size diverges geometrically fast. In contrast, without cumulation ($c = 1$) and with $\lambda = 2$, a random walk on $\ln(\sigma)$ occurs, like for the $(1, 2)$-σSA-ES [9] (and also for the same symmetry reason). We derive an expression for the variance of the step-size increment. On linear functions when $c = 1/n^\alpha$, for $\alpha \geq 0$ ($\alpha = 0$ meaning c constant) and for $n \to \infty$ the standard deviation is about $\sqrt{(n^{2\alpha} + n)/n^{3\alpha}}$ times larger than the step-size increment. From this follows that keeping $c < 1/n^{1/3}$ ensures that the standard deviation of $\ln(\sigma_{t+1}/\sigma_t)$ becomes negligible compared to $\ln(\sigma_{t+1}/\sigma_t)$ when the dimensions goes to infinity. That means, the signal to noise ratio goes to zero, giving the algorithm strong stability. The result confirms that even the largest default cumulation parameter $c = 1/\sqrt{n}$ is a stable choice.

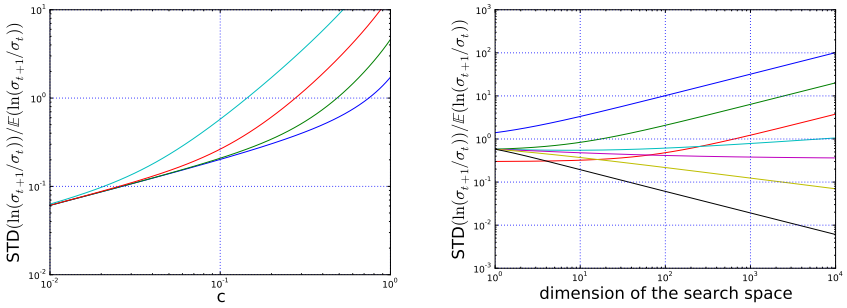


Fig. 2: Standard deviation of $\ln(\sigma_{t+1}/\sigma_t)$ relatively to its expectation. Here $\lambda = 8$. The curves were plotted using Eq. (24) and Eq. (25). On the left, curves for (right to left) $n = 2$, 20, 200 and 2000. On the right, different curves for (top to bottom) $c = 1$, 0.5, 0.2, $1/(1+n^{1/4})$, $1/(1+n^{1/3})$, $1/(1+n^{1/2})$ and $1/(1+n)$.

Acknowledgments

This work was partially supported by the ANR-2010-COSI-002 grant (SIMINOLE) of the French National Research Agency and the ANR COSINUS project ANR-08-COSI-007-12.

References

1. Dirk V Arnold. Cumulative step length adaptation on ridge functions. In *Parallel Problem Solving from Nature PPSN IX*, pages 11–20. Springer, 2006.
2. Dirk V Arnold and H-G Beyer. Performance analysis of evolutionary optimization with cumulative step length adaptation. *IEEE Transactions on Automatic Control*, 49(4):617–622, 2004.
3. Dirk V Arnold and Hans-Georg Beyer. On the behaviour of evolution strategies optimising cigar functions. *Evolutionary Computation*, 18(4):661–682, 2010.
4. D.V. Arnold. On the behaviour of the (1,λ)-ES for a simple constrained problem. In *Foundations of Genetic Algorithms - FOGA 11*, pages 15–24. ACM, 2011.
5. D.V. Arnold and H.G. Beyer. Random dynamics optimum tracking with evolution strategies. In *Parallel Problem Solving from Nature - PPSN VII*, pages 3–12. Springer, 2002.
6. D.V. Arnold and H.G. Beyer. Optimum tracking with evolution strategies. *Evolutionary Computation*, 14(3):291–308, 2006.
7. D.V. Arnold and H.G. Beyer. Evolution strategies with cumulative step length adaptation on the noisy parabolic ridge. *Natural Computing*, 7(4):555–587, 2008.
8. A. Chotard, A. Auger, and N. Hansen. Cumulative step-size adaptation on linear functions: Technical report. Technical report, Inria, 2012. http://www.lri.fr/~chotard/chotard2012TRcumulative.pdf.
9. N. Hansen. An analysis of mutative σ-self-adaptation on linear fitness functions. *Evolutionary Computation*, 14(3):255–275, 2006.
10. N. Hansen and A. Ostermeier. Completely derandomized self-adaptation in evolution strategies. *Evolutionary Computation*, 9(2):159–195, 2001.

11. Nikolaus Hansen, Andreas Gawelczyk, and Andreas Ostermeier. Sizing the population with respect to the local progress in (1, lambda)-evolution strategies - a theoretical analysis. In *In Proceedings of the 1995 IEEE Conference on Evolutionary Computation*, pages 80–85. IEEE Press, 1995.
12. S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, second edition, 1993.
13. A. Ostermeier, A. Gawelczyk, and N. Hansen. Step-size adaptation based on non-local use of selection information. In *Proceedings of Parallel Problem Solving from Nature — PPSN III*, volume 866 of *Lecture Notes in Computer Science*, pages 189–198. Springer, 1994.

4.3 Linear Functions with Linear Constraints

In this section we present two analyses of $(1,\lambda)$-ESs on a linear function f with a linear constraint g, where the constraint is handled through the resampling of unfeasible points until λ feasible points have been sampled. The problem reads

$$\begin{aligned} &\text{maximize } f(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \mathbb{R}^n \\ &\text{subject to } g(\boldsymbol{x}) \geq 0 \quad . \end{aligned}$$

An important characteristic of the problem is the constraint angle $(\nabla f, -\nabla g)$, denoted θ. The two analyses study the problem for values of θ in $(0, \pi/2)$; lower values of θ correspond to a higher conflict between the objective function and the constraint, making the problem more difficult. Linear constraints are a very frequent type of constraint (e.g. non-negative variables from problems in physics or biology). Despite that a linear function with a linear constraint seems to be an easy problem, a $(1,\lambda)$-ES with σSA or CSA step-size adaptation fails to solve the problem when the value of the constraint angle θ is too low [14, 15].
This section first presents a study of a $(1,\lambda)$-ES with constant step-size and with cumulative step-size adaptation (as defined with (2.12) and (4.11)) on a linear function with a linear constraint. Then this section presents a study of a $(1,\lambda)$-ES with constant step-size and with a general sampling distribution that can be non-Gaussian on a linear function with a linear constraint.

4.3.1 Paper: Markov Chain Analysis of Cumulative Step-size Adaptation on a Linear Constraint Problem

The article presented here [45] has been accepted for publication at the Evolutionary Computation Journal in 2015, and is an extension of [44] which was published at the conference Congress on Evolutionary Computation in 2014. It was inspired by [14], which assumes the positivity (i.e. the existence of an invariant probability measure) of the sequence $(\delta_t)_{t\in\mathbb{N}}$ defined as a signed distance from the mean of the sampling distribution to the constraint normalized by the step-size, i.e. $\delta_t := -g(\boldsymbol{X}_t)/\sigma_t$. The results of [14] are obtained through a few approximations (mainly, the invariant distribution π of $(\delta_t)_{t\in\mathbb{N}}$ is approximated as a Dirac distribution at $\mathbf{E}_\pi(\delta_t)$) and the accuracy of these results is then verified through Monte Carlo simulations. The ergodicity of the sequence $(\delta_t)_{t\in\mathbb{N}}$ studied is therefore a crucial underlying hypothesis since it justifies that the Monte Carlo simulations do converge independently of their initialization to what they are supposed to measure. Therefore we aim in the following paper to prove the ergodicity of the sequence $(\delta_t)_{t\in\mathbb{N}}$.
Note that the problem of a linear function with a linear constraint is much more complex than in the unconstrained case of Section 4.2, due to the fact that the sequence of random vectors $(\boldsymbol{N}_t^\star)_{t\in\mathbb{N}} := ((\boldsymbol{X}_{t+1} - \boldsymbol{X}_t)/\sigma_t)_{t\in\mathbb{N}}$ is not i.i.d., contrarily as in Section 4.2. Instead, the distribution of $\boldsymbol{N}_t^\star$ can be shown to be a function of δ_t, and to prove the log-linear divergence or convergence of the step-size for a $(1,\lambda)$-CSA-ES, a study of the full Markov chain $(\delta_t, \boldsymbol{p}_{t+1}^\sigma)_{t\in\mathbb{N}}$ is required. Furthermore, due to the resampling, sampling $\boldsymbol{N}_t^\star$ involves an unbounded number of samples.

The problem being complex, the paper [44] starts by studying the more simple $(1,\lambda)$-ES with constant step-size in order to investigate the problem and establish a methodology. In order to avoid any problem with the unbounded number of samples required to sample $\boldsymbol{N}_t^\star$, the paper considers a random vector sampled with a constant and bounded number of random variables, and which is equal in distribution to $\boldsymbol{N}_t^\star$. The paper shows in this context the geometric ergodicity of the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ and that the sequence $(f(\boldsymbol{X}_t))_{t\in\mathbb{N}}$ diverges in probability to $+\infty$ at constant speed $s>0$, i.e.

$$\frac{f(\boldsymbol{X}_t)-f(\boldsymbol{X}_0)}{t} \xrightarrow[t\to+\infty]{P} s>0 \ . \tag{4.12}$$

Note that the divergence cannot be log-linear since the step-size is kept constant. The paper then sketches results for the $(1,\lambda)$-CSA-ES on the same problem.
In [45], which is the article presented here, the $(1,\lambda)$-CSA-ES is also investigated. The article shows that $(\delta_t,\boldsymbol{p}_t^\sigma)_{t\in\mathbb{N}}$ is a time-homogeneous Markov chain, which for reasons suggested in Section 4.1 is difficult to analyse. Therefore the paper analyses the algorithm in the simpler case where the cumulation parameter c_σ equals to 1, which implies that the evolution path $\boldsymbol{p}_{t+1}^\sigma$ equals $\boldsymbol{N}_t^\star$, and that $(\delta_t)_{t\in\mathbb{N}}$ is a Markov chain. The Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is shown to be a geometrically ergodic Markov chain, from which it is deduced that the step-size diverges or converges log-linearly in probability to a rate r whose sign indicates whether divergence or convergence occurs. This rate is then estimated through Monte Carlo simulations, and its dependence with the constraint angle θ and parameters of the algorithm such as the population size λ and the cumulation parameter c_σ is investigated. It appears that for small values of θ this rate is negative which shows the log-linear convergence of the algorithm, although this effect can be countered by taking large enough values of λ or low enough values of c_σ. Hence critical values of λ and c_σ, between the values implying convergence and the values implying divergence, exist, and are plotted as a function of the constraint angle θ.

Markov Chain Analysis of Cumulative Step-size Adaptation on a Linear Constrained Problem

Alexandre Chotard chotard@lri.fr
Univ. Paris-Sud, LRI, Rue Noetzlin, Bat 660, 91405 Orsay Cedex France

Anne Auger auger@lri.fr
Inria, Univ. Paris-Sud, LRI, Rue Noetzlin, Bat 660, 91405 Orsay Cedex France

Nikolaus Hansen hansen@lri.fr
Inria, Univ. Paris-Sud, LRI, Rue Noetzlin, Bat 660, 91405 Orsay Cedex France

Abstract

This paper analyses a $(1, \lambda)$-Evolution Strategy, a randomised comparison-based adaptive search algorithm, optimizing a linear function with a linear constraint. The algorithm uses resampling to handle the constraint. Two cases are investigated: first the case where the step-size is constant, and second the case where the step-size is adapted using cumulative step-size adaptation. We exhibit for each case a Markov chain describing the behavior of the algorithm. Stability of the chain implies, by applying a law of large numbers, either convergence or divergence of the algorithm. Divergence is the desired behavior. In the constant step-size case, we show stability of the Markov chain and prove the divergence of the algorithm. In the cumulative step-size adaptation case, we prove stability of the Markov chain in the simplified case where the cumulation parameter equals 1, and discuss steps to obtain similar results for the full (default) algorithm where the cumulation parameter is smaller than 1. The stability of the Markov chain allows us to deduce geometric divergence or convergence, depending on the dimension, constraint angle, population size and damping parameter, at a rate that we estimate. Our results complement previous studies where stability was assumed.

Keywords

Continuous Optimization, Evolution Strategies, CMA-ES, Cumulative Step-size Adaptation, Constrained problem.

1 Introduction

Derivative Free Optimization (DFO) methods are tailored for the optimization of numerical problems in a black-box context, where the objective function $f : \mathbb{R}^n \to \mathbb{R}$ is pictured as a black-box that *solely* returns f values (in particular no gradients are available).

Evolution Strategies (ES) are *comparison-based* randomised DFO algorithms. At iteration t, solutions are sampled from a multivariate normal distribution centered in a vector $\mathbf{X}_t$. The candidate solutions are ranked according to f, and the updates of $\mathbf{X}_t$ and other parameters of the distribution (usually a step-size σ_t and a covariance matrix) are performed using solely the ranking information given by the candidate solutions. Since ES do not directly use the function values of the new points, but only how the objective function f ranks the different samples, they are invariant to the composition (to the left) of the objective function by a strictly increasing function $h : \mathbb{R} \to \mathbb{R}$.

This property and the black-box scenario make Evolution Strategies suited for a wide class of real-world problems, where constraints on the variables are often imposed. Different techniques for handling constraints in randomised algorithms have been proposed, see (Mezura-Montes and Coello, 2011) for a survey. For ES, common techniques are resampling, i.e. resample a solution until it lies in the feasible domain, repair of solutions that project unfeasible points onto the feasible domain (e.g. (Arnold, 2011b, 2013)), penalty methods where unfeasible solutions are penalised either by a quantity that depends on the distance to the constraint (e.g. (Hansen et al., 2009) with adaptive penalty weights) (if this latter one can be computed) or by the constraint value itself (e.g. stochastic ranking (Runarsson and Yao, 2000)) or methods inspired from multi-objective optimization (e.g. (Mezura-Montes and Coello, 2008)).

In this paper we focus on the resampling method and study it on a simple constrained problem. More precisely, we study a $(1, \lambda)$-ES optimizing *a linear function with a linear constraint* and resampling any infeasible solution until a feasible solution is sampled. The linear function models the situation where the current point is, relatively to the step-size, far from the optimum and "solving" this function means diverging. The linear constraint models being close to the constraint relatively to the step-size and far from other constraints. Due to the invariance of the algorithm to the composition of the objective function by a strictly increasing map, the linear function could be composed by a function without derivative and with many discontinuities without any impact on our analysis.

The problem we address was studied previously for different step-size adaptation mechanisms and different constraint handling methods: with constant step-size, self-adaptation, and cumulative step-size adaptation, resampling or repairing unfeasible solutions (Arnold, 2011a, 2012, 2013). The drawn conclusion is that when adapting the step-size the $(1, \lambda)$-ES fails to diverge unless some requirements on internal parameters of the algorithm are met. However, the approach followed in the aforementioned studies relies on finding simplified theoretical models to explain the behaviour of the algorithm: typically those models arise by doing some approximations (considering some random variables equal to their expected value, etc.) and assuming some mathematical properties like the existence of stationary distributions of underlying Markov chains.

In contrast, our motivation is to study the real—i.e., not simplified—-algorithm and prove rigorously different mathematical properties of the algorithm allowing to deduce the exact behaviour of the algorithm, as well as to provide tools and methodology for such studies. Our theoretical studies need to be complemented by simulations of the convergence/divergence rates. The mathematical properties that we derive show that these numerical simulations converge fast. Our results are largely in agreement with the aforementioned studies of simplified models thereby backing up their validity.

As for the step-size adaptation mechanism, our aim is to study the cumulative step-size adaptation (CSA) also called path-length control, default step-size mechanism for the CMA-ES algorithm (Hansen and Ostermeier, 2001). The mathematical object to study for this purpose is a discrete time, continuous state space Markov chain that is defined as the pair: evolution path and distance to the constraint normalized by the step-size. More precisely stability properties like irreducibility, existence of a stationary distribution of this Markov chain need to be studied to deduce the geometric divergence of the CSA and have a rigorous mathematical framework to perform Monte Carlo simulations allowing to study the influence of different parameters of the algorithm. We start by illustrating in details the methodology on the simpler case where the

step-size is constant. We show in this case that the distance to the constraint reaches a stationary distribution. This latter property was assumed in a previous study (see (Arnold, 2011a)). We then prove that the algorithm diverges at a constant speed. We then apply this approach to the case where the step-size is adapted using path length control. We show that in the special case where the cumulation parameter c equals to 1, the expected logarithmic step-size change, $\mathbf{E}\ln(\sigma_{t+1}/\sigma_t)$, converges to a constant r, and the average logarithmic step-size change, $\ln(\sigma_t/\sigma_0)/t$, converges in probability to the same constant, which depends on parameters of the problem and of the algorithm. This implies geometric divergence (if $r > 0$) or convergence (if $r < 0$) at the rate r that we estimate.

This paper is organized as follows. In Section 2 we define the $(1, \lambda)$-ES using resampling and the problem. In Section 3 we provide some preliminary derivations on the distributions that come into play for the analysis. In Section 4 we analyze the constant step-size case. In Section 5 we analyse the cumulative step-size adaptation case. Finally we discuss our results and our methodology in Section 6.

A preliminary version of this paper appeared in the conference proceedings (Chotard et al., 2014). The analysis of path-length control with cumulation parameter equal to 1 is however fully new, as well as the discussion on how to analyze the case with cumulation parameter smaller than one. Also Figures 4–11 are new as well as the convergence of the progress rate in Theorem 1.

Notations

Throughout this article, we denote by φ the density function of the standard multivariate normal distribution (the dimension being clarified within the context), and Φ the cumulative distribution function of a standard univariate normal distribution. The standard (unidimensional) normal distribution is denoted $\mathcal{N}(0, 1)$, the (n-dimensional) multivariate normal distribution with covariance matrix identity is denoted $\mathcal{N}(\mathbf{0}, \mathrm{Id}_n)$ and the i^{th} order statistic of λ i.i.d. standard normal random variables is denoted $\mathcal{N}_{i:\lambda}$. The uniform distribution on an interval I is denoted $\mathcal{U}_I$. We denote μ_{Leb} the Lebesgue measure. The set of natural numbers (including 0) is denoted $\mathbb{N}$, and the set of real numbers $\mathbb{R}$. We denote $\mathbb{R}_+$ the set $\{x \in \mathbb{R} | x \geq 0\}$, and for $A \subset \mathbb{R}^n$, the set A^* denotes $A \backslash \{\mathbf{0}\}$ and $\mathbf{1}_A$ denotes the indicator function of A. For two vectors $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$, we denote $[\mathbf{x}]_i$ the i^{th}-coordinate of $\mathbf{x}$, and $\mathbf{x}.\mathbf{y}$ the scalar product of $\mathbf{x}$ and $\mathbf{y}$. Take $(a, b) \in \mathbb{N}^2$ with $a \leq b$, we denote $[a..b]$ the interval of integers between a and b. For a topological set $\mathcal{X}$, $\mathcal{B}(\mathcal{X})$ denotes the Borel algebra of $\mathcal{X}$. For $\mathbf{X}$ and $\mathbf{Y}$ two random vectors, we denote $\mathbf{X} \overset{d}{=} \mathbf{Y}$ if $\mathbf{X}$ and $\mathbf{Y}$ are equal in distribution. For $(X_t)_{t\in\mathbb{N}}$ a sequence of random variables and X a random variable we denote $X_t \overset{a.s.}{\to} X$ if X_t converges almost surely to X and $X_t \overset{P}{\to} X$ if X_t converges in probability to X.

2 Problem statement and algorithm definition

2.1 $(1, \lambda)$-ES with resampling

In this paper, we study the behaviour of a $(1, \lambda)$-Evolution Strategy *maximizing* a function f: $\mathbb{R}^n \to \mathbb{R}$, $\lambda \geq 2$, $n \geq 2$, with a constraint defined by a function $g : \mathbb{R}^n \to \mathbb{R}$ restricting the feasible space to $X_{\text{feasible}} = \{\mathbf{x} \in \mathbb{R}^n | g(\mathbf{x}) > 0\}$. To handle the constraint, the algorithm resamples any unfeasible solution until a feasible solution is found.

From iteration $t \in \mathbb{N}$, given the vector $\mathbf{X}_t \in \mathbb{R}^n$ and step-size $\sigma_t \in \mathbb{R}_+^*$, the algo-

A. Chotard, A. Auger, N. Hansen

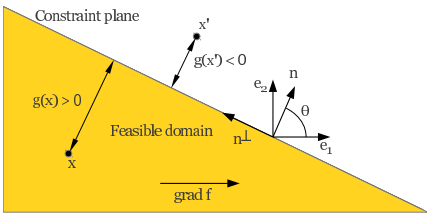


Figure 1: Linear function with a linear constraint, in the plane generated by ∇f and $\mathbf{n}$, a normal vector to the constraint hyperplane with angle $\theta \in (0, \pi/2)$ with ∇f. The point $\mathbf{x}$ is at distance $g(\mathbf{x})$ from the constraint.

rithm generates λ new candidates:

$$\mathbf{Y}_t^i = \mathbf{X}_t + \sigma_t \mathbf{N}_t^i \ , \tag{1}$$

with $i \in [1..\lambda]$, and $(\mathbf{N}_t^i)_{i\in[1..\lambda]}$ i.i.d. standard multivariate normal random vectors. If a new sample $\mathbf{Y}_t^i$ lies outside the feasible domain, that is $g(\mathbf{Y}_t^i) \leq 0$, then it is resampled until it lies within the feasible domain. The first feasible i^{th} candidate solution is denoted $\tilde{\mathbf{Y}}_t^i$ and the realization of the multivariate normal distribution giving $\tilde{\mathbf{Y}}_t^i$ is $\tilde{\mathbf{N}}_t^i$, i.e.

$$\tilde{\mathbf{Y}}_t^i = \mathbf{X}_t + \sigma_t \tilde{\mathbf{N}}_t^i \tag{2}$$

The vector $\tilde{\mathbf{N}}_t^i$ is called a feasible step. Note that $\tilde{\mathbf{N}}_t^i$ is not distributed as a multivariate normal distribution, further details on its distribution are given later on.

We define $\star = \operatorname{argmax}_{i\in[1..\lambda]} f(\tilde{\mathbf{Y}}_t^i)$ as the index realizing the maximum objective function value, and call $\tilde{\mathbf{N}}_t^\star$ the selected step. The vector $\mathbf{X}_t$ is then updated as the solution realizing the maximum value of the objective function, i.e.

$$\mathbf{X}_{t+1} = \tilde{\mathbf{Y}}_t^\star = \mathbf{X}_t + \sigma_t \tilde{\mathbf{N}}_t^\star \ . \tag{3}$$

The step-size and other internal parameters are then adapted. We denote for the moment in a non specific manner the adaptation as

$$\sigma_{t+1} = \sigma_t \xi_t \tag{4}$$

where ξ_t is a random variable whose distribution is a function of the selected steps $(\tilde{\mathbf{N}}_i^\star)_{i\leq t}$, $\mathbf{X}_0$, σ_0 and of internal parameters of the algorithm. We will define later on specific rules for this adaptation.

2.2 Linear fitness function with linear constraint

In this paper, we consider the case where f, the function that we optimize, and g, the constraint, are linear functions. W.l.o.g., we assume that $\|\nabla f\| = \|\nabla g\| = 1$. We denote $\mathbf{n} := -\nabla g$ a normal vector to the constraint hyperplane. We choose an orthonormal Euclidean coordinate system with basis $(\mathbf{e}_i)_{i\in[1..n]}$ with its origin located on the constraint hyperplane where $\mathbf{e}_1$ is equal to the gradient ∇f, hence

$$f(\mathbf{x}) = [\mathbf{x}]_1 \tag{5}$$

and the vector $\mathbf{e}_2$ lives in the plane generated by ∇f and $\mathbf{n}$ and is such that the angle between $\mathbf{e}_2$ and $\mathbf{n}$ is positive. We define θ the angle between ∇f and $\mathbf{n}$, and restrict

our study to $\theta \in (0, \pi/2)$. The function g can be seen as a signed distance to the linear constraint as

$$g(\mathbf{x}) = \mathbf{x}.\nabla g = -\mathbf{x}.\mathbf{n} = -[\mathbf{x}]_1 \cos\theta - [\mathbf{x}]_2 \sin\theta \ . \tag{6}$$

A point is feasible if and only if $g(\mathbf{x})>0$ (see Figure 1). Overall the problem reads

$$\begin{array}{l} \text{maximize} \quad f(\mathbf{x}) = [\mathbf{x}]_1 \quad \text{subject to} \\ g(\mathbf{x}) = -[\mathbf{x}]_1 \cos\theta - [\mathbf{x}]_2 \sin\theta > 0 \ . \end{array} \tag{7}$$

Although $\tilde{\mathbf{N}}_t^i$ and $\tilde{\mathbf{N}}_t^\star$ are in $\mathbb{R}^n$, due to the choice of the coordinate system and the independence of the sequence $([\mathbf{N}_t^i]_k)_{k\in[1..n]}$, only the two first coordinates of these vectors are affected by the resampling implied by g and the selection according to f. Therefore $[\tilde{\mathbf{N}}_t^\star]_k \sim \mathcal{N}(0,1)$ for $k \in [3..n]$. With an abuse of notations, the vector $\tilde{\mathbf{N}}_t^i$ will denote the 2-dimensional vector $([\tilde{\mathbf{N}}_t^i]_1, [\tilde{\mathbf{N}}_t^i]_2)$, likewise $\tilde{\mathbf{N}}_t^\star$ will also denote the 2-dimensional vector $([\tilde{\mathbf{N}}_t^\star]_1, [\tilde{\mathbf{N}}_t^\star]_2)$, and $\mathbf{n}$ will denote the 2-dimensional vector $(\cos\theta, \sin\theta)$. The coordinate system will also be used as $(\mathbf{e}_1, \mathbf{e}_2)$ only.

Following (Arnold, 2011a, 2012; Arnold and Brauer, 2008), we denote the normalized signed distance to the constraint as δ_t, that is

$$\delta_t = \frac{g(\mathbf{X}_t)}{\sigma_t} \ . \tag{8}$$

We initialize the algorithm by choosing $\mathbf{X}_0 = -\mathbf{n}$ and $\sigma_0 = 1$, which implies that $\delta_0 = 1$.

3 Preliminary results and definitions

Throughout this section we derive the probability density functions of the random vectors $\tilde{\mathbf{N}}_t^i$ and $\tilde{\mathbf{N}}_t^\star$ and give a definition of $\tilde{\mathbf{N}}_t^i$ and of $\tilde{\mathbf{N}}_t^\star$ as a function of δ_t and of an i.i.d. sequence of random vectors.

3.1 Feasible steps

The random vector $\tilde{\mathbf{N}}_t^i$, the i^{th} feasible step, is distributed as the standard multivariate normal distribution truncated by the constraint, as stated in the following lemma.

Lemma 1. *Let a $(1,\lambda)$-ES with resampling optimize a function f under a constraint function g. If g is a linear form determined by a vector $\mathbf{n}$ as in* (6)*, then the distribution of the feasible step $\tilde{\mathbf{N}}_t^i$ only depends on the normalized distance to the constraint δ_t and its density given that δ_t equals δ reads*

$$p_\delta(\mathbf{x}) = \frac{\varphi(\mathbf{x})\mathbf{1}_{\mathbb{R}_+^*}(\delta - \mathbf{x}.\mathbf{n})}{\Phi(\delta)} \ . \tag{9}$$

Proof. A solution $\mathbf{Y}_t^i$ is feasible if and only if $g(\mathbf{Y}_t^i)>0$, which is equivalent to $-(\mathbf{X}_t + \sigma_t \mathbf{N}_t^i).\mathbf{n}>0$. Hence dividing by σ_t, a solution is feasible if and only if $\delta_t = -\mathbf{X}_t.\mathbf{n}/\sigma_t > \mathbf{N}_t^i.\mathbf{n}$. Since a standard multivariate normal distribution is rotational invariant, $\mathbf{N}_t^i.\mathbf{n}$ follows a standard (unidimensional) normal distribution. Hence the probability that a solution $\mathbf{Y}_t^i$ or a step $\mathbf{N}_t^i$ is feasible is given by

$$\Pr(\mathcal{N}(0,1) < \delta_t) = \Phi(\delta_t) \ .$$

Therefore the probability density function of the random variable $\tilde{\mathbf{N}}_t^i.\mathbf{n}$ for $\delta_t = \delta$ is $x \mapsto \varphi(x)\mathbf{1}_{\mathbb{R}_+^*}(\delta - x)/\Phi(\delta)$. For any vector $\mathbf{n}^\perp$ orthogonal to $\mathbf{n}$ the random variable

$\tilde{\mathbf{N}}_t^i.\mathbf{n}^\perp$ was not affected by the resampling and is therefore still distributed as a standard (unidimensional) normal distribution. With a change of variables using the fact that the standard multivariate normal distribution is rotational invariant we obtain the joint distribution of Eq. (9). □

Then the marginal density function $p_{1,\delta}$ of $[\tilde{\mathbf{N}}_t^i]_1$ can be computed by integrating Eq. (9) over $[\mathbf{x}]_2$ and reads

$$p_{1,\delta}(x) = \varphi(x) \frac{\Phi\left(\frac{\delta - x\cos\theta}{\sin\theta}\right)}{\Phi(\delta)} \quad , \tag{10}$$

(see (Arnold, 2011a, Eq. 4) for details) and we denote $F_{1,\delta}$ its cumulative distribution function.

It will be important in the sequel to be able to express the vector $\tilde{\mathbf{N}}_t^i$ as a function of δ_t and of a *finite* number of random samples. Hence we give an alternative way to sample $\tilde{\mathbf{N}}_t^i$ rather than the resampling technique that involves an unbounded number of samples.

Lemma 2. *Let a $(1,\lambda)$-ES with resampling optimize a function f under a constraint function g, where g is a linear form determined by a vector $\mathbf{n}$ as in* (6)*. Let the feasible step $\tilde{\mathbf{N}}_t^i$ be the random vector described in Lemma 1 and $\mathbf{Q}$ be the 2-dimensional rotation matrix of angle θ. Then*

$$\tilde{\mathbf{N}}_t^i \overset{d}{=} \tilde{F}_{\delta_t}^{-1}(U_t^i)\mathbf{n} + \mathcal{N}_t^i\mathbf{n}^\perp = \mathbf{Q}^{-1}\begin{pmatrix}\tilde{F}_{\delta_t}^{-1}(U_t^i)\\ \mathcal{N}_t^i\end{pmatrix} \tag{11}$$

where $\tilde{F}_{\delta_t}^{-1}$ denotes the generalized inverse of the cumulative distribution of $\tilde{\mathbf{N}}_t^i.\mathbf{n}$ [1]*, $U_t^i \sim \mathcal{U}_{[0,1]}$, $\mathcal{N}_t^i \sim \mathcal{N}(0,1)$ with $(U_t^i)_{i\in[1..\lambda],t\in\mathbb{N}}$ i.i.d. and $(\mathcal{N}_t^i)_{i\in[1..\lambda],t\in\mathbb{N}}$ i.i.d. random variables.*

Proof. We define a new coordinate system $(\mathbf{n},\mathbf{n}^\perp)$ (see Figure 1). It is the image of $(\mathbf{e}_1,\mathbf{e}_2)$ by $\mathbf{Q}$. In the new basis $(\mathbf{n},\mathbf{n}^\perp)$, only the coordinate along $\mathbf{n}$ is affected by the resampling. Hence the random variable $\tilde{\mathbf{N}}_t^i.\mathbf{n}$ follows a truncated normal distribution with cumulative distribution function $\tilde{F}_{\delta_t}$ equal to $\min(1,\Phi(x)/\Phi(\delta_t))$, while the random variable $\tilde{\mathbf{N}}_t^i.\mathbf{n}^\perp$ follows an independent standard normal distribution, hence $\tilde{\mathbf{N}}_t^i \overset{d}{=} (\tilde{\mathbf{N}}_t^i.\mathbf{n})\mathbf{n} + \mathcal{N}_t^i\mathbf{n}^\perp$. Using the fact that if a random variable has a cumulative distribution F, then for F^{-1} the generalized inverse of F, $F^{-1}(U)$ with $U \sim \mathcal{U}_{[0,1]}$ has the same distribution as this random variable, we get that $\tilde{F}_{\delta_t}^{-1}(U_t^i) \overset{d}{=} \tilde{\mathbf{N}}_t^i.\mathbf{n}$, so we obtain Eq. (11). □

We now extend our study to the selected step $\tilde{\mathbf{N}}_t^\star$.

3.2 Selected step

The selected step $\tilde{\mathbf{N}}_t^\star$ is chosen among the different feasible steps $(\tilde{\mathbf{N}}_t^i)_{i\in[1..\lambda]}$ to maximize the function f, and has the density described in the following lemma.

Lemma 3. *Let a $(1,\lambda)$-ES with resampling optimize the problem* (7)*. Then the distribution of the selected step $\tilde{\mathbf{N}}_t^\star$ only depends on the normalized distance to the constraint δ_t and its density*

[1] The generalized inverse of $\tilde{F}_\delta$ is $\tilde{F}_\delta^{-1}(y) := \inf_{x\in\mathbb{R}}\{\tilde{F}_\delta(x) \geq y\}$.

given that δ_t equals δ reads

$$p^\star_\delta(\mathbf{x}) = \lambda p_\delta(\mathbf{x})\, F_{1,\delta}([\mathbf{x}]_1)^{\lambda-1} \ , \tag{12}$$
$$= \lambda \frac{\varphi(\mathbf{x})\mathbf{1}_{\mathbb{R}^*_+}(\delta - \mathbf{x}.\mathbf{n})}{\Phi(\delta)} \left(\int_{-\infty}^{[\mathbf{x}]_1} \varphi(u) \frac{\Phi(\frac{\delta - u\cos\theta}{\sin\theta})}{\Phi(\delta)} \mathrm{d}u \right)^{\lambda-1}$$

where p_δ is the density of $\tilde{\mathbf{N}}^i_t$ given that $\delta_t = \delta$ given in Eq. (9) *and $F_{1,\delta}$ the cumulative distribution function of $[\tilde{\mathbf{N}}^i_t]_1$ whose density is given in Eq.* (10) *and* $\mathbf{n}$ *the vector* $(\cos\theta, \sin\theta)$.

Proof. The function f being linear, the rankings on $(\tilde{\mathbf{N}}^i_t)_{i\in[1..\lambda]}$ correspond to the order statistic on $([\tilde{\mathbf{N}}^i_t]_1)_{i\in[1..\lambda]}$. If we look at the joint cumulative distribution $F^\star_\delta$ of $\tilde{\mathbf{N}}^\star_t$

$$F^\star_\delta(x,y) = \Pr\left([\tilde{\mathbf{N}}^\star_t]_1 \le x, [\tilde{\mathbf{N}}^\star_t]_2 \le y \right)$$
$$= \sum_{i=1}^{\lambda} \Pr\left(\tilde{\mathbf{N}}^i_t \le \begin{pmatrix} x \\ y \end{pmatrix}, [\tilde{\mathbf{N}}^j_t]_1 < [\tilde{\mathbf{N}}^i_t]_1 \text{ for } j \ne i \right)$$

by summing disjoints events. The vectors $(\tilde{\mathbf{N}}^i_t)_{i\in[1..\lambda]}$ being independent and identically distributed

$$F^\star_\delta(x,y) = \lambda \Pr\left(\tilde{\mathbf{N}}^1_t \le \begin{pmatrix} x \\ y \end{pmatrix}, [\tilde{\mathbf{N}}^j_t]_1 < [\tilde{\mathbf{N}}^1_t]_1 \text{ for } j \ne 1 \right)$$
$$= \lambda \int_{-\infty}^{x}\int_{-\infty}^{y} p_\delta(u,v) \prod_{j=2}^{\lambda} \Pr([\tilde{\mathbf{N}}^j_t]_1 < u) \mathrm{d}v\mathrm{d}u$$
$$= \lambda \int_{-\infty}^{x}\int_{-\infty}^{y} p_\delta(u,v) F_{1,\delta}(u)^{\lambda-1} \mathrm{d}v\mathrm{d}u \ .$$

Deriving $F^\star_\delta$ on x and y yields the density of $\tilde{\mathbf{N}}^\star_t$ of Eq. (12). □

We may now obtain the marginal of $[\tilde{\mathbf{N}}^\star_t]_1$ and $[\tilde{\mathbf{N}}^\star_t]_2$.

Corollary 1. *Let a $(1,\lambda)$-ES with resampling optimize the problem* (7)*. Then the marginal distribution of $[\tilde{\mathbf{N}}^\star_t]_1$ only depends on δ_t and its density given that δ_t equals δ reads*

$$p^\star_{1,\delta}(x) = \lambda p_{1,\delta}(x) F_{1,\delta}(x)^{\lambda-1} \ , \tag{13}$$
$$= \lambda\varphi(x) \frac{\Phi\left(\frac{\delta - x\cos\theta}{\sin\theta}\right)}{\Phi(\delta)} F_{1,\delta}(x)^{\lambda-1} \ ,$$

and the same holds for $[\tilde{\mathbf{N}}^\star_t]_2$ whose marginal density reads

$$p^\star_{2,\delta}(y) = \lambda \frac{\varphi(y)}{\Phi(\delta)} \int_{-\infty}^{\frac{\delta - y\sin\theta}{\cos\theta}} \varphi(u) F_{1,\delta}(u)^{\lambda-1} \mathrm{d}u \ . \tag{14}$$

Proof. Integrating Eq. (12) directly yields Eq. (13).

The conditional density function of $[\tilde{\mathbf{N}}^\star_t]_2$ is

$$p^\star_{2,\delta}(y | [\tilde{\mathbf{N}}^\star_t]_1 = x) = \frac{p^\star_\delta((x,y))}{p^\star_{1,\delta}(x)} \ .$$

As $p^{\star}_{2,\delta}(y) = \int_{\mathbb{R}} p^{\star}_{2,\delta}(y|[\tilde{\mathbf{N}}^{\star}_t]_1 = x) p^{\star}_{1,\delta}(x)\mathrm{d}x$, using the previous equation with Eq. (12) gives that $p^{\star}_{2,\delta}(y) = \int_{\mathbb{R}} \lambda p_{\delta}((x,y)) F_{1,\delta}(x)^{\lambda-1}\mathrm{d}x$, which with Eq. (9) gives

$$p^{\star}_{2,\delta}(y) = \lambda \frac{\varphi(y)}{\Phi(\delta)} \int_{\mathbb{R}} \varphi(x) \mathbf{1}_{\mathbb{R}^*_+}\left(\delta - \begin{pmatrix} x \\ y \end{pmatrix}.\mathbf{n}\right) F_{1,\delta}(x)^{\lambda-1}\mathrm{d}x.$$

The condition $\delta - x\cos\theta - y\sin\theta \geq 0$ is equivalent to $x \leq (\delta - y\sin\theta)/\cos\theta$, hence Eq. (14) holds. □

We will need in the next sections an expression of the random vector $\tilde{\mathbf{N}}^{\star}_t$ as a function of δ_t and a random vector composed of a *finite* number of i.i.d. random variables. To do so, using notations of Lemma 2, we define the function $\tilde{\mathcal{G}} : \mathbb{R}^*_+ \times ([0,1]\times\mathbb{R}) \to \mathbb{R}^2$ as

$$\tilde{\mathcal{G}}(\delta, \mathbf{w}) = \mathbf{Q}^{-1} \begin{pmatrix} \tilde{F}^{-1}_{\delta}([\mathbf{w}]_1) \\ [\mathbf{w}]_2 \end{pmatrix} . \tag{15}$$

According to Lemma 2, given that $U \sim \mathcal{U}_{[0,1]}$ and $\mathcal{N} \sim \mathcal{N}(0,1)$, $(\tilde{F}^{-1}_{\delta}(U), \mathcal{N})$ (resp. $\tilde{\mathcal{G}}(\delta, (U,\mathcal{N}))$) is distributed as the resampled step $\tilde{\mathbf{N}}^i_t$ in the coordinate system $(\mathbf{n}, \mathbf{n}^{\perp})$ (resp. $(\mathbf{e}_1, \mathbf{e}_2)$). Finally, let $(\mathbf{w}_i)_{i\in[1..\lambda]} \in ([0,1]\times\mathbb{R})^{\lambda}$ and let $\mathcal{G} : \mathbb{R}^*_+ \times ([0,1]\times\mathbb{R})^{\lambda} \to \mathbb{R}^2$ be the function defined as

$$\mathcal{G}(\delta, (\mathbf{w}_i)_{i\in[1..\lambda]}) = \underset{\mathbf{N}\in\left\{\tilde{\mathcal{G}}(\delta,\mathbf{w}_i)|i\in[1..\lambda]\right\}}{\operatorname{argmax}} f(\mathbf{N}) . \tag{16}$$

As shown in the following proposition, given that $\mathbf{W}^i_t \sim (\mathcal{U}_{[0,1]}, \mathcal{N}(0,1))$ and $\mathcal{W}_t = (\mathbf{W}^i_t)_{i\in[1..\lambda]}$, the function $\mathcal{G}(\delta_t, \mathcal{W}_t)$ is distributed as the selected step $\tilde{\mathbf{N}}^{\star}_t$.

Proposition 1. *Let a $(1,\lambda)$-ES with resampling optimize the problem defined in Eq.* (7)*, and let $(\mathbf{W}^i_t)_{i\in[1..\lambda],t\in\mathbb{N}}$ be an i.i.d. sequence of random vectors with $\mathbf{W}^i_t \sim (\mathcal{U}_{[0,1]}, \mathcal{N}(0,1))$, and $\mathcal{W}_t = (\mathbf{W}^i_t)_{i\in[1..\lambda]}$. Then*

$$\tilde{\mathbf{N}}^{\star}_t \overset{d}{=} \mathcal{G}(\delta_t, \mathcal{W}_t) , \tag{17}$$

where the function $\mathcal{G}$ is defined in Eq. (16)*.*

Proof. Since f is a linear function $f(\tilde{\mathbf{Y}}^i_t) = f(\mathbf{X}_t) + \sigma_t f(\tilde{\mathbf{N}}^i_t)$, so $f(\tilde{\mathbf{Y}}^i_t) \leq f(\tilde{\mathbf{Y}}^j_t)$ is equivalent to $f(\tilde{\mathbf{N}}^i_t) \leq f(\tilde{\mathbf{N}}^j_t)$. Hence $\star = \operatorname{argmax}_{i\in[1..\lambda]} f(\tilde{\mathbf{N}}^i_t)$ and therefore $\tilde{\mathbf{N}}^{\star}_t = \operatorname{argmax}_{\mathbf{N}\in\{\tilde{\mathbf{N}}^i_t|i\in[1..\lambda]\}} f(\mathbf{N})$. From Lemma 2 and Eq. (15), $\tilde{\mathbf{N}}^i_t \overset{d}{=} \tilde{\mathcal{G}}(\delta_t, \mathbf{W}^i_t)$, so $\tilde{\mathbf{N}}^{\star}_t \overset{d}{=} \operatorname{argmax}_{\mathbf{N}\in\{\tilde{\mathcal{G}}(\delta_t,\mathbf{W}^i_t)|i\in[1..\lambda]\}} f(\mathbf{N})$, which from (16) is $\mathcal{G}(\delta_t, \mathcal{W}_t)$. □

4 Constant step-size case

We illustrate in this section our methodology on the simple case where the step-size is constantly equal to σ and prove that $(\mathbf{X}_t)_{t\in\mathbb{N}}$ diverges in probability at constant speed and that the progress rate $\varphi^* := \mathbf{E}([\mathbf{X}_{t+1}]_1 - [\mathbf{X}_t]_1) = \sigma\mathbf{E}([\tilde{\mathbf{N}}^{\star}_t]_1)$ (see Arnold 2011a, Eq. 2) converges to a strictly positive constant (Theorem 1). The analysis of the CSA is then a generalization of the results presented here, with more technical results to derive. Note that the progress rate definition coincides with the fitness gain, i.e. $\varphi^* = \mathbf{E}(f(\mathbf{X}_{t+1}) - f(\mathbf{X}_t))$.

As suggested in (Arnold, 2011a), the sequence $(\delta_t)_{t\in\mathbb{N}}$ plays a central role for the analysis, and we will show that it admits a stationary measure. We first prove that this sequence is a homogeneous Markov chain.

Proposition 2. *Consider the* $(1, \lambda)$*-ES with resampling and with constant step-size* σ *optimizing the constrained problem* (7)*. Then the sequence* $\delta_t = g(\mathbf{X}_t)/\sigma$ *is a homogeneous Markov chain on* $\mathbb{R}_+^*$ *and*

$$\delta_{t+1} = \delta_t - \tilde{\mathbf{N}}_t^\star.\mathbf{n} \overset{d}{=} \delta_t - \mathcal{G}(\delta_t, \mathcal{W}_t).\mathbf{n} \ , \tag{18}$$

where $\mathcal{G}$ *is the function defined in* (16) *and* $(\mathcal{W}_t)_{t\in\mathbb{N}} = (\mathbf{W}_t^i)_{i\in[1..\lambda],t\in\mathbb{N}}$ *is an i.i.d. sequence with* $\mathbf{W}_t^i \sim (\mathcal{U}_{[0,1]}, \mathcal{N}(0,1))$ *for all* $(i,t) \in [1..\lambda] \times \mathbb{N}$*.*

Proof. It follows from the definition of δ_t that $\delta_{t+1} = \frac{g(\mathbf{X}_{t+1})}{\sigma_{t+1}} = \frac{-\left(\mathbf{X}_t + \sigma\tilde{\mathbf{N}}_t^\star\right).\mathbf{n}}{\sigma} = \delta_t - \tilde{\mathbf{N}}_t^\star.\mathbf{n}$, and in Proposition 1 we state that $\tilde{\mathbf{N}}_t^\star \overset{d}{=} \mathcal{G}(\delta_t, \mathcal{W}_t)$. Since δ_{t+1} has the same distribution as a time independent function of δ_t and of $\mathcal{W}_t$ where $(\mathcal{W}_t)_{t\in\mathbb{N}}$ are i.i.d., it is a homogeneous Markov chain. □

The Markov Chain $(\delta_t)_{t\in\mathbb{N}}$ comes into play for investigating the divergence of $f(\mathbf{X}_t) = [\mathbf{X}_t]_1$. Indeed, we can express $\frac{[\mathbf{X}_t - \mathbf{X}_0]_1}{t}$ in the following manner:

$$\begin{aligned}\frac{[\mathbf{X}_t - \mathbf{X}_0]_1}{t} &= \frac{1}{t}\sum_{k=0}^{t-1}\left([\mathbf{X}_{k+1}]_1 - [\mathbf{X}_k]_1\right) \\ &= \frac{\sigma}{t}\sum_{k=0}^{t-1}[\tilde{\mathbf{N}}_k^\star]_1 \overset{d}{=} \frac{\sigma}{t}\sum_{k=0}^{t-1}[\mathcal{G}(\delta_k, \mathcal{W}_k)]_1 \ . \end{aligned} \tag{19}$$

The latter term suggests the use of a Law of Large Numbers (LLN) to prove the convergence of $\frac{[\mathbf{X}_t - \mathbf{X}_0]_1}{t}$ which will in turn imply–-if the limit is positive–-the divergence of $[\mathbf{X}_t]_1$ at a constant rate. Sufficient conditions on a Markov chain to be able to apply the LLN include the existence of an invariant probability measure π. The limit term is then expressed as an expectation over the stationary distribution. More precisely, assume the LLN can be applied, the following limit will hold

$$\frac{[\mathbf{X}_t - \mathbf{X}_0]_1}{t} \underset{t\to\infty}{\overset{a.s.}{\to}} \sigma \int_{\mathbb{R}_+^*} \mathbf{E}\left([\mathcal{G}(\delta, \mathcal{W})]_1\right) \pi(d\delta) \ . \tag{20}$$

If the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is also V-ergodic with $|\mathbf{E}([G(\delta, \mathcal{W})]_1)| \leq V(\delta)$ then the progress rate converges to the same limit.

$$\mathbf{E}([\mathbf{X}_{t+1}]_1 - [\mathbf{X}_t]_1) \underset{t\to+\infty}{\longrightarrow} \sigma \int_{\mathbb{R}_+^*} \mathbf{E}\left([\mathcal{G}(\delta, \mathcal{W})]_1\right) \pi(d\delta) \ . \tag{21}$$

We prove formally these two equations in Theorem 1.

The invariant measure π is also underlying the study carried out in (Arnold, 2011a, Section 4) where more precisely it is stated: *"Assuming for now that the mutation strength* σ *is held constant, when the algorithm is iterated, the distribution of* δ*-values tends to a stationary limit distribution."*. We will now provide a formal proof that indeed $(\delta_t)_{t\in\mathbb{N}}$ admits a stationary limit distribution π, as well as prove some other useful properties that will allow us in the end to conclude to the divergence of $([\mathbf{X}_t]_1)_{t\in\mathbb{N}}$.

4.1 Study of the stability of $(\delta_t)_{t\in\mathbb{N}}$

We study in this section the stability of $(\delta_t)_{t\in\mathbb{N}}$. We first derive its transition kernel $P(\delta, A) := \Pr(\delta_{t+1} \in A | \delta_t = \delta)$ for all $\delta \in \mathbb{R}_+^*$ and $A \in \mathcal{B}(\mathbb{R}_+^*)$. Since $\Pr(\delta_{t+1} \in A | \delta_t =$

$\delta) = \Pr(\delta_t - \tilde{\mathbf{N}}_t^\star.\mathbf{n} \in A | \delta_t = \delta)$,

$$P(\delta, A) = \int_{\mathbb{R}^2} \mathbf{1}_A \left(\delta - \mathbf{u}.\mathbf{n}\right) p_\delta^\star \left(\mathbf{u}\right) \mathrm{d}\mathbf{u} \tag{22}$$

where $p_\delta^\star$ is the density of $\tilde{\mathbf{N}}_t^\star$ given in (12). For $t \in \mathbb{N}^*$, the t-steps transition kernel P^t is defined by $P^t(\delta, A) := \Pr(\delta_t \in A | \delta_0 = \delta)$.

From the transition kernel, we will now derive the first properties on the Markov chain $(\delta_t)_{t\in\mathbb{N}}$. First of all we investigate the so-called ψ-irreducible property.

A Markov chain $(\delta_t)_{t\in\mathbb{N}}$ on a state space $\mathbb{R}_+^*$ is *ψ-irreducible* if there exists a non-trivial measure ψ such that for all sets $A \in \mathcal{B}(\mathbb{R}_+^*)$ with $\psi(A) > 0$ and for all $\delta \in \mathbb{R}_+^*$, there exists $t \in \mathbb{N}^*$ such that $P^t(\delta, A) > 0$. We denote $\mathcal{B}^+(\mathbb{R}_+^*)$ the set of Borel sets of $\mathbb{R}_+^*$ with strictly positive ψ-measure.

We also need the notion of *small sets* and *petite sets*. A set $C \in \mathcal{B}(\mathbb{R}_+^*)$ is called a small set if there exists $m \in \mathbb{N}^*$ and a non trivial measure ν_m such that for all sets $A \in \mathcal{B}(\mathbb{R}_+^*)$ and all $\delta \in C$

$$P^m(\delta, A) \geq \nu_m(A) \ . \tag{23}$$

A set $C \in \mathcal{B}(\mathbb{R}_+^*)$ is called a petite set if there exists α a probability measure on $\mathbb{N}$ It seems and a non trivial measure ν_α such that for all sets $A \in \mathcal{B}(\mathbb{R}_+^*)$ and all $\delta \in C$

$$K_\alpha(x, A) := \sum_{m\in\mathbb{N}} P^m(\mathbf{x}, A)\alpha(m) \geq \nu_\alpha(A) \ . \tag{24}$$

A small set is therefore automatically a petite set. If there exists C a ν_1-small set such that $\nu_1(C) > 0$ then the Markov chain is said *strongly aperiodic*.

Proposition 3. *Consider a $(1, \lambda)$-ES with resampling and with constant step-size optimizing the constrained problem* (7) *and let $(\delta_t)_{t\in\mathbb{N}}$ be the Markov chain exhibited in* (18)*. Then $(\delta_t)_{t\in\mathbb{N}}$ is μ_{Leb}-irreducible, strongly aperiodic, and compact sets of $\mathbb{R}_+^*$ and sets of the form $(0, M]$ with $M > 0$ are small sets.*

Proof. Using Eq. (22) and Eq. (12) the transition kernel can be written

$$P(\delta, A) = \lambda \int_{\mathbb{R}^2} \mathbf{1}_A(\delta - \begin{pmatrix} x \\ y \end{pmatrix}.\mathbf{n}) \frac{\varphi(x)\varphi(y)}{\Phi(\delta)} F_{1,\delta}(x)^{\lambda-1} \mathrm{d}y\mathrm{d}x \ .$$

We remove δ from the indicator function by a substitution of variables $u = \delta - x\cos\theta - y\sin\theta$, and $v = x\sin\theta - y\cos\theta$. As this substitution is the composition of a rotation and a translation the determinant of its Jacobian matrix is 1. We denote $h_\delta : (u, v) \mapsto (\delta - u)\cos\theta + v\sin\theta$, $h_\delta^\perp : (u, v) \mapsto (\delta - u)\sin\theta - v\cos\theta$ and $g(\delta, u, v) \mapsto \lambda\varphi(h_\delta(u, v))\varphi(h_\delta^\perp(u, v))/\Phi(\delta)F_{1,\delta}(h_\delta(u, v))^{\lambda-1}$. Then $x = h_\delta(u, v)$, $y = h_\delta^\perp(u, v)$ and

$$P(\delta, A) = \int_{\mathbb{R}} \int_{\mathbb{R}} 1_A(u) g(\delta, u, v) \mathrm{d}v\mathrm{d}u \ . \tag{25}$$

For all δ, u, v the function $g(\delta, u, v)$ is strictly positive hence for all A with $\mu_{Leb}(A) > 0$, $P(\delta, A) > 0$. Hence $(\delta_t)_{t\in\mathbb{N}}$ is irreducible with respect to the Lebesgue measure.

In addition, the function $(\delta, u, v) \mapsto g(\delta, u, v)$ is continuous as the composition of continuous functions (the continuity of $\delta \mapsto F_{1,\delta}(x)$ for all x coming from the dominated convergence theorem). Given a compact C of $\mathbb{R}_+^*$, we hence know that there

exists $g_C > 0$ such that for all $(\delta, u, v) \in C \times [0,1]^2$, $g(\delta, u, v) \geq g_C > 0$. Hence for all $\delta \in C$,

$$P(\delta, A) \geq \underbrace{g_C \mu_{Leb}(A \cap [0,1])}_{:=\nu_C(A)} \ .$$

The measure ν_C being non-trivial, the previous equation shows that compact sets of $\mathbb{R}_+^*$, are small and that for C a compact such that $\mu_{Leb}(C \cap [0,1]) > 0$, we have $\nu_C(C) > 0$ hence the chain is strongly aperiodic. Note also that since $\lim_{\delta \to 0} g(\delta, u, v) > 0$, the same reasoning holds for $(0, M]$ instead of C (where $M > 0$). Hence the set $(0, M]$ is also a small set. □

The application of the LLN for a ψ-irreducible Markov chain $(\delta_t)_{t \in \mathbb{N}}$ on a state space $\mathbb{R}_+^*$ requires the existence of an *invariant measure* π, that is satisfying for all $A \in \mathcal{B}(\mathbb{R}_+^*)$

$$\pi(A) = \int_{\mathbb{R}_+^*} P(\delta, A) \pi(\mathrm{d}\delta) \ . \tag{26}$$

If a Markov chain admits an invariant probability measure then the Markov chain is called positive.

A typical assumption to apply the LLN is positivity and Harris-recurrence. A ψ-irreducible chain $(\delta_t)_{t \in \mathbb{N}}$ on a state space $\mathbb{R}_+^*$ is *Harris-recurrent* if for all sets $A \in \mathcal{B}^+(\mathbb{R}_+^*)$ and for all $\delta \in \mathbb{R}_+^*$, $\Pr(\eta_A = \infty | \delta_0 = \delta) = 1$ where η_A is the occupation time of A, i.e. $\eta_A = \sum_{t=1}^{\infty} 1_A(\delta_t)$. We will show that the Markov chain $(\delta_t)_{t \in \mathbb{N}}$ is positive and Harris-recurrent by using so-called Foster-Lyapunov drift conditions: define the *drift* operator for a positive function V as

$$\Delta V(\delta) = \mathbf{E}[V(\delta_{t+1}) | \delta_t = \delta] - V(\delta) \ .$$

Drift conditions translate that outside a small set, the drift operator is negative. We will show a drift condition for V-geometric ergodicity where given a function $f \geq 1$, a positive and Harris-recurrent chain $(\delta_t)_{t \in \mathbb{N}}$ with invariant measure π is called *f-geometrically ergodic* if $\pi(f) := \int_{\mathbb{R}} f(\delta) \pi(\mathrm{d}\delta) < \infty$ and there exists $r_f > 1$ such that

$$\sum_{t \in \mathbb{N}} r_f^t \| P^t(\delta, \cdot) - \pi \|_f < \infty \ , \forall \delta \in \mathbb{R}_+^* \ , \tag{27}$$

where for ν a signed measure $\|\nu\|_f$ denotes $\sup_{g : |g| \leq f} | \int_{\mathbb{R}_+^*} g(x) \nu(\mathrm{d}x) |$.

To prove the V-geometric ergodicity, we will prove that there exists a small set C, constants $b \in \mathbb{R}$, $\epsilon \in \mathbb{R}_+^*$ and a function $V \geq 1$ finite for at least some $\delta_0 \in \mathbb{R}_+^*$ such that for all $\delta \in \mathbb{R}_+^*$

$$\Delta V(\delta) \leq -\epsilon V(\delta) + b \mathbf{1}_C(\delta) \ . \tag{28}$$

If the Markov chain $(\delta_t)_{t \in \mathbb{N}}$ is ψ-irreducible and aperiodic, this drift condition implies that the chain is V-geometrically ergodic (Meyn and Tweedie, 1993, Theorem 15.0.1)[2] as well as positive and Harris-recurrent[3].

Because sets of the form $(0, M]$ with $M > 0$ are small sets and drift conditions investigate the negativity outside a small set, we need to study the chain for δ large. The following lemma is a technical lemma studying the limit of $\mathbf{E}(\exp(\mathcal{G}(\delta, \mathcal{W}).\mathbf{n}))$ for δ to infinity.

[2] The condition $\pi(V) < \infty$ is given by (Meyn and Tweedie, 1993, Theorem 14.0.1).

[3] The function V of (28) is unbounded off small sets (Meyn and Tweedie, 1993, Lemma 15.2.2) with (Meyn and Tweedie, 1993, Proposition 5.5.7), hence with (Meyn and Tweedie, 1993, Theorem 9.1.8) the Markov chain is Harris-recurrent.

Lemma 4. *Consider the* $(1, \lambda)$*-ES with resampling optimizing the constrained problem* (7)*, and let* $\mathcal{G}$ *be the function defined in* (16)*. We denote* K *and* $\bar{K}$ *the random variables* $\exp(\mathcal{G}(\delta, \mathcal{W}).(a, b))$ *and* $\exp(a|[\mathcal{G}(\delta, \mathcal{W})]_1| + b|[\mathcal{G}(\delta, \mathcal{W})]_2|)$*. For* $\mathcal{W} \sim (\mathcal{U}_{[0,1]}, \mathcal{N}(0, 1))^\lambda$ *and any* $(a, b) \in \mathbb{R}^2$ $\lim_{\delta \to +\infty} \mathbf{E}(K) = \mathbf{E}(\exp(a\mathcal{N}_{\lambda:\lambda}))\mathbf{E}(\exp(b\mathcal{N}(0, 1))) < \infty$ *and* $\lim_{\delta \to +\infty} \mathbf{E}(\bar{K}) < \infty$

For the proof see the appendix. We are now ready to prove a drift condition for geometric ergodicity.

Proposition 4. *Consider a* $(1, \lambda)$*-ES with resampling and with constant step-size optimizing the constrained problem* (7) *and let* $(\delta_t)_{t\in\mathbb{N}}$ *be the Markov chain exhibited in* (18)*. The Markov chain* $(\delta_t)_{t\in\mathbb{N}}$ *is* V*-geometrically ergodic with* $V : \delta \mapsto \exp(\alpha\delta)$ *for* $\alpha > 0$ *small enough, and is Harris-recurrent and positive with invariant probability measure* π*.*

Proof. Take the function $V : \delta \mapsto \exp(\alpha\delta)$, then

$$\begin{aligned}\Delta V(\delta) &= \mathbf{E}\left(\exp\left(\alpha(\delta - \mathcal{G}(\delta, \mathcal{W}).\mathbf{n})\right)\right) - \exp(\alpha\delta) \\ \frac{\Delta V}{V}(\delta) &= \mathbf{E}\left(\exp\left(-\alpha\mathcal{G}(\delta, \mathcal{W}).\mathbf{n}\right)\right) - 1 \ .\end{aligned}$$

With Lemma 4 we obtain that

$$\lim_{\delta \to +\infty} \mathbf{E}\left(\exp\left(-\alpha\mathcal{G}(\delta, \mathcal{W}).\mathbf{n}\right)\right) = \mathbf{E}\left(\exp(-\alpha\mathcal{N}_{\lambda:\lambda}\cos\theta)\right)\mathbf{E}(\exp(-\alpha\mathcal{N}(0, 1)\sin\theta)) < \infty \ .$$

As the right hand side of the previous equation is finite we can invert integral with series with Fubini's theorem, so with Taylor series

$$\lim_{\delta \to +\infty} \mathbf{E}\left(\exp\left(-\alpha\mathcal{G}(\delta, \mathcal{W}).\mathbf{n}\right)\right) = \left(\sum_{i\in\mathbb{N}} \frac{(-\alpha\cos\theta)^i \, \mathbf{E}\left(\mathcal{N}_{\lambda:\lambda}^i\right)}{i!}\right)\left(\sum_{i\in\mathbb{N}} \frac{(-\alpha\sin\theta)^i \, \mathbf{E}\left(\mathcal{N}(0, 1)^i\right)}{i!}\right) \ ,$$

which in turns yields

$$\begin{aligned}\lim_{\delta \to +\infty} \frac{\Delta V}{V}(\delta) &= (1 - \alpha\mathbf{E}(\mathcal{N}_{\lambda:\lambda})\cos\theta + o(\alpha))\,(1 + o(\alpha)) - 1 \\ &= -\alpha\mathbf{E}(\mathcal{N}_{\lambda:\lambda})\cos\theta + o(\alpha) \ .\end{aligned}$$

Since for $\lambda \geq 2$, $\mathbf{E}(\mathcal{N}_{\lambda:\lambda}) > 0$, for $\alpha > 0$ and small enough we get $\lim_{\delta \to +\infty} \frac{\Delta V}{V}(\delta) < -\epsilon < 0$. Hence there exists $\epsilon > 0$, $M > 0$ and $b \in \mathbb{R}$ such that

$$\Delta V(\delta) \leq -\epsilon V(\delta) + b\mathbf{1}_{(0,M]}(\delta) \ .$$

According to Proposition 3, $(0, M]$ is a small set, hence it is petite (Meyn and Tweedie, 1993, Proposition 5.5.3). Furthermore $(\delta_t)_{t\in\mathbb{N}}$ is a ψ-irreducible aperiodic Markov chain so $(\delta_t)_{t\in\mathbb{N}}$ satisfies the conditions of Theorem 15.0.1 from (Meyn and Tweedie, 1993), which with Lemma 15.2.2, Theorem 9.1.8 and Theorem 14.0.1 of (Meyn and Tweedie, 1993) proves the proposition. □

We now proved rigorously the existence (and unicity) of an invariant measure π for the Markov chain $(\delta_t)_{t\in\mathbb{N}}$, which provides the so-called steady state behaviour in (Arnold, 2011a, Section 4). As the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is positive and Harris-recurrent we may now apply a Law of Large Numbers (Meyn and Tweedie, 1993, Theorem 17.1.7) in Eq (19) to obtain the divergence of $f(\mathbf{X}_t)$ and an exact expression of the divergence rate.

Theorem 1. *Consider a $(1,\lambda)$-ES with resampling and with constant step-size optimizing the constrained problem* (7) *and let $(\delta_t)_{t\in\mathbb{N}}$ be the Markov chain exhibited in* (18)*. The sequence $([\mathbf{X}_t]_1)_{t\in\mathbb{N}}$ diverges in probability and expectation to $+\infty$ at constant speed, that is*

$$\frac{[\mathbf{X}_t - \mathbf{X}_0]_1}{t} \xrightarrow[t\to+\infty]{P} \sigma\mathbf{E}_{\pi\otimes\mu_\mathcal{W}}\left([\mathcal{G}\left(\delta,\mathcal{W}\right)]_1\right) > 0 \tag{29}$$

$$\varphi^* = \mathbf{E}\left([\mathbf{X}_{t+1} - \mathbf{X}_t]_1\right) \xrightarrow[t\to+\infty]{} \sigma\mathbf{E}_{\pi\otimes\mu_\mathcal{W}}\left([\mathcal{G}\left(\delta,\mathcal{W}\right)]_1\right) > 0 \ , \tag{30}$$

where φ^* is the progress rate defined in (Arnold, 2011a, Eq. (2)), $\mathcal{G}$ is defined in (16)*, $\mathcal{W} = (\mathbf{W}^i)_{i\in[1..\lambda]}$ with $(\mathbf{W}^i)_{i\in[1..\lambda]}$ an i.i.d. sequence such that $\mathbf{W}^i \sim (\mathcal{U}_{[0,1]}, \mathcal{N}(0,1))$, π is the stationary measure of $(\delta_t)_{t\in\mathbb{N}}$ whose existence is proven in Proposition 4 and $\mu_\mathcal{W}$ is the probability measure of $\mathcal{W}$.*

Proof. From Proposition 4 the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is Harris-recurrent and positive, and since $(\mathcal{W}_t)_{t\in\mathbb{N}}$ is i.i.d., the chain $(\delta_t, \mathcal{W}_t)$ is also Harris-recurrent and positive with invariant probability measure $\pi \times \mu_\mathcal{W}$, so to apply the Law of Large Numbers (Meyn and Tweedie, 1993, Theorem 17.0.1) to $[\mathcal{G}]_1$ we only need $[\mathcal{G}]_1$ to be $\pi\otimes\mu_\mathcal{W}$-integrable.

With Fubini-Tonelli's theorem $\mathbf{E}_{\pi\otimes\mu_\mathcal{W}}(|[\mathcal{G}(\delta,\mathcal{W})]_1|)$ equals to $\mathbf{E}_\pi(\mathbf{E}_{\mu_\mathcal{W}}(|[\mathcal{G}(\delta,\mathcal{W})]_1|))$. As $\delta \geq 0$, we have $\Phi(\delta) \geq \Phi(0) = 1/2$, and for all $x \in \mathbb{R}$ as $\Phi(x) \leq 1$, $F_{1,\delta}(x) \leq 1$ and $\varphi(x) \leq \exp(-x^2/2)$ with Eq. (13) we obtain that $|x|p^\star_{1,\delta}(x) \leq 2\lambda|x|\exp(-x^2/2)$ so the function $x \mapsto |x|p^\star_{1,\delta}(x)$ is integrable. Hence for all $\delta \in \mathbb{R}_+$, $\mathbf{E}_{\mu_\mathcal{W}}(|[\mathcal{G}(\delta,\mathcal{W})]_1|)$ is finite. Using the dominated convergence theorem, the function $\delta \mapsto F_{1,\delta}(x)$ is continuous, hence so is $\delta \mapsto p^\star_{1,\delta}(x)$. From (13) $|x|p^\star_{1,\delta}(x) \leq 2\lambda|x|\varphi(x)$, which is integrable, so the dominated convergence theorem implies that the function $\delta \mapsto \mathbf{E}_{\mu_\mathcal{W}}(|[\mathcal{G}(\delta,\mathcal{W}]_1|)$ is continuous. Finally, using Lemma 4 with Jensen's inequality shows that $\lim_{\delta\to+\infty}\mathbf{E}_{\mu_\mathcal{W}}(|[\mathcal{G}(\delta,\mathcal{W})]_1|)$ is finite. Therefore the function $\delta \mapsto \mathbf{E}_{\mu_\mathcal{W}}(|[\mathcal{G}(\delta,\mathcal{W}]_1|)$ is bounded by a constant $M \in \mathbb{R}_+$. As π is a probability measure $\mathbf{E}_\pi(\mathbf{E}_{\mu_\mathcal{W}}(|[\mathcal{G}(\delta,\mathcal{W})]_1|)) \leq M < \infty$, meaning $[\mathcal{G}]_1$ is $\pi\otimes\mu_\mathcal{W}$-integrable. Hence we may apply the LLN on Eq. (19)

$$\frac{\sigma}{t}\sum_{k=0}^{t-1}[\mathcal{G}(\delta_k,\mathcal{W}_k)]_1 \xrightarrow[t\to+\infty]{a.s.} \sigma\mathbf{E}_{\pi\otimes\mu_\mathcal{W}}\left([\mathcal{G}(\delta,\mathcal{W})]_1\right) < \infty \ .$$

The equality in distribution in (19) allows us to deduce the convergence in probability of the left hand side of (19) to the right hand side of the previous equation.

From (19) $[\mathbf{X}_{t+1} - \mathbf{X}_t]_1 \overset{d}{=} \sigma\mathcal{G}(\delta_t,\mathcal{W}_t)$ so $\mathbf{E}([\mathbf{X}_{t+1} - \mathbf{X}_t]_1|\mathbf{X}_0 = \mathbf{x}) = \sigma\mathbf{E}(\mathcal{G}(\delta_t,\mathcal{W}_t)|\delta_0 = \mathbf{x}/\sigma)$. As $\mathcal{G}$ is integrable with Fubini's theorem $\mathbf{E}(\mathcal{G}(\delta_t,\mathcal{W}_t)|\delta_0 = \mathbf{x}/\sigma) = \int_{\mathbb{R}_+^*}\mathbf{E}_{\mu_\mathcal{W}}(\mathcal{G}(\mathbf{y},\mathcal{W}))P^t(\mathbf{x}/\sigma,d\mathbf{y})$, so $\mathbf{E}(\mathcal{G}(\delta_t,\mathcal{W}_t)|\delta_0 = \mathbf{x}/\sigma) - \mathbf{E}_{\pi\times\mu_\mathcal{W}}(\mathcal{G}(\delta,\mathcal{W})) = \int_{\mathbb{R}_+^*}\mathbf{E}_{\mu_\mathcal{W}}(\mathcal{G}(\mathbf{y},\mathcal{W}))(P^t(\mathbf{x}/\sigma,d\mathbf{y}) - \pi(d\mathbf{y}))$. According to Proposition 4 $(\delta_t)_{t\in\mathbb{N}}$ is V-geometrically ergodic with $V : \delta \mapsto \exp(\alpha\delta)$, so there exists M_δ and $r > 1$ such that $\|P^t(\delta,\cdot) - \pi\|_V \leq M_\delta r^{-t}$. We showed that the function $\delta \mapsto \mathbf{E}(|[\mathcal{G}(\delta,\mathcal{W})]_1|)$ is bounded, so since $V(\delta) \geq 1$ for all $\delta \in \mathbb{R}_+^*$ and $\lim_{\delta\to+\infty}V(\delta) = +\infty$, there exists k such that $\mathbf{E}_{\mu_\mathcal{W}}(|[\mathcal{G}(\delta,\mathcal{W})]_1|) \leq kV(\delta)$ for all δ. Hence $|\int\mathbf{E}_{\mu_\mathcal{W}}(|[\mathcal{G}(x,\mathcal{W})]_1|)(P^t(\delta,dx) - \pi(dx))| \leq k\|P^t(\delta,\cdot) - \pi\|_V \leq kM_\delta r^{-t}$. And therefore $|\mathbf{E}(\mathcal{G}(\delta_t,\mathcal{W}_t)|\delta_0 = \mathbf{x}/\sigma) - \mathbf{E}_{\pi\times\mu_\mathcal{W}}(\mathcal{G}(\delta,\mathcal{W}))| \leq kM_\delta r^{-t}$ which converges to 0 when t goes to infinity.

As the measure π is an invariant measure for the Markov chain $(\delta_t)_{t\in\mathbb{N}}$, using (18), $\mathbf{E}_{\pi\otimes\mu_\mathcal{W}}(\delta) = \mathbf{E}_{\pi\otimes\mu_\mathcal{W}}(\delta - \mathcal{G}(\delta,\mathcal{W}).\mathbf{n})$, hence $\mathbf{E}_{\pi\otimes\mu_\mathcal{W}}(\mathcal{G}(\delta,\mathcal{W}).\mathbf{n}) = 0$ and thus

$$\mathbf{E}_{\pi\otimes\mu_\mathcal{W}}\left([\mathcal{G}(\delta,\mathcal{W})]_1\right) = -\tan\theta\mathbf{E}_{\pi\otimes\mu_\mathcal{W}}\left([\mathcal{G}(\delta,\mathcal{W})]_2\right) \ .$$

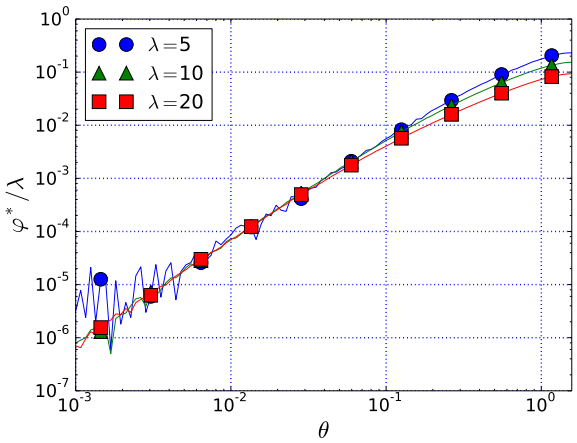


Figure 2: Normalized progress rate $\varphi^\star = \mathbf{E}(f(\mathbf{X}_{t+1}) - f(\mathbf{X}_t))$ divided by λ for the $(1, \lambda)$-ES with constant step-size $\sigma = 1$ and resampling, plotted against the constraint angle θ, for $\lambda \in \{5, 10, 20\}$.

We see from Eq. (14) that for $y > 0$, $p^\star_{2,\delta}(y) < p^\star_{2,\delta}(-y)$ hence the expected value $\mathbf{E}_{\pi\otimes\mu_\mathcal{W}}([\mathcal{G}(\delta, \mathcal{W})]_2)$ is strictly negative. With the previous equation it implies that $\mathbf{E}_{\pi\otimes\mu_\mathcal{W}}([\mathcal{G}(\delta, \mathcal{W})]_1)$ is strictly positive.

□

We showed rigorously the divergence of $[\mathbf{X}_t]_1$ and gave an exact expression of the divergence rate, and that the progress rate φ^* converges to the same rate. The fact that the chain $(\delta_t)_{t\in\mathbb{N}}$ is V-geometrically ergodic gives that there exists a constant $r > 1$ such that $\sum_t r^t \|P^t(\delta, \cdot) - \pi\|_V < \infty$. This implies that the distribution π can be simulated efficiently by a Monte Carlo simulation allowing to have precise estimations of the divergence rate of $[\mathbf{X}_t]_1$.

A Monte Carlo simulation of the divergence rate in the right hand side of (29) and (30) and for 10^6 time steps gives the progress rate of (Arnold, 2011a) $\varphi^\star = \mathbf{E}([\mathbf{X}_{t+1} - \mathbf{X}_t]_1)$, which once normalized by σ and λ yields Fig. 2. We normalize per λ as in evolution strategies the cost of the algorithm is assumed to be the number of f-calls. We see that for small values of θ, the normalized serial progress rate assumes roughly $\varphi^\star/\lambda \approx \theta^2$. Only for larger constraint angles the serial progress rate depends on λ where smaller λ are preferable.

Fig. 3 is obtained through simulations of the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ defined in Eq. (18) for 10^6 time steps where the values of $(\delta_t)_{t\in\mathbb{N}}$ are averaged over time. We see that when $\theta \to \pi/2$ then $\mathbf{E}_\pi(\delta) \to +\infty$ since the selection does not attract $\mathbf{X}_t$ towards the constraint anymore. With a larger population size the algorithm is closer to the constraint, as better samples are more likely to be found close to the constraint.

5 Cumulative Step size Adaptation

In this section we apply the techniques introduced in the previous section to the case where the step-size is adapted using Cumulative Step-size Adaptation, CSA (Hansen and Ostermeier, 2001). This technique was studied on sphere functions (Arnold and Beyer, 2004) and on ridge functions (Arnold and MacLeod, 2008).

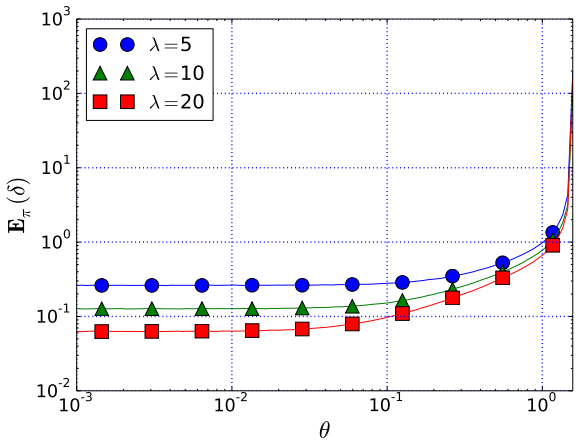


Figure 3: Average normalized distance δ from the constraint for the $(1, \lambda)$-ES with constant step-size and resampling plotted against the constraint angle θ for $\lambda \in \{5, 10, 20\}$.

In CSA, the step-size is adapted using a path $\mathbf{p}_t$, vector of $\mathbb{R}^n$, that sums up the different selected steps $\tilde{\mathbf{N}}_t^\star$ with a discount factor. More precisely the evolution path $\mathbf{p}_t \in \mathbb{R}^n$ is defined by $\mathbf{p}_0 \sim \mathcal{N}(\mathbf{0}, \mathrm{Id}_n)$ and

$$\mathbf{p}_{t+1} = (1-c)\mathbf{p}_t + \sqrt{c(2-c)}\tilde{\mathbf{N}}_t^\star \ . \tag{31}$$

The variable $c \in (0, 1]$ is called the cumulation parameter, and determines the "memory" of the evolution path, with the importance of a step $\tilde{\mathbf{N}}_0^\star$ decreasing in $(1-c)^t$. The backward time horizon is consequently about $1/c$. The coefficients in Eq (31) are chosen such that if $\mathbf{p}_t$ follows a standard normal distribution, and if f ranks uniformly randomly the different samples $(\tilde{\mathbf{N}}_t^i)_{i\in[[1..\lambda]]}$ and that these samples are normally distributed, then $\mathbf{p}_{t+1}$ will also follow a standard normal distribution independently of the value of c.

The length of the evolution path is compared to the expected length of a Gaussian vector (that corresponds to the expected length under random selection) (see (Hansen and Ostermeier, 2001)). To simplify the analysis we study here a modified version of CSA introduced in (Arnold, 2002) where the squared length of the evolution path is compared with the expected squared length of a Gaussian vector, that is n, since it would be the distribution of the evolution path under random selection. If $\|\mathbf{p}_t\|^2$ is greater (respectively lower) than n, then the step-size is increased (respectively decreased) following

$$\sigma_{t+1} = \sigma_t \exp\left(\frac{c}{2d_\sigma}\left(\frac{\|\mathbf{p}_{t+1}\|^2}{n} - 1\right)\right) \ , \tag{32}$$

where the damping parameter d_σ determines how much the step-size can change and can be set here to $d_\sigma = 1$.

As $[\tilde{\mathbf{N}}_t^\star]_i \sim \mathcal{N}(0,1)$ for $i \geq 3$, we also have $[\mathbf{p}_t]_i \sim \mathcal{N}(0,1)$. It is convenient in the sequel to also denote by $\mathbf{p}_t$ the two dimensional vector $([\mathbf{p}_t]_1, [\mathbf{p}_t]_2)$. With this (small) abuse of notations, (32) is rewritten as

$$\sigma_{t+1} = \sigma_t \exp\left(\frac{c}{2d_\sigma}\left(\frac{\|\mathbf{p}_{t+1}\|^2 + K_t}{n} - 1\right)\right) \ , \tag{33}$$

with $(K_t)_{t\in\mathbb{N}}$ an i.i.d. sequence of random variables following a chi-squared distribution with $n-2$ degrees of freedom. We shall denote $\eta_c^\star$ the multiplicative step-size change σ_{t+1}/σ_t, that is the function

$$\eta_c^\star(\mathbf{p}_t, \delta_t, \mathcal{W}_t, K_t) = \exp\left(\frac{c}{2d_\sigma}\left(\frac{\|(1-c)\mathbf{p}_t + \sqrt{c(2-c)}\mathcal{G}(\delta_t, \mathcal{W}_t)\|^2 + K_t}{n} - 1\right)\right) . \quad (34)$$

Note that for $c = 1$, $\eta_1^\star$ is a function of only δ_t, $\mathcal{W}_t$ and K_t that we will hence denote $\eta_1^\star(\delta_t, \mathcal{W}_t, K_t)$.

We prove in the next proposition that for $c < 1$ the sequence $(\delta_t, \mathbf{p}_t)_{t\in\mathbb{N}}$ is an homogeneous Markov chain and explicit its update function. In the case where $c = 1$ the chain reduces to δ_t.

Proposition 5. *Consider a $(1, \lambda)$-ES with resampling and cumulative step-size adaptation maximizing the constrained problem* (7)*. Take $\delta_t = g(\mathbf{X}_t)/\sigma_t$. The sequence $(\delta_t, \mathbf{p}_t)_{t\in\mathbb{N}}$ is a time-homogeneous Markov chain and*

$$\delta_{t+1} \overset{d}{=} \frac{\delta_t - \mathcal{G}(\delta_t, \mathcal{W}_t).\mathbf{n}}{\eta_c^\star(\mathbf{p}_t, \delta_t, \mathcal{W}_t, K_t)} , \quad (35)$$

$$\mathbf{p}_{t+1} \overset{d}{=} (1-c)\mathbf{p}_t + \sqrt{c(2-c)}\mathcal{G}(\delta_t, \mathcal{W}_t) , \quad (36)$$

with $(K_t)_{t\in\mathbb{N}}$ a i.i.d. sequence of random variables following a chi squared distribution with $n-2$ degrees of freedom, $\mathcal{G}$ defined in Eq. (16) *and $\mathcal{W}_t$ defined in Proposition 1.*

If $c = 1$ then the sequence $(\delta_t)_{t\in\mathbb{N}}$ is a time-homogeneous Markov chain and

$$\delta_{t+1} \overset{d}{=} \frac{\delta_t - \mathcal{G}(\delta_t, \mathcal{W}_t).\mathbf{n}}{\exp\left(\frac{c}{2d_\sigma}\left(\frac{\|\mathcal{G}(\delta_t, \mathcal{W}_t)\|^2}{n} - 1\right)\right)} \quad (37)$$

Proof. With Eq. (31) and Eq. (17) we get Eq. (36).

From Eq. (8) and Proposition 1 it follows that

$$\begin{aligned}\delta_{t+1} = -\frac{\mathbf{X}_{t+1}.\mathbf{n}}{\sigma_{t+1}} &\overset{d}{=} -\frac{\mathbf{X}_t.\mathbf{n} + \sigma_t \tilde{\mathbf{N}}_t^\star.\mathbf{n}}{\sigma_t \eta_c^\star(\mathbf{p}_t, \delta_t, \mathcal{W}_t, K_t)} \\ &\overset{d}{=} \frac{\delta_t - \mathcal{G}(\delta_t, \mathcal{W}_t).\mathbf{n}}{\eta_c^\star(\mathbf{p}_t, \delta_t, \mathcal{W}_t, K_t)} .\end{aligned}$$

So $(\delta_{t+1}, \mathbf{p}_{t+1})$ is a function of only $(\delta_t, \mathbf{p}_t)$ and i.i.d. random variables, hence $(\delta_t, \mathbf{p}_t)_{t\in\mathbb{N}}$ is a time-homogeneous Markov chain.

Fixing $c = 1$ in (35) and (36) immediately yields (37), and then δ_{t+1} is a function of only δ_t and i.i.d. random variables, so in this case $(\delta_t)_{t\in\mathbb{N}}$ is a time-homogeneous Markov chain. □

As for the constant step-size case, the Markov chain is important when investigating the convergence or divergence of the step size of the algorithm. Indeed from Eq. (33) we can express $\ln(\sigma_t/\sigma_0)/t$ as

$$\frac{1}{t}\ln\frac{\sigma_t}{\sigma_0} = \frac{c}{2d_\sigma}\left(\frac{\frac{1}{t}\left(\sum_{i=0}^{t-1}\|\mathbf{p}_{i+1}\|^2 + K_i\right)}{n} - 1\right) \quad (38)$$

The right hand side suggests to use the LLN. The convergence of $\ln(\sigma_t/\sigma_0)/t$ to a strictly positive limit (resp. negative) will imply the divergence (resp. convergence) of σ_t at a geometrical rate.

It turns out that the dynamic of the chain $(\delta_t, \mathbf{p}_t)_{t\in\mathbb{N}}$ looks complex to analyze. Establishing drift conditions looks particularly challenging. We therefore restrict the rest of the study to the more simple case where $c = 1$, hence the Markov chain of interest is $(\delta_t)_{t\in\mathbb{N}}$. Then (38) becomes

$$\frac{1}{t}\ln\frac{\sigma_t}{\sigma_0} \overset{d}{=} \frac{c}{2d_\sigma}\left(\frac{\frac{1}{t}\sum_{i=0}^{t-1}\|\mathcal{G}(\delta_i,\mathcal{W}_i)\|^2 + K_i}{n} - 1\right) \quad . \tag{39}$$

To apply the LLN we will need the Markov chain to be Harris positive, and the properties mentioned in the following lemma.

Lemma 5 (Chotard and Auger 201, Proposition 7)**.** *Consider a $(1,\lambda)$-ES with resampling and cumulative step-size adaptation maximizing the constrained problem* (7)*. For $c = 1$ the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ from Proposition 5 is ψ-irreducible, strongly aperiodic, and compact sets of $\mathbb{R}_+^*$ are small sets for this chain.*

We believe that the latter result can be generalized to the case $c < 1$ if for any $(\delta_0, \mathbf{p}_0) \in \mathbb{R}_+^* \times \mathbb{R}^n$ there exists $t_{\delta_0,\mathbf{p}_0}$ such that for all $t \geq t_{\delta_0,\mathbf{p}_0}$ there exists a path of events of length t from $(\delta_0, \mathbf{p}_0)$ to any point of the set $[0, M] \times B(\mathbf{0}, r)$.

To show the Harris positivity of $(\delta_t)_{t\in\mathbb{N}}$ we first need to study the behaviour of the drift operator we want to use when $\delta \to +\infty$, that is far from the constraint. Then, intuitively, as $[\tilde{\mathbf{N}}_t^\star]_2$ would not be influenced by the resampling anymore, it would be distributed as a random normal variable, and $[\tilde{\mathbf{N}}_t^\star]_1$ would be distributed as the last order statistic of λ normal random variables. This is used in the following technical lemma.

Lemma 6. *For $\alpha > 0$ small enough*

$$\frac{1}{\delta^\alpha + \delta^{-\alpha}}\mathbf{E}\left(\frac{(\delta - \mathcal{G}(\delta,\mathcal{W}).\mathbf{n})^\alpha}{\eta_1^\star(\delta,\mathcal{W},K)^\alpha}\right) \underset{\delta\to+\infty}{\longrightarrow} E_1E_2E_3 < \infty \tag{40}$$

$$\frac{1}{\delta^\alpha + \delta^{-\alpha}}\mathbf{E}\left(\frac{(\delta - \mathcal{G}(\delta,\mathcal{W}).\mathbf{n})^\alpha}{\eta_1^\star(\delta,\mathcal{W},K)^\alpha}\right) \underset{\delta\to0}{\longrightarrow} 0 \tag{41}$$

$$\frac{1}{\delta^\alpha + \delta^{-\alpha}}\mathbf{E}\left(\frac{\eta_1^\star(\delta,\mathcal{W},K)^\alpha}{(\delta - \mathcal{G}(\delta,\mathcal{W}).\mathbf{n})^\alpha}\right) \underset{\delta\to+\infty}{\longrightarrow} 0 \tag{42}$$

$$\frac{1}{\delta^\alpha + \delta^{-\alpha}}\mathbf{E}\left(\frac{\eta_1^\star(\delta,\mathcal{W},K)^\alpha}{(\delta - \mathcal{G}(\delta,\mathcal{W}).\mathbf{n})^\alpha}\right) \underset{\delta\to0}{\longrightarrow} 0 \quad , \tag{43}$$

where $E_1 = \mathbf{E}(\exp(-\frac{\alpha}{2d_\sigma n}(\mathcal{N}_{\lambda:\lambda}^2 - 1)))$, $E_2 = \mathbf{E}(\exp(-\frac{\alpha}{2d_\sigma n}(\mathcal{N}(0,1)^2 - 1)))$, and $E_3 = \mathbf{E}(\exp(-\frac{\alpha}{2d_\sigma n}(K - (n-2))))$; where $\mathcal{G}$ is the function defined in Eq. (16) *and $\eta_1^\star$ is defined in Eq.* (34) *(for $c = 1$), K is a random variable following a chi-squared distribution with $n-2$ degrees of freedom and $\mathcal{W} \sim (\mathcal{U}_{[0,1]}, \mathcal{N}(0,1))^\lambda$ is a random vector.*

The proof of this lemma consists in applications of Lebesgue's dominated convergence theorem, and can be found in the appendix.

We now prove the Harris positivity of $(\delta_t)_{t\in\mathbb{N}}$ by proving a stronger property, namely the geometric ergodicity that we show using the drift inequality (28).

Proposition 6. *Consider a $(1,\lambda)$-ES with resampling and cumulative step-size adaptation maximizing the constrained problem* (7)*. For $c = 1$ the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ from Proposition 5 is V-geometrically ergodic with $V : \delta \in \mathbb{R}_+^* \mapsto \delta^\alpha + \delta^{-\alpha}$ for $\alpha > 0$ small enough, and positive Harris with invariant measure π_1.*

Proof. Take V the positive function $V(\delta) = \delta^\alpha + \delta^{-\alpha}$ (the parameter α is strictly positive and will be specified later), $\mathcal{W} \sim (\mathcal{U}_{[0,1]}, \mathcal{N}(0,1))^\lambda$ a random vector and K a random variable following a chi squared distribution with $n-2$ degrees of freedom. We first study $\Delta V/V(\delta)$ when $\delta \to +\infty$. From Eq. (37) we then have the following drift quotient

$$\frac{\Delta V(\delta)}{V(\delta)} = \frac{1}{V(\delta)}\mathbf{E}\left(\frac{(\delta - \mathcal{G}(\delta,\mathcal{W}).\mathbf{n})^\alpha}{\eta_1^\star(\delta,\mathcal{W},K)^\alpha}\right) + \frac{1}{V(\delta)}\mathbf{E}\left(\frac{\eta_1^\star(\delta,\mathcal{W},K)^\alpha}{(\delta - \mathcal{G}(\delta,\mathcal{W}).\mathbf{n})^\alpha}\right) - 1 \ , \qquad (44)$$

with $\eta_1^\star$ defined in Eq. (34) and $\mathcal{G}$ in Eq. (16). From Lemma 6, following the same notations than in the lemma, when $\delta \to +\infty$ and if $\alpha > 0$ is small enough, the right hand side of the previous equation converges to $E_1 E_2 E_3 - 1$. With Taylor series

$$E_1 = \mathbf{E}\left(\sum_{k\in\mathbb{N}} \frac{\left(-\frac{\alpha}{2d_\sigma n}\left(\mathcal{N}_{\lambda:\lambda}^2 - 1\right)\right)^k}{k!}\right) \ .$$

Furthermore, as the density of $\mathcal{N}_{\lambda:\lambda}$ at x equals to $\lambda\varphi(x)\Phi(x)^{\lambda-1}$ and that $\exp|\alpha/(2d_\sigma n)(x^2-1)|\lambda\varphi(x)\Phi(x)^{\lambda-1} \leq \lambda\exp(\alpha/(2d_\sigma n)x^2 - x^2/2)$ which for α small enough is integrable,

$$\mathbf{E}\left(\sum_{k\in\mathbb{N}} \frac{\left|-\frac{\alpha}{2d_\sigma n}\left(\mathcal{N}_{\lambda:\lambda}^2 - 1\right)\right|^k}{k!}\right) = \int_{\mathbb{R}} \exp\left|\frac{\alpha}{2d_\sigma n}\left(x^2-1\right)\right| \lambda\varphi(x)\Phi(x)^{\lambda-1}\mathrm{d}x < \infty \ .$$

Hence we can use Fubini's theorem to invert series (which are integrals for the counting measure) and integral. The same reasoning holding for E_2 and E_3 (for E_3 with the chi-squared distribution we need $\alpha/(2d_\sigma n)x - x/2 < 0$) we have

$$\lim_{\delta\to+\infty}\frac{\Delta V}{V}(\delta) = \left(1 - \frac{\alpha}{2d_\sigma n}\mathbf{E}(\mathcal{N}_{\lambda:\lambda}^2 - 1) + o(\alpha)\right)\left(1 - \frac{\alpha}{2d_\sigma n}\mathbf{E}(\mathcal{N}(0,1)^2 - 1) + o(\alpha)\right)$$
$$\left(1 - \frac{\alpha}{2d_\sigma n}\mathbf{E}(\chi_{n-2}^2 - (n-2)) + o(\alpha)\right) - 1 \ ,$$

and as $\mathbf{E}(\mathcal{N}(0,1)^2) = 1$ and $\mathbf{E}(\chi_{n-2}^2) = n-2$

$$\lim_{\delta\to+\infty}\frac{\Delta V}{V}(\delta) = -\frac{\alpha}{2d_\sigma n}\mathbf{E}\left(\mathcal{N}_{\lambda:\lambda}^2 - 1\right) + o(\alpha) \ .$$

From (Chotard et al., 2012a) if $\lambda > 2$ then $\mathbf{E}(\mathcal{N}_{\lambda:\lambda}^2) > 1$. Therefore, for α small enough, we have $\lim_{\delta\to+\infty}\frac{\Delta V}{V}(\delta) < 0$ so there exists $\epsilon_1 > 0$ and $M > 0$ such that $\Delta V(\delta) \leq -\epsilon_1 V(\delta)$ whenever $\delta > M$.

Similarly, when α is small enough, using Lemma 6, $\lim_{\delta\to 0}\mathbf{E}((\delta - \mathcal{G}(\delta,\mathcal{W}))^\alpha/\eta_1^\star(\delta,\mathcal{W},K)^\alpha)/V(\delta) = 0$ and $\lim_{\delta\to 0}\mathbf{E}(\eta_1^\star(\delta,\mathcal{W},K)^\alpha/(\delta - \mathcal{G}(\delta,\mathcal{W}))^\alpha)/V(\delta) = 0$. Hence using (44), $\lim_{\delta\to 0}\Delta V(\delta)/V(\delta) = -1$. So there exists ϵ_2 and $m > 0$ such that $\Delta V(\delta) \leq -\epsilon_2 V(\delta)$ for all $\delta \in (0, m)$. And since $\Delta V(\delta)$ and $V(\delta)$ are bounded functions on compacts of $\mathbb{R}_+^*$, there exists $b \in R$ such that

$$\Delta V(\delta) \leq -\min(\epsilon_1, \epsilon_2)V(\delta) + b\mathbf{1}_{[m,M]}(\delta) \ .$$

With Lemma 5, $[m, M]$ is a small set, and $(\delta_t)_{t\in\mathbb{N}}$ is a ψ-irreducible aperiodic Markov chain. So $(\delta_t)_{t\in\mathbb{N}}$ satisfies the assumptions of (Meyn and Tweedie, 1993, Theorem 15.0.1), which proves the proposition.

□

The same results for $c < 1$ are difficult to obtain, as then both δ_t and $\mathbf{p}_t$ must be controlled together. For $\mathbf{p}_t = 0$ and $\delta_t \geq M$, $\|\mathbf{p}_{t+1}\|$ and δ_{t+1} will in average increase, so either we need that $[M, +\infty) \times B(\mathbf{0}, r)$ is a small set (although it is not compact), or we need to look τ steps in the future with τ large enough to see $\delta_{t+\tau}$ decrease for all possible values of $\mathbf{p}_t$ outside of a small set.

Note that although in Proposition 4 and Proposition 6 we show the existence of a stationary measure for $(\delta_t)_{t\in\mathbb{N}}$, these are not the same measures, and not the same Markov chains as they have different update rules (compare Eq. (18) and Eq. (35)) The chain $(\delta_t)_{t\in\mathbb{N}}$ being Harris positive we may now apply a LLN to Eq. (39) to get an exact expression of the divergence/convergence rate of the step-size.

Theorem 2. *Consider a $(1, \lambda)$-ES with resampling and cumulative step-size adaptation maximizing the constrained problem* (7)*, and for $c = 1$ take $(\delta_t)_{t\in\mathbb{N}}$ the Markov chain from Proposition 5. Then the step-size diverges or converges geometrically in probability*

$$\frac{1}{t}\ln\left(\frac{\sigma_t}{\sigma_0}\right) \xrightarrow[t\to\infty]{P} \frac{1}{2d_\sigma n}\left(\mathbf{E}_{\pi_1\otimes\mu_\mathcal{W}}\left(\|\mathcal{G}\left(\delta, \mathcal{W}\right)\|^2\right) - 2\right) \quad , \tag{45}$$

and in expectation

$$\mathbf{E}\left(\ln\left(\frac{\sigma_{t+1}}{\sigma_t}\right)\right) \xrightarrow[t\to+\infty]{} \frac{1}{2d_\sigma n}\left(\mathbf{E}_{\pi_1\otimes\mu_\mathcal{W}}\left(\|\mathcal{G}\left(\delta, \mathcal{W}\right)\|^2\right) - 2\right) \tag{46}$$

with $\mathcal{G}$ defined in (16) *and $\mathcal{W} = (\mathbf{W}^i)_{i\in[1..\lambda]}$ where $(\mathbf{W}^i)_{i\in[1..\lambda]}$ is an i.i.d. sequence such that $\mathbf{W}^i \sim (\mathcal{U}_{[0,1]}, \mathcal{N}(0,1))$, $\mu_\mathcal{W}$ is the probability measure of $\mathcal{W}$ and π_1 is the invariant measure of $(\delta_t)_{t\in\mathbb{N}}$ whose existence is proved in Proposition 6.*

Furthermore, the change in fitness value $f(\mathbf{X}_{t+1}) - f(\mathbf{X}_t)$ diverges or converges geometrically in probability

$$\frac{1}{t}\ln\left|\frac{f(\mathbf{X}_{t+1}) - f(\mathbf{X}_t)}{\sigma_0}\right| \xrightarrow[t\to\infty]{P} \frac{1}{2d_\sigma n}\left(\mathbf{E}_{\pi_1\otimes\mu_\mathcal{W}}\left(\|\mathcal{G}\left(\delta, \mathcal{W}\right)\|^2\right) - 2\right) \quad . \tag{47}$$

Proof. From Proposition 6 the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is Harris positive, and since $(\mathcal{W}_t)_{t\in\mathbb{N}}$ is i.i.d., the chain $(\delta_t, \mathcal{W}_t)_{t\in\mathbb{N}}$ is also Harris positive with invariant probability measure $\pi_1 \times \mu_\mathcal{W}$, so to apply the Law of Large Numbers of (Meyn and Tweedie, 1993, Theorem 17.0.1) to Eq. (38) we only need the function $(\delta, \mathbf{w}) \mapsto \|G(\delta, \mathbf{w})\|^2 + K$ to be $\pi_1 \times \mu_\mathcal{W}$-integrable.

Since K has chi-squared distribution with $n - 2$ degrees of freedom, $\mathbf{E}_{\pi_1\times\mu_\mathcal{W}}(\|G(\delta, \mathcal{W})\|^2 + K)$ equals to $\mathbf{E}_{\pi_1\times\mu_\mathcal{W}}(\|G(\delta, \mathcal{W})\|^2) + n - 2$. With Fubini-Tonelli's theorem, $\mathbf{E}_{\pi_1\times\mu_\mathcal{W}}(\|\mathcal{G}(\delta, \mathcal{W})\|^2)$ is equal to $\mathbf{E}_{\pi_1}(\mathbf{E}_{\mu_\mathcal{W}}(\|\mathcal{G}(\delta, \mathcal{W})\|^2))$. From Eq. (12) and from the proof of Lemma 4 the function $\mathbf{x} \mapsto \|\mathbf{x}\|^2 p^\star_\delta(\mathbf{x})$ converges simply to $\|\mathbf{x}\|^2 p_{\mathcal{N}_{\lambda:\lambda}}([\mathbf{x}]_1)\varphi([x]_2)$ while being dominated by $\lambda/\Phi(0)\exp(-\|\mathbf{x}\|^2)$ which is integrable. Hence we may apply Lebesgue's dominated convergence theorem showing that the function $\delta \mapsto \mathbf{E}_{\mu_\mathcal{W}}(\|\mathcal{G}(\delta, \mathcal{W})\|^2)$ is continuous and has a finite limit and is therefore bounded by a constant $M_{\mathcal{G}^2}$. As the measure π_1 is a probability measure (so $\pi_1(\mathbb{R}) = 1$), $\mathbf{E}_{\pi_1}(\mathbf{E}_{\mu_\mathcal{W}}(\|\mathcal{G}(\delta, \mathcal{W})\|^2 | \delta_t = \delta)) \leq M_{\mathcal{G}^2} < \infty$. Hence we may apply the Law of Large Numbers

$$\sum_{i=0}^{t-1} \frac{\|\mathcal{G}(\delta_i, \mathcal{W}_i)\|^2 + K_i}{t} \xrightarrow[t\to\infty]{a.s} \mathbf{E}_{\pi_1\times\mu_\mathcal{W}}\left(\|\mathcal{G}(\delta, \mathcal{W})\|^2\right) + n - 2 \ .$$

Combining this equation with Eq. (39) yields Eq. (45).

From Proposition 1, (31) for $c = 1$ and (33), $\ln(\sigma_{t+1}/\sigma_t) \overset{d}{=} 1/(2d_\sigma n)(\|\mathcal{G}(\delta_t, \mathcal{W}_t)\|^2 + \chi^2_{n-2} - n)$ so $\mathbf{E}(\ln(\sigma_{t+1}/\sigma_t)|(\delta_0, \sigma_0)) = 1/(2d_\sigma n)(\mathbf{E}(\|\mathcal{G}(\delta_t, \mathcal{W}_t)\|^2|(\delta_0, \sigma_0)) - 2)$. As $\|\mathcal{G}\|^2$ is integrable with Fubini's theorem $\mathbf{E}(\|\mathcal{G}(\delta_t, \mathcal{W}_t)\|^2|(\delta_0, \sigma_0)) = \int_{\mathbb{R}^*_+} \mathbf{E}_{\mu_\mathcal{W}}(\|\mathcal{G}(\mathbf{y}, \mathcal{W})\|^2) P^t(\delta_0, d\mathbf{y})$, so $\mathbf{E}(\|\mathcal{G}(\delta_t, \mathcal{W}_t)\|^2|(\delta_0, \sigma_0)) - \mathbf{E}_{\pi_1 \times \mu_\mathcal{W}}(\|\mathcal{G}(\delta, \mathcal{W})\|^2) = \int_{\mathbb{R}^*_+} \mathbf{E}_{\mu_\mathcal{W}}(\|\mathcal{G}(\mathbf{y}, \mathcal{W})\|^2)(P^t(\mathbf{x}/\sigma, d\mathbf{y}) - \pi_1(d\mathbf{y}))$. According to Proposition 6 $(\delta_t)_{t\in\mathbb{N}}$ is V-geometrically ergodic with $V : \delta \mapsto \delta^\alpha + /\delta^{-\alpha}$, so there exists M_δ and $r > 1$ such that $\|P^t(\delta, \cdot) - \pi_1\|_V \leq M_\delta r^{-t}$. We showed that the function $\delta \mapsto \mathbf{E}(\|\mathcal{G}(\delta, \mathcal{W})\|^2)$ is bounded, so since $V(\delta) \geq 1$ for all $\delta \in \mathbb{R}^*_+$ there exists k such that $\mathbf{E}_{\mu_\mathcal{W}}(\|\mathcal{G}(\delta, \mathcal{W})\|^2) \leq kV(\delta)$ for all δ. Hence $|\int \mathbf{E}_{\mu_\mathcal{W}}(\|\mathcal{G}(x, \mathcal{W})\|^2)(P^t(\delta, dx) - \pi_1(dx))| \leq k\|P^t(\delta, \cdot) - \pi_1\|_V \leq kM_\delta r^{-t}$. And therefore $|\mathbf{E}(\|\mathcal{G}(\delta_t, \mathcal{W}_t)\|^2|(\delta_0, \sigma_0)) - \mathbf{E}_{\pi_1 \times \mu_\mathcal{W}}(\|\mathcal{G}(\delta, \mathcal{W}))\|^2) \leq kM_\delta r^{-t}$ which converges to 0 when t goes to infinity, which shows Eq. (46).

For (47) we have that $\mathbf{X}_{t+1} - \mathbf{X}_t \overset{d}{=} \sigma_t \mathcal{G}(\delta_t, \mathcal{W}_t)$ so $(1/t) \ln |(f(\mathbf{X}_{t+1}) - f(\mathbf{X}_t))/\sigma_0| \overset{d}{=} (1/t) \ln(\sigma_t/\sigma_0) + (1/t) \ln |f(\mathcal{G}(\delta_t, \mathcal{W}_t))/\sigma_0|$. From (13), since $1/2 \leq \Phi(x) \leq 1$ for all $x \geq 0$ and that $F_{1,\delta}(x) \leq 1$, the probability density function of $f(\mathcal{G}(\delta_t, \mathcal{W}_t)) = [\mathcal{G}(\delta_t, \mathcal{W}_t)]_1$ is dominated by $2\lambda\varphi(x)$. Hence

$$\begin{aligned}\Pr(\ln |[\mathcal{G}(\delta, \mathcal{W})]_1|/t \geq \epsilon) &\leq \int_{\mathbb{R}} \mathbf{1}_{[\epsilon t, +\infty)}(\ln |x|) 2\lambda\varphi(x) dx \\ &\leq \int_{\exp(\epsilon t)}^{+\infty} 2\lambda\varphi(x) dx + \int_{-\infty}^{-\exp(\epsilon t)} 2\lambda\varphi(x) dx\end{aligned}$$

For all $\epsilon > 0$ since φ is integrable with the dominated convergence theorem both members of the previous inequation converges to 0 when $t \to \infty$, which shows that $\ln |f(\mathcal{G}(\delta_t, \mathcal{W}_t))|/t$ converges in probability to 0. Since $\ln(\sigma_t/\sigma_0)/t$ converges in probability to the right hand side of (47) we get (47). □

If, for $c < 1$, the chain $(\delta_t, \mathbf{p}_t)_{t\in\mathbb{N}}$ was positive Harris with invariant measure π_c and V-ergodic such that $\|\mathbf{p}_{t+1}\|^2$ is dominated by V then we would obtain similar results with a convergence/divergence rate equal to $c/(2d_\sigma n)(\mathbf{E}_{\pi_c \otimes \mu_\mathcal{W}}(\|\mathbf{p}\|^2) - 2)$.

If the sign of the RHS of Eq. (45) is strictly positive then the step size diverges geometrically. The Law of Large Numbers entails that Monte Carlo simulations will converge to the RHS of Eq. 45, and the fact that the chain is V-geometrically ergodic (see Proposition 6) means sampling from the t-steps transition kernel P^t will get close exponentially fast to sampling directly from the stationary distribution π_1. We could apply a Central Limit Theorem for Markov chains (Meyn and Tweedie, 1993, Theorem 17.0.1), and get an approximate confidence interval for $\ln(\sigma_t/\sigma_0)/t$, given that we find a function V for which the chain $(\delta_t, \mathcal{W}_t)_{t\in\mathbb{N}}$ is V-uniformly ergodic and such that $\|\mathcal{G}(\delta, \mathbf{w})\|^4 \leq V(\delta, \mathbf{w})$. The question of the sign of $\lim_{t\to+\infty} f(\mathbf{X}_t) - f(\mathbf{X}_0)$ is not adressed in Theorem 2, but simulations indicate that for $d_\sigma \geq 1$ the probability that $f(\mathbf{X}_t) > f(\mathbf{X}_0)$ converges to 1 as $t \to +\infty$. For low enough values of d_σ and of θ this probability appears to converge to 0.

As in Fig. 3 we simulate the Markov chain $(\delta_t, \mathbf{p}_t)_{t\in\mathbb{N}}$ defined in Eq. (35) to obtain Fig. 4 after an average of δ_t over 10^6 time steps. The expected value $\mathbf{E}_{\pi_1}(\delta)$ shows the same dependency in λ that in the constant case, with larger population size, the algorithm follows the constraint from closer, as better samples are available closer to the constraint, which a larger population helps to find. The difference between $\mathbf{E}_{\pi_c}(\delta)$ and $\mathbf{E}_\pi(\delta)$ appears small except for large values of the constraint angle. When $\mathbf{E}_\pi(\delta) > \mathbf{E}_{\pi_c}(\delta)$ we observe on Fig. 6 that $\mathbf{E}_{\pi_c}(\ln(\sigma_{t+1}/\sigma_t)) > 0$.

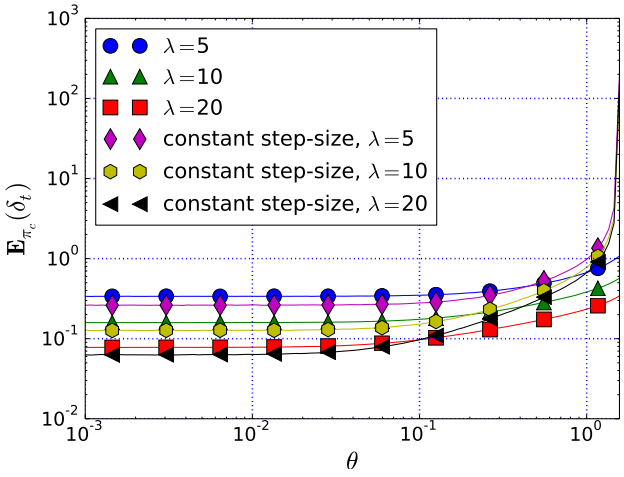


Figure 4: Average normalized distance δ from the constraint for the $(1, \lambda)$-CSA-ES plotted against the constraint angle θ, for $\lambda \in \{5, 10, 20\}$, $c = 1/\sqrt{2}$, $d_\sigma = 1$ and dimension 2.

In Fig. 5 the average of δ_t over 10^6 time steps is again plotted with $\lambda = 5$, this time for different values of the cumulation parameter, and compared with the constant step-size case. A lower value of c makes the algorithm follow the constraint from closer. When θ goes to 0 the value $\mathbf{E}_{\pi_c}(\delta)$ converges to a constant, and $\lim_{\theta\to 0} \mathbf{E}_{\pi}(\delta)$ for constant step-size seem to be $\lim_{\theta\to 0} \mathbf{E}_{\pi_c}(\delta)$ when c goes to 0. As in Fig. 4 the difference between $\mathbf{E}_{\pi_c}(\delta)$ and $\mathbf{E}_{\pi}(\delta)$ appears small except for large values of the constraint angle. This suggests that the difference between the distributions π and π_c is small. Therefore the approximation made in (Arnold, 2011a) where π is used instead of π_c to estimate $\ln(\sigma_{t+1}/\sigma_t)$ is accurate for not large values of the constraint angle.

In Fig. 6 the left hand side of Eq. (45) is simulated for 10^6 time steps against the constraint angle θ for different population sizes. This is the same as making an average of $\Delta_i = \ln(\sigma_{t+1}/\sigma_t)$ for i from 0 to $t-1$. If this value is below zero the step-size converges, which means a premature convergence of the algorithm. We see that a larger population size helps to achieve a faster divergence rate and for the step-size adaptation to succeed for a wider interval of values of θ.

In Fig. 7 like in the previous Fig. 6 the left hand side of Eq. (45) is simulated for 10^6 time steps against the constraint angle θ, this time for different values of the cumulation parameter c. A lower value of c yields a higher divergence rate for the step-size although $\mathbf{E}_{\pi_c}(\ln(\sigma_{t+1}/\sigma_t))$ appears to converge quickly when $c \to 0$. Lower values of c hence also allow success of the step-size adaptation for wider range values of θ, and in case of premature convergence a lower value of c means a lower convergence rate.

In Fig. 8 the left hand side of Eq. (45) is simulated for 10^4 time steps for the $(1, \lambda)$-CSA-ES plotted against the constraint angle θ, for $\lambda = 5$, $c = 1/\sqrt{2}$, $d_\sigma \in \{1, 0.5, 0.2, 0.1, 0.05\}$ and dimension 2. A lower value of d_σ allows larger change of step-size and induces here a bias towards increasing the step-size. This is confirmed in Fig. 8 where a low enough value of d_σ implies geometric divergence of the step-size regardless of the constraint angle. However simulations suggest that while for $d_\sigma \geq 1$

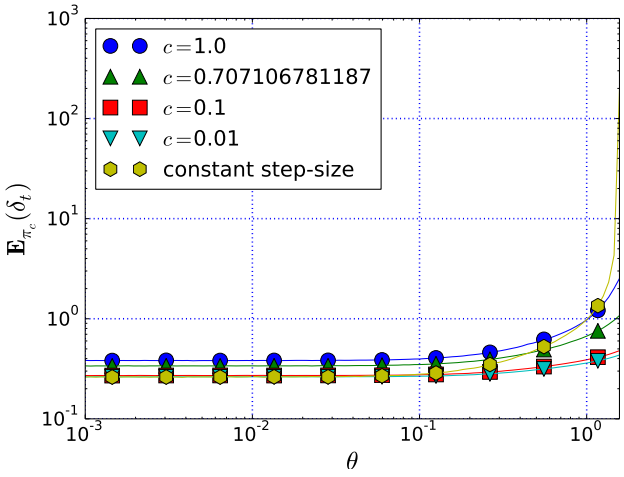


Figure 5: Average normalized distance δ from the constraint for the $(1, \lambda)$-CSA-ES plotted against the constraint angle θ with $c \in \{1, 1/\sqrt{2}, 0.1, 0.01\}$ and for constant step-size, where $\lambda = 5$, $d_\sigma = 1$ and dimension 2.

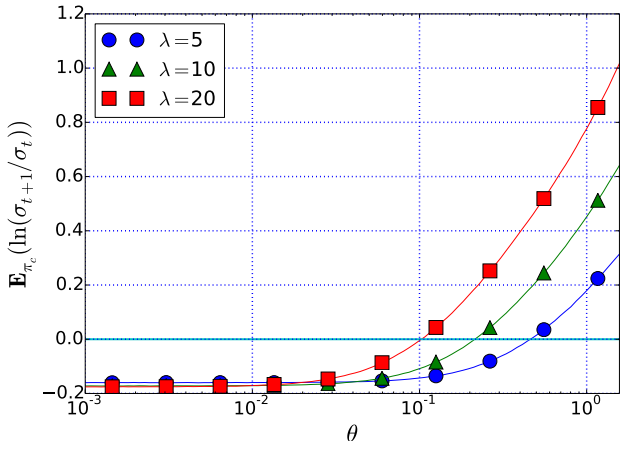


Figure 6: Average of the logarithmic adaptation response $\Delta_t = \ln(\sigma_{t+1}/\sigma_t)$ for the $(1, \lambda)$-CSA-ES plotted against the constraint angle θ, for $\lambda \in \{5, 10, 20\}$, $c = 1/\sqrt{2}$, $d_\sigma = 1$ and dimension 2. Values below zero (straight line) indicate premature convergence.

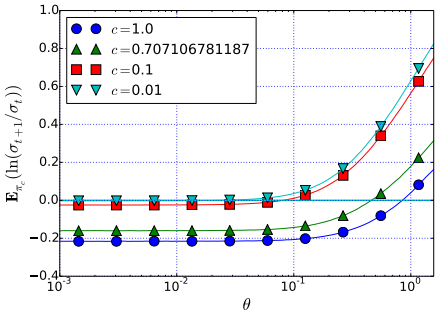


Figure 7: Average of the logarithmic adaptation response $\Delta_t = \ln(\sigma_{t+1}/\sigma_t)$ for the $(1, \lambda)$-CSA-ES plotted against the constraint angle θ, for $\lambda = 5$, $c \in \{1, 1/\sqrt{2}, 0.1, 0.01\}$, $d_\sigma = 1$ and dimension 2. Values below zero (straight line) indicate premature convergence.

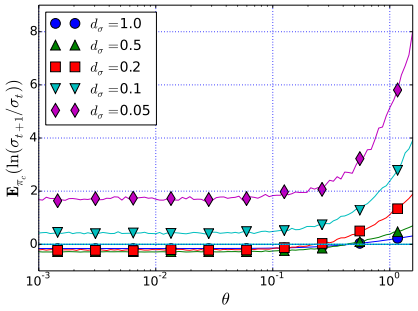


Figure 8: Average of the logarithmic adaptation response $\Delta_t = \ln(\sigma_{t+1}/\sigma_t)$ for the $(1, \lambda)$-CSA-ES plotted against the constraint angle θ, for $\lambda = 5$, $c = 1/\sqrt{2}$, $d_\sigma \in \{1, 0.5, 0.2, 0.1, 0.05\}$ and dimension 2. Values below zero (straight line) indicate premature convergence.

the probability that $f(\mathbf{X}_t) < f(\mathbf{X}_0)$ is close to 1, this probability decreases with smaller values of d_σ. The bias induced by a low value of d_σ may also prevent convergence when it is desired, as shown in Fig. 9.

In Fig. 9 the average of $\ln(\sigma_{t+1}/\sigma_t)$ is plotted against d_σ for the $(1, \lambda)$-CSA-ES minimizing a sphere function $f_{\text{sphere}} : \mathbf{x} \mapsto \|\mathbf{x}\|$, for $\lambda = 5$, $c \in \{1, 0.5, 0.2, 0.1\}$ and dimension 5, averaged over 10 runs. Low values of d_σ induce a bias towards increasing the step-size, which makes the algorithm diverge while convergence here is desired.

In Fig. 10 the smallest population size allowing geometric divergence is plotted against the constraint angle for different values of c. Any value of λ above the curve implies the geometric divergence of the step-size for the corresponding values of θ and c. We see that lower values of c allow for lower values of λ. It appears that the required value of λ scales inversely proportionally with θ. These curves were plotted by simulating runs of the algorithm for different values of θ and λ, and stopping the runs when the logarithm of the step-size had decreased or increased by 100 (for $c = 1$)

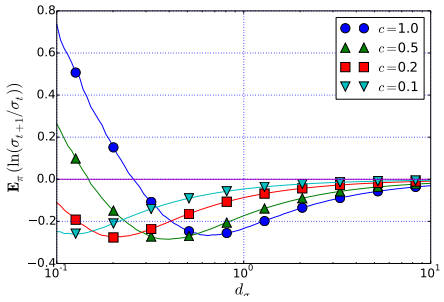


Figure 9: Average of the logarithmic adaptation response $\Delta_t = \ln(\sigma_{t+1}/\sigma_t)$ against d_σ for the $(1, \lambda)$-CSA-ES minimizing a sphere function for $\lambda = 5$, $c \in \{1, 0.5, 0.2, 0.1\}$, and dimension 5.

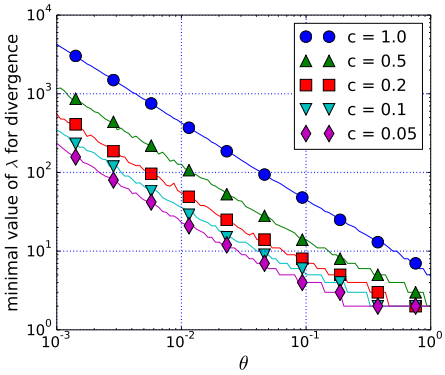


Figure 10: Minimal value of λ allowing geometric divergence for the $(1, \lambda)$-CSA-ES plotted against the constraint angle θ, for $c \in \{1., 0.5, 0.2, 0.05\}$, $d_\sigma = 1$ and dimension 2.

or 20 (for the other values of c). If the step-size had decreased (resp. increased) then this value of λ became a lower (resp. upper) bound for λ and a larger (resp. smaller) value of λ would be tested until the estimated upper and lower bounds for λ would meet. Also, simulations suggest that for increasing values of λ the probability that $f(\mathbf{X}_t) < f(\mathbf{X}_0)$ increases to 1, so large enough values of λ appear to solve the linear function on this constrained problem.

In Fig. 11 the highest value of c leading to geometric divergence of the step-size is plotted against the constraint angle θ for different values of λ. We see that larger values of λ allow higher values of c to be taken, and when $\theta \to 0$ the critical value of c appears proportional to θ^2. These curves were plotted following a similar scheme than with Fig. 10. For a certain θ the algorithm is ran with a certain value of c, and when the logarithm of the step-size has increased (resp. decreased) by more than $1000\sqrt{c}$ the run is stopped, the value of c tested becomes the new lower (resp. upper) bound for c and a new c taken between the lower and upper bounds is tested, until the lower and upper bounds are distant by less than the precision $\theta^2/10$. Similarly as with λ, simulations

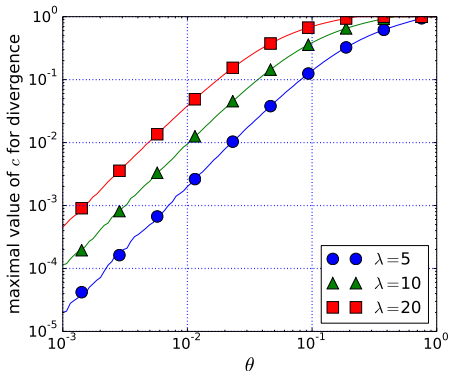


Figure 11: Transition boundary for c between convergence and divergence (lower value of c is divergence) for the $(1, \lambda)$-CSA-ES plotted against the constraint angle θ, for $\lambda \in \{5, 10, 20\}$ and dimension 2.

suggest that for decreasing values of c the probability that $f(\mathbf{X}_t) < f(\mathbf{X}_0)$ increases to 1, so small enough values of c appear to solve the linear function on this constrained problem.

6 Discussion

We investigated the $(1, \lambda)$-ES with constant step-size and cumulative step-size adaptation optimizing a linear function under a linear constraint handled by resampling unfeasible solutions. In the case of constant step-size or cumulative step-size adaptation when $c = 1$ we prove the stability (formally V-geometric ergodicity) of the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ defined as the normalised distance to the constraint, which was *pressumed* in Arnold (2011a). This property implies the divergence of the algorithm with constant step-size at a constant speed (see Theorem 1), and the geometric divergence or convergence of the algorithm with step-size adaptation (see Theorem 2). In addition, it ensures (fast) convergence of Monte Carlo simulations of the divergence rate, justifying their use.

In the case of cumulative step-size adaptation simulations suggest that geometric divergence occurs for a small enough cumulation parameter, c, or large enough population size, λ. In simulations we find the critical values for $\theta \to 0$ following $c \propto \theta^2$ and $\lambda \propto 1/\theta$. Smaller values of the constraint angle seem to increase the difficulty of the problem arbitrarily, i.e. no given values for c and λ solve the problem for *every* $\theta \in (0, \pi/2)$. However, when using a repair method instead of resampling in the $(1, \lambda)$-CSA-ES, fixed values of λ and c can solve the problem for every $\theta \in (0, \pi/2)$ (Arnold, 2013).

Using a different covariance matrix to generate new samples implies a change of the constraint angle (see Chotard and Holena 2014 for more details). Therefore, adaptation of the covariance matrix may render the problem arbitrarily close to the one with $\theta = \pi/2$. The unconstrained linear function case has been shown to be solved by a $(1, \lambda)$-ES with cumulative step-size adaptation for a population size larger than 3, regardless of other internal parameters (Chotard et al., 2012b). We believe this is one reason for using covariance matrix adaptation with ES when dealing with constraints,

as has been done in (Arnold and Hansen, 2012), as pure step-size adaptation has been shown to be liable to fail on even a very basic problem.

This work provides a methodology that can be applied to many ES variants. It demonstrates that a rigorous analysis of the constrained problem can be achieved. It relies on the theory of Markov chains for a continuous state space that once again proves to be a natural theoretical tool for analysing ESs, complementing particularly well previous studies (Arnold, 2011a, 2012; Arnold and Brauer, 2008).

Acknowledgments

This work was supported by the grants ANR-2010-COSI-002 (SIMINOLE) and ANR-2012-MONU-0009 (NumBBO) of the French National Research Agency.

References

Arnold, D. (2011a). On the behaviour of the (1,λ)-ES for a simple constrained problem. In *Foundations of Genetic Algorithms - FOGA 11*, pages 15–24. ACM.

Arnold, D. (2012). On the behaviour of the $(1, \lambda)$-σSA-ES for a constrained linear problem. In *Parallel Problem Solving from Nature - PPSN XII*, pages 82–91. Springer.

Arnold, D. and Brauer, D. (2008). On the behaviour of the $(1 + 1)$-ES for a simple constrained problem. In et al., G. R., editor, *Parallel Problem Solving from Nature - PPSN X*, pages 1–10. Springer.

Arnold, D. V. (2002). *Noisy Optimization with Evolution Strategies*. Kluwer Academic Publishers.

Arnold, D. V. (2011b). Analysis of a repair mechanism for the $(1, \lambda)$-ES applied to a simple constrained problem. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, GECCO 2011, pages 853–860, New York, NY, USA. ACM.

Arnold, D. V. (2013). Resampling versus repair in evolution strategies applied to a constrained linear problem. *Evolutionary computation*, 21(3):389–411.

Arnold, D. V. and Beyer, H.-G. (2004). Performance analysis of evolutionary optimization with cumulative step length adaptation. *IEEE Transactions on Automatic Control*, 49(4):617–622.

Arnold, D. V. and MacLeod, A. (2008). Step length adaptation on ridge functions. *Evolutionary Computation*, 16(2):151–184.

Arnold, Dirk, V. and Hansen, N. (2012). A (1+1)-CMA-ES for Constrained Optimisation. In Soule, T. and Moore, J. H., editors, *GECCO*, pages 297–304, Philadelphia, United States. ACM, ACM Press.

Chotard, A. and Auger, A. (201). Verifiable conditions for irreducibility, aperiodicity and weak feller property of a general markov chain. *(submitted) pre-print available at* `http://www.lri.fr/~auger/pdf/ChotardAugerBernoulliSub.pdf`.

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

Chotard, A., Auger, A., and Hansen, N. (2012b). Cumulative step-size adaptation on linear functions: Technical report. Technical report, Inria.

Chotard, A., Auger, A., and Hansen, N. (2014). Markov chain analysis of evolution strategies on a linear constraint optimization problem. In *Evolutionary Computation (CEC), 2014 IEEE Congress on*, pages 159–166.

Chotard, A. and Holena, M. (2014). A generalized markov-chain modelling approach to $(1, \lambda)$-es linear optimization. In Bartz-Beielstein, T., Branke, J., Filipič, B., and Smith, J., editors, *Parallel Problem Solving from Nature – PPSN XIII*, volume 8672 of *Lecture Notes in Computer Science*, pages 902–911. Springer International Publishing.

Hansen, N., Niederberger, S., Guzzella, L., and Koumoutsakos, P. (2009). A method for handling uncertainty in evolutionary optimization with an application to feedback control of combustion. *IEEE Transactions on Evolutionary Computation*, 13(1):180–197.

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

Meyn, S. P. and Tweedie, R. L. (1993). *Markov chains and stochastic stability*. Cambridge University Press, second edition.

Mezura-Montes, E. and Coello, C. A. C. (2008). Constrained optimization via multiobjective evolutionary algorithms. In *Multiobjective problem solving from nature*, pages 53–75. Springer.

Mezura-Montes, E. and Coello, C. A. C. (2011). Constraint-handling in nature-inspired numerical optimization: past, present and future. *Swarm and Evolutionary Computation*, 1(4):173–194.

Runarsson, T. P. and Yao, X. (2000). Stochastic ranking for constrained evolutionary optimization. *Evolutionary Computation, IEEE Transactions on*, 4(3):284–294.

Appendix

Proof of Lemma 4.

Proof. From Proposition 1 and Lemma 3 the density probability function of $\mathcal{G}(\delta, \mathcal{W})$ is $p^\star_\delta$, and from Eq. (12)

$$p^\star_\delta\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) = \lambda \frac{\varphi(x)\varphi(y)\mathbf{1}_{\mathbb{R}^*_+}\left(\delta - \begin{pmatrix} x \\ y \end{pmatrix}.\mathbf{n}\right)}{\Phi(\delta)} F_{1,\delta}(x)^{\lambda-1} \ .$$

From Eq. (10) $p_{1,\delta}(x) = \varphi(x)\Phi((\delta - x\cos\theta)/\sin\theta)/\Phi(\delta)$, so as $\delta{>}0$ we have $1 \geq \Phi(\delta){>}\Phi(0) = 1/2$, hence $p_{1,\delta}(x){<}2\varphi(x)$. So $p_{1,\delta}(x)$ converges when $\delta \to +\infty$ to $\varphi(x)$ while being bounded by $2\varphi(x)$ which is integrable. Therefore we can apply Lebesgue's dominated convergence theorem: $F_{1,\delta}$ converges to Φ when $\delta \to +\infty$ and is finite.

For $\delta \in \mathbb{R}^*_+$ and $(x, y) \in \mathbb{R}^2$ let $h_{\delta,y}(x)$ be $\exp(ax)p^\star_\delta((x, y))$. With Fubini-Tonelli's theorem $\mathbf{E}(\exp(\mathcal{G}(\delta, \mathcal{W}).(a, b))) = \int_{\mathbb{R}} \int_{\mathbb{R}} \exp(by)h_{\delta,y}(x)\mathrm{d}x\mathrm{d}y$. For $\delta \to +\infty$, $h_{\delta,y}(x)$ converges to $\exp(ax)\lambda\varphi(x)\varphi(y)\Phi(x)^{\lambda-1}$ while being dominated by $2\lambda\exp(ax)\varphi(x)\varphi(y)$, which is integrable. Therefore by the dominated convergence theorem and as the density of $\mathcal{N}_{\lambda:\lambda}$ is $x \mapsto \lambda\varphi(x)\Phi(x)^{\lambda-1}$, when $\delta \to +\infty$, $\int_{\mathbb{R}} h_{\delta,y}(x)\mathrm{d}x$ converges to $\varphi(y)\mathbf{E}(\exp(a\mathcal{N}_{\lambda:\lambda})) < \infty$.

So the function $y \mapsto \exp(by)\int_{\mathbb{R}} h_{\delta,y}(x)\mathrm{d}x$ converges to $y \mapsto \exp(by)\varphi(y)\mathbf{E}(\exp(a\mathcal{N}_{\lambda:\lambda}))$ while being dominated by $y \mapsto 2\lambda\varphi(y)\exp(by)\int_{\mathbb{R}}\exp(ax)\varphi(x)\mathrm{d}x$ which is integrable. Therefore we may apply the dominated convergence theorem: $\mathbf{E}(\exp(\mathcal{G}(\delta, \mathcal{W}).(a, b)))$ converges to $\int_{\mathbb{R}}\exp(by)\varphi(y)\mathbf{E}(\exp(a\mathcal{N}_{\lambda:\lambda}))\mathrm{d}y$ which equals to $\mathbf{E}(\exp(a\mathcal{N}_{\lambda:\lambda}))\mathbf{E}(\exp(b\mathcal{N}(0, 1)))$; and this quantity is finite.

The same reasoning gives that $\lim_{\delta\to\infty}\mathbf{E}(\bar{K}) < \infty$. □

Proof of Lemma 6.

Proof. As in Lemma 4, let E_1, E_2 and E_3 denote respectively $\mathbf{E}(\exp(-\frac{\alpha}{2d_\sigma n}(\mathcal{N}^2_{\lambda:\lambda} - 1)))$, $\mathbf{E}(\exp(-\frac{\alpha}{2d_\sigma n}(\mathcal{N}(0, 1)^2 - 1)))$, and $\mathbf{E}(\exp(-\frac{\alpha}{2d_\sigma n}(K - n + 2)))$, where K is a random variable following a chi-squared distribution with $n - 2$ degrees of freedom. Let us denote φ_χ the probability density function of K. Since $\varphi_\chi(z) = (1/2)^{(n-2)/2}/\Gamma((n - 2)/2)z^{(n-2)/2}\exp(-z/2)$, E_3 is finite.

Let h_δ be a function such that for $(x, y) \in \mathbb{R}^2$

$$h_\delta(x,y) = \frac{|\delta - ax - by|^\alpha}{\exp\left(\frac{\alpha}{2d_\sigma n}\left(x^2+y^2-2\right)\right)} \ ,$$

where $a := \cos\theta$ and $b := \sin\theta$.

From Proposition 1 and Lemma 3, the probability density function of $(\mathcal{G}(\delta, \mathcal{W}_t), K)$ is $p^\star_\delta \varphi_\chi$. Using the theorem of Fubini-Tonelli the expected value of the random variable $\frac{(\delta - \mathcal{G}(\delta,\mathcal{W}_t).\mathbf{n})^\alpha}{\eta^\star_1(\delta,\mathcal{W},K)^\alpha}$, that we denote E_δ, is

$$\begin{aligned}
E_\delta &= \int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}} \frac{|\delta - ax - by|^\alpha p^\star_\delta((x,y))\varphi_\chi(z)}{\exp\left(\frac{\alpha}{2d_\sigma}\left(\frac{\|(x,y)\|^2+z}{n} - 1\right)\right)} \mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}} \frac{|\delta - ax - by|^\alpha p^\star_\delta((x,y))\varphi_\chi(z)}{\exp\left(\frac{\alpha}{2d_\sigma n}\left(x^2+y^2-2\right)\right)\exp\left(\frac{\alpha}{2d_\sigma n}\left(z-(n-2)\right)\right)} \mathrm{d}z\mathrm{d}y\mathrm{d}x \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}} \frac{h_\delta(x,y) p^\star_\delta((x,y))\varphi_\chi(z)}{\exp\left(\frac{\alpha}{2d_\sigma n}\left(z-(n-2)\right)\right)} \mathrm{d}z\mathrm{d}y\mathrm{d}x \ .
\end{aligned}$$

Integration over z yields $E_\delta = \int_{\mathbb{R}}\int_{\mathbb{R}} h_\delta(x,y) p^\star_\delta((x,y))\mathrm{d}y\mathrm{d}x E_3$.

We now study the limit when $\delta \to +\infty$ of E_δ/δ^α. Let $\varphi_{\mathcal{N}_{\lambda:\lambda}}$ denote the probability density function of $\mathcal{N}_{\lambda:\lambda}$. For all $\delta \in \mathbb{R}^*_+$, $\Phi(\delta) > 1/2$, and for all $x \in \mathbb{R}$, $F_{1,\delta}(x) \leq 1$, hence with (9) and (12)

$$p^\star_\delta(x,y) = \lambda \frac{\varphi(x)\varphi(y)\mathbf{1}_{\mathbb{R}^*_+}(\delta - ax - by)}{\Phi(\delta)} F_{1,\delta}(x)^{\lambda-1} \leq \lambda \frac{\varphi(x)\varphi(y)}{\Phi(0)} \ , \tag{48}$$

and when $\delta \to +\infty$, as shown in the proof of Lemma 4, $p^\star_\delta((x,y))$ converges to $\varphi_{\mathcal{N}_{\lambda:\lambda}}(x)\varphi(y)$. For $\delta \geq 1$, $|\delta - ax - by|/\delta \leq 1 + |ax + by|$ with the triangular inequality. Hence

$$p^\star_\delta((x,y))\frac{h_\delta(x,y)}{\delta^\alpha} \leq \lambda \frac{\varphi(x)\varphi(y)}{\Phi(0)} \frac{(1+|ax+by|)^\alpha}{\exp\left(\frac{\alpha}{2d_\sigma n}\left(x^2+y^2-2\right)\right)} \quad \text{for } \delta \geq 1, \text{ and} \tag{49}$$

$$p^\star_\delta((x,y))\frac{h_\delta(x,y)}{\delta^\alpha} \underset{\delta\to+\infty}{\longrightarrow} \varphi_{\mathcal{N}_{\lambda:\lambda}}(x)\varphi(y) \frac{1}{\exp\left(\frac{\alpha}{2d_\sigma n}\left(x^2+y^2-2\right)\right)} \ . \tag{50}$$

Since the right hand side of (49) is integrable, we can use Lebesgue's dominated convergence theorem, and deduce from (50) that

$$\frac{E_\delta}{\delta^\alpha} = \int_{\mathbb{R}}\int_{\mathbb{R}} \frac{h_\delta(x,y)}{\delta^\alpha} p^\star_\delta((x,y))\mathrm{d}y\mathrm{d}x E_3 \underset{\delta\to+\infty}{\longrightarrow} \int_{\mathbb{R}}\int_{\mathbb{R}} \frac{\varphi_{\mathcal{N}_{\lambda:\lambda}}(x)\varphi(y)}{\exp\left(\frac{\alpha}{2d_\sigma n}\left(x^2+y^2-2\right)\right)} \mathrm{d}y\mathrm{d}x E_3$$

$$\text{and so} \quad \frac{E_\delta}{\delta^\alpha} \underset{\delta\to+\infty}{\longrightarrow} E_1E_2E_3 < \infty \ .$$

Since $\delta^\alpha/(\delta^\alpha + \delta^{-\alpha})$ converges to 1 when $\delta \to +\infty$, $E_\delta/(\delta^\alpha + \delta^{-\alpha})$ converges to $E_1E_2E_3$ when $\delta \to +\infty$, and $E_1E_2E_3$ is finite.

We now study the limit when $\delta \to 0$ of $\delta^\alpha E_\delta$, and restrict δ to $(0, 1]$. When $\delta \to 0$, $\delta^\alpha h_\delta(x,y)p^\star_\delta((x,y))$ converges to 0. Since we took $\delta \leq 1$, $|\delta + ax + by| \leq 1 + |ax + by|$,

and with (48) we have

$$\delta^\alpha h_\delta(x,y)p^\star_\delta((x,y)) \leq \lambda \frac{(1+|ax+by|)^\alpha \varphi(x)\varphi(y)}{\Phi(0)\exp\left(\frac{\alpha}{2d_\sigma n}\left(x^2+y^2-2\right)\right)} \quad \text{for } 0<\delta\leq 1 \ . \tag{51}$$

The right hand side of (51) is integrable, so we can apply Lebesgue's dominated convergence theorem, which shows that $\delta^\alpha E_\delta$ converges to 0 when $\delta \to 0$. And since $(1/\delta^\alpha)/(\delta^\alpha+\delta^{-\alpha})$ converges to 1 when $\delta \to 0$, $E_\delta/(\delta^\alpha+\delta^{-\alpha})$ also converges to 0 when $\delta \to 0$.

Let H_3 denote $\mathbf{E}(\exp(\alpha/(2d_\sigma n)(K-(n+2))))$. Since $\varphi_\chi(z) = (1/2)^{(n-2)/2}/\Gamma((n-2)/2)z^{(n-2)/2}\exp(-z/2)$, when α is close enough to 0, H_3 is finite. Let H_δ denote $\mathbf{E}(\delta_{t+1}^{-\alpha}|\delta_t=\delta)$, then

$$H_\delta = \int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}} \frac{p^\star_\delta((x,y))\varphi_\chi(z)\exp\left(\frac{\alpha}{2d_\sigma n}\left(z-(n-2)\right)\right)}{h_\delta(x,y)}\mathrm{d}z\mathrm{d}y\mathrm{d}x \ .$$

Integrating over z yields $H_\delta = \int_{\mathbb{R}}\int_{\mathbb{R}} \frac{p^\star_\delta((x,y))}{h_\delta(x,y)}\mathrm{d}y\mathrm{d}x H_3$.

We now study the limit when $\delta \to +\infty$ of H_δ/δ^α. With (48), we have that

$$\frac{p^\star_\delta((x,y))}{\delta^\alpha h_\delta(x,y)} \leq \lambda \frac{\varphi(x)\varphi(y)}{\Phi(0)} \frac{\exp\left(\frac{\alpha}{2d_\sigma n}\left(x^2+y^2-2\right)\right)}{\delta^\alpha|\delta-ax-by|^\alpha} \ .$$

With the change of variables $\tilde{x} = x-\delta/a$ we get

$$\begin{aligned}
\frac{p^\star_\delta((\tilde{x}+\frac{\delta}{a},y))}{\delta^\alpha h_\delta(\tilde{x}+\frac{\delta}{a},y)} &\leq \lambda \frac{\exp\left(-\frac{(\tilde{x}+\frac{\delta}{a})^2}{2}\right)\varphi(y)}{\sqrt{2\pi}\Phi(0)} \frac{\exp\left(\frac{\alpha}{2d_\sigma n}\left(\left(\tilde{x}+\frac{\delta}{a}\right)^2+y^2-2\right)\right)}{\delta^\alpha|a\tilde{x}+by|^\alpha} \\
&\leq \lambda \frac{\varphi(\tilde{x})\varphi(y)}{\Phi(0)} \frac{\exp\left(\frac{\alpha}{2d_\sigma n}\left(\tilde{x}^2+y^2-2\right)\right)}{|a\tilde{x}+by|^\alpha} \frac{\exp\left(\left(\frac{\alpha}{2d_\sigma n}-\frac{1}{2}\right)\left(2\frac{\delta}{a}\tilde{x}+\frac{\delta^2}{a^2}\right)\right)}{\delta^\alpha} \\
&\leq \lambda \frac{\varphi(\tilde{x})\varphi(y)}{\Phi(0)} \frac{1}{h_0(\tilde{x},y)} \frac{\exp\left(\left(\frac{\alpha}{2d_\sigma n}-\frac{1}{2}\right)\left(2\frac{\delta}{a}\tilde{x}+\frac{\delta^2}{a^2}\right)\right)}{\exp(\alpha\ln(\delta))} \ .
\end{aligned}$$

An upper bound for all $\delta \in \mathbb{R}_+^*$ of the right hand side of the previous inequation is related to an upper bound of the function $l : \delta \in \mathbb{R}_+^* \mapsto (\alpha/(2d_\sigma n)-1/2)(2(\delta/a)\tilde{x}+\delta^2/a^2)-\alpha\ln(\delta)$. And since we are interested in a limit when $\delta \to +\infty$, we can restrict our search of an upper bound of l to $\delta \geq 1$. Let $c := \alpha/(2d_\sigma n)-1/2$. We take α small enough to ensure that c is negative. An upper bound to l can be found through derivation:

$$\begin{aligned}
\frac{\partial l(\delta)}{\partial\delta} = 0 &\Leftrightarrow 2\frac{c}{a^2}\delta + 2\frac{c}{a}\tilde{x} - \frac{\alpha}{\delta} = 0 \\
&\Leftrightarrow 2\frac{c}{a^2}\delta^2 + 2\frac{c}{a}\tilde{x}\delta - \alpha = 0
\end{aligned}$$

The discriminant of the quadratic equation is $\Delta = 4(c^2/a^2)\tilde{x}^2 + 8\alpha c/a^2$. The derivative of l multiplied by δ is a quadratic function with a negative quadratic coefficient $2c/a^2$. Since we restricted δ to $[1,+\infty)$, multiplying the derivative of l by δ leaves its sign unchanged. So the maximum of l is attained for δ equal to 1 or for

δ equal to $\delta_M := (-2c/a\tilde{x} - \sqrt{\Delta})/(4c/a^2)$, and so $l(\delta) \leq \max(l(1), l(\delta_M))$ for all $\delta \in [1, +\infty)$. We also have that $\lim_{\tilde{x}\to\infty} \sqrt{\Delta}/\tilde{x} = 2|c|/a = -2ca$, so $\lim_{\tilde{x}\to\infty} \delta_M/\tilde{x} = (-2c/a - (-2c/a))/(4c/a^2) = 0$. Hence when $|\tilde{x}|$ is large enough, $\delta_M \leq 1$, so since we restricted δ to $[1, +\infty)$ there exists $m > 0$ such that if $|\tilde{x}| > m$, $l(\delta) \leq l(1)$ for all $\delta \in [1, +\infty)$. And trivially, $l(\delta)$ is bounded for all $\tilde{x}$ in the compact set $[-m, m]$ by a constant $M > 0$, so $l(\delta) \leq \max(M, l(1)) \leq M + |l(1)|$ for all $\tilde{x} \in \mathbb{R}$ and all $\delta \in [1, +\infty)$. Therefore

$$\begin{aligned}\frac{p^\star_\delta((\tilde{x} + \frac{\delta}{a}, y))}{\delta^\alpha h_\delta(\tilde{x} + \frac{\delta}{a}, y)} &\leq \lambda \frac{\varphi(\tilde{x})\varphi(y)}{\Phi(0)} \frac{1}{h_0(\tilde{x}, y)} \exp(M + |l(1)|) \\ &\leq \lambda \frac{\varphi(\tilde{x})\varphi(y)}{\Phi(0)} \frac{1}{h_0(\tilde{x}, y)} \exp\left(M + \left|2\frac{c}{a}\tilde{x} + \frac{c}{a^2}\right|\right) \ .\end{aligned}$$

For α small enough, the right hand side of the previous inequation is integrable. And since the left hand side of this inequation converges to 0 when $\delta \to +\infty$, according to Lebesgue's dominated convergence theorem H_δ/δ^α converges to 0 when $\delta \to +\infty$. And since $\delta^\alpha/(\delta^\alpha + \delta^{-\alpha})$ converges to 1 when $\delta \to +\infty$, $H_\delta/(\delta^\alpha + \delta^{-\alpha})$ also converges to 0 when $\delta \to +\infty$.

We now study the limit when $\delta \to 0$ of $H_\delta/(\delta^\alpha + \delta^{-\alpha})$. Since we are interested in the limit for $\delta \to 0$, we restrict δ to $(0, 1]$. Similarly as what was done previously, with the change of variables $\tilde{x} = x - \delta/a$,

$$\begin{aligned}\frac{p^\star_\delta((\tilde{x} + \frac{\delta}{a}, y))}{(\delta^\alpha + \delta^{-\alpha})h_\delta(\tilde{x} + \frac{\delta}{a}, y)} &\leq \lambda \frac{\varphi(\tilde{x})\varphi(y)}{\Phi(0)} \frac{1}{h_0(\tilde{x}, y)} \frac{\exp\left(\left(\frac{\alpha}{2d_\sigma n} - \frac{1}{2}\right)\left(2\frac{\delta}{a}\tilde{x} + \frac{\delta^2}{a^2}\right)\right)}{\delta^\alpha + \delta^{-\alpha}} \\ &\leq \lambda \frac{\varphi(\tilde{x})\varphi(y)}{\Phi(0)h_0(\tilde{x}, y)} \exp\left(\left(\frac{\alpha}{2d_\sigma n} - \frac{1}{2}\right)\left(2\frac{\delta}{a}\tilde{x} + \frac{\delta^2}{a^2}\right)\right) \ .\end{aligned}$$

Take α small enough to ensure that $\alpha/(2d_\sigma n) - 1/2$ is negative. Then an upper bound for $\delta \in (0, 1]$ of the right hand side of the previous inequality is related to an upper bound of the function $k : \delta \in (0, 1] \mapsto 2\delta\tilde{x}/a + \delta^2/a^2$. This maximum can be found through derivation: $\partial k(\delta)/\partial\delta = 0$ is equivalent to $2\tilde{x}/a + 2\delta/a^2 = 0$, and so the maximum of k is realised at $\delta_M := -a\tilde{x}$. However, since we restricted δ to $(0, 1]$, for $\tilde{x} \geq 0$ we have $\delta_M \leq 0$ so an upper bound of k in $(0, 1]$ is realized at 0, and for $\tilde{x} \leq -1/a$ we have $\delta_M \geq 1$ so the maximum of k in $(0, 1]$ is realized at 1. Furthermore, $k(\delta_M) = -2\tilde{x}^2 + \tilde{x}^2 = -\tilde{x}^2$ so when $-1/a < \tilde{x} < 0$, $k(\delta) < 1/a^2$. Therefore $k(\delta) \leq \max(k(0), k(1), 1/a^2)$. Note that $k(0) = 0$ which is inferior to $1/a^2$, and note that $k(1) = 2c\tilde{x}/a + /a^2$. Hence $k(\delta) \leq \max(2\tilde{x}/a + 1/a^2, 1/a^2) \leq |2\tilde{x}/a + 1/a^2| + 1/a^2$, and so

$$\frac{p^\star_\delta((\tilde{x} + \frac{\delta}{a}, y))}{(\delta^\alpha + \delta^{-\alpha})h_\delta(\tilde{x} + \frac{\delta}{a}, y)} \leq \lambda \frac{\varphi(\tilde{x})\varphi(y)}{\Phi(0)h_0(\tilde{x}, y)} \exp\left(\left(\frac{\alpha}{2d_\sigma n} - \frac{1}{2}\right)\left(\left|2\frac{\tilde{x}}{a} + \frac{1}{a^2}\right| + \frac{1}{a^2}\right)\right) \ .$$

For α small enough the right hand side of the previous inequation is integrable. Since the left hand side of this inequation converges to 0 when $\delta \to 0$, we can apply Lebesgue's dominated convergence theorem, which proves that $H_\delta/(\delta^\alpha + \delta^{-\alpha})$ converges to 0 when $\delta \to 0$.

□

4.3.2 Paper: A Generalized Markov Chain Modelling Approach to $(1,\lambda)$-ES Linear Optimization

The article presented here is a technical report [48] which includes [47], which was published at the conference Parallel Problem Solving from Nature in 2014, and the full proofs for every proposition of [47]. The subject of this paper has been proposed by the second author as an extension of the study conducted in [44] on the $(1,\lambda)$-ES with constant step-size on a linear function with a linear constraint to more general sampling distributions, i.e. for H a distribution, the sampling of the candidates writes

$$\boldsymbol{Y}_t^{i,j} = \boldsymbol{X}_t + \sigma \boldsymbol{M}_t^{i,j} \quad , \quad (\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda],t\in\mathbb{N},j\in\mathbb{N}} \text{ i.i.d., } \boldsymbol{M}_t^{i,j} \sim H \quad , \tag{4.13}$$

where $\boldsymbol{Y}_t^{i,j}$ denotes at iteration $t \in \mathbb{N}$ the sample obtained after j resampling for the i^{th} feasible sample, in case all the previous samples $(\boldsymbol{Y}_t^{i,k})_{k\in[1..j-1]}$ were unfeasible. Although the use of H as Gaussian distributions is justified in the black-box optimization context as Gaussian distributions are maximum entropy probability distributions, when more information is available (e.g. separability of the function or multimodality) the use of other sampling distributions may be preferable (e.g. see [64] for an analysis of some heavy-tailed distributions). Furthermore, since in the study presented in 4.3.1 the Gaussian sampling distribution is assumed to have identity covariance matrix, in this article different covariance matrices can be taken, and so the influence of the covariance matrix on the problem can be investigated.
The article presented here starts by analysing how the sampling distribution H is impacted by the resampling. It then shows that the sequence $(\delta_t)_{t\in\mathbb{N}}$ which is defined as the signed distance from $\boldsymbol{X}_t$ to the constraint normalized by the step-size (i.e. $\delta_t := -g(\boldsymbol{X}_t)/\sigma$) is a Markov chain. It then gives sufficient conditions on the distribution H for $(\delta_t)_{t\in\mathbb{N}}$ to be positive Harris recurrent and ergodic or geometrically ergodic (note that heavy-tailed distributions do not follow the condition for geometrical ergodicity). The positivity and Harris recurrence of the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is then used to show that the sequence $(f(\boldsymbol{X}_t))_{t\in\mathbb{N}}$ diverges almost surely similarly as in (4.12). The paper then investigates more specific distributions: it recovers the results of [44] with isotropic Gaussian distributions for a $(1,\lambda)$-ES with constant step-size, and shows that a different covariance matrix for the sampling distribution is equivalent to a different norm on the search space, which implies a different constraint angle θ. Since, as seen in [14, 15, 45] small values of the constraint angle cause the step-size adaptation to fail, adapting the covariance matrix to the problem could allow the step-size to successfully diverge log-linearly on the linear function with a linear constraint. Finally, sufficient conditions on the marginals of the sampling distribution and the copula combining them are given to get the absolute continuity of the sampling distribution.

A Generalized Markov-Chain Modelling Approach to $(1, \lambda)$-ES Linear Optimization

Alexandre Chotard [1] and Martin Holeňa [2]

[1] INRIA Saclay-Ile-de-France, LRI, alexandre.chotard@lri.fr University Paris-Sud, France
[2] Institute of Computer Science, Academy of Sciences, Pod vodárenskou věží 2, Prague, Czech Republic, martin@cs.cas.cz

Abstract. Several recent publications investigated Markov-chain modelling of linear optimization by a $(1, \lambda)$-ES, considering both unconstrained and linearly constrained optimization, and both constant and varying step size. All of them assume normality of the involved random steps, and while this is consistent with a black-box scenario, information on the function to be optimized (e.g. separability) may be exploited by the use of another distribution. The objective of our contribution is to complement previous studies realized with normal steps, and to give sufficient conditions on the distribution of the random steps for the success of a constant step-size $(1, \lambda)$-ES on the simple problem of a linear function with a linear constraint. The decomposition of a multidimensional distribution into its marginals and the copula combining them is applied to the new distributional assumptions, particular attention being paid to distributions with Archimedean copulas.

Keywords: evolution strategies, continuous optimization, linear optimization, linear constraint, linear function, Markov chain models, Archimedean copulas

1 Introduction

Evolution Strategies (ES) are Derivative Free Optimization (DFO) methods, and as such are suited for the optimization of numerical problems in a black-box context, where the algorithm has no information on the function f it optimizes (e.g. existence of gradient) and can only query the function's values. In such a context, it is natural to assume normality of the random steps, as the normal distribution has maximum entropy for given mean and variance, meaning that it is the most general assumption one can make without the use of additional information on f. However such additional information may be available, and then using normal steps may not be optimal. Cases where different distributions have been studied include so-called Fast Evolution Strategies [1] or SNES [2, 3] which exploits the separability of f, or heavy-tail distributions on multimodal problems [4, 3].

In several recent publications [5–8], attention has been paid to Markov-chain modelling of linear optimization by a $(1, \lambda)$-ES, i.e. by an evolution strategy in

which λ children are generated from a single parent $\boldsymbol{X} \in \mathbb{R}^n$ by adding normally distributed n-dimensional random steps $\boldsymbol{M}$,

$$\boldsymbol{X} \leftarrow \boldsymbol{X} + \sigma \mathbf{C}^{\frac{1}{2}} \boldsymbol{M}, \text{ where } \boldsymbol{M} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n). \tag{1}$$

Here, σ is called step size, $\mathbf{C}$ is a covariance matrix, and $\mathcal{N}(\mathbf{0}, \mathbf{I}_n)$ denotes the n-dimensional standard normal distribution with zero mean and covariance matrix identity. The best among the λ children, i.e. the one with the highest fitness, becomes the parent of the next generation, and the step-size σ and the covariance matrix $\mathbf{C}$ may then be adapted to increase the probability of sampling better children. In this paper we relax the normality assumption of the movement $\boldsymbol{M}$ to a more general distribution H.

The linear function models a situation where the step-size is relatively small compared to the distance towards a local optimum. This is a simple problem that must be solved by any effective evolution strategy by diverging with positive increments of $\nabla f.\boldsymbol{M}$. This unconstrained case was studied in [7] for normal steps with cumulative step-size adaptation (the step-size adaptation mechanism in CMA-ES [9]).

Linear constraints naturally arise in real-world problems (e.g. need for positive values, box constraints) and also model a step-size relatively small compared to the curvature of the constraint. Many techniques to handle constraints in randomised algorithms have been proposed (see [10]). In this paper we focus on the resampling method, which consists in resampling any unfeasible candidate until a feasible one is sampled. We chose this method as it makes the algorithm easier to study, and is consistent with the previous studies assuming normal steps [11, 5, 6, 8], studying constant step-size, self adaptation and cumulative step-size adaptation mechanisms (with fixed covariance matrix).

Our aim is to study the $(1, \lambda)$-ES with constant step-size, constant covariance matrix and random steps with a general absolutely continuous distribution H optimizing a linear function under a linear constraint handled through resampling. We want to extend the results obtained in [5, 8] using the theory of Markov chains. It is our hope that such results will help in designing new algorithms using information on the objective function to make non-normal steps. We pay a special attention to distributions with Archimedean copulas, which are a particularly well transparent alternative to the normal distribution. Such distributions have been recently considered in the Estimation of Distribution Algorithms [12, 13], continuing the trend of using copulas in that kind of evolutionary optimization algorithms [14].

In the next section, the basic setting for modelling the considered evolutionary optimization task is formally defined. In Section 3, the distributions of the feasible steps and of the selected steps are linked to the distribution of the random steps, and another way to sample them is provided. In Section 4, it is shown that, under some conditions on the distribution of the random steps, the normalized distance to the constraint defined in (5) is a ergodic Markov chain, and a law of large numbers for Markov chains is applied. Finally, Section 5 gives properties on the distribution of the random steps under which some of the aforementioned conditions are verified.

Notations

For $(a, b) \in \mathbb{N}^2$ with $a < b$, $[a..b]$ denotes the set of integers i such that $a \leq i \leq b$. For X and Y two random vectors, $X \overset{(d)}{=} Y$ denotes that these variables are equal in distribution, $X \overset{a.s.}{\to} Y$ and $X \overset{\mathcal{P}}{\to} Y$ denote, respectively, almost sure convergence and convergence in probability. For $(\boldsymbol{x}, \boldsymbol{y}) \in \mathbb{R}^n$, $\boldsymbol{x}.\boldsymbol{y}$ denotes the scalar product between the vectors $\boldsymbol{x}$ and $\boldsymbol{y}$, and for $i \in [1..n]$, $[\boldsymbol{x}]_i$ denotes the i^{th} coordinate of $\boldsymbol{x}$. For A a subset of $\mathbb{R}^n$, $\mathbb{1}_A$ denotes the indicator function of A. For $\mathcal{X}$ a topological set, $\mathcal{B}(\mathcal{X})$ denotes the Borel algebra on $\mathcal{X}$.

2 Problem setting and algorithm definition

Throughout this paper, we study a $(1, \lambda)$-ES optimizing a linear function $f : \mathbb{R}^n \to \mathbb{R}$ where $\lambda \geq 2$ and $n \geq 2$, with a linear constraint $g : \mathbb{R}^n \to \mathbb{R}$, handling the constraint by resampling unfeasible solutions until a feasible solution is sampled.

Take $(\boldsymbol{e}_k)_{k\in[1..n]}$ a orthonormal basis of $\mathbb{R}^n$. We may assume ∇f to be normalized as the behaviour of an ES is invariant to the composition of the objective function by a strictly increasing function (e.g. $h : x \mapsto x/\|\nabla f\|$), and the same holds for ∇g since our constraint handling method depends only on the inequality $g(\boldsymbol{x}) \leq 0$ which is invariant to the composition of g by a homothetic transformation. Hence w.l.o.g. we assume that $\nabla f = \boldsymbol{e}_1$ and $\nabla g = \cos\theta \boldsymbol{e}_1 + \sin\theta \boldsymbol{e}_2$ with the set of feasible solutions $\mathcal{X}_{\text{feasible}} := \{\boldsymbol{x} \in \mathbb{R}^n | g(\boldsymbol{x}) \leq 0\}$. We restrict our study to $\theta \in (0, \pi/2)$. Overall the problem reads

$$\begin{aligned} &\text{maximize} \quad f(\boldsymbol{x}) = [\boldsymbol{x}]_1 \quad \text{subject to} \\ &g(\boldsymbol{x}) = [\boldsymbol{x}]_1 \cos\theta + [\boldsymbol{x}]_2 \sin\theta \leq 0 \ . \end{aligned} \tag{2}$$

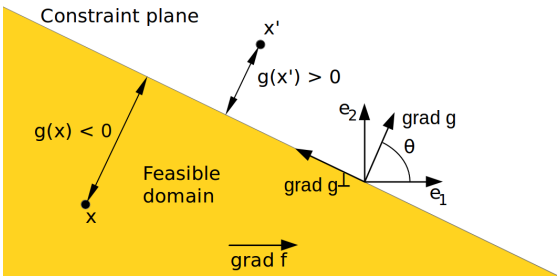


Fig. 1. Linear function with a linear constraint, in the plane spanned by ∇f and ∇g, with the angle from ∇f to ∇g equal to $\theta \in (0, \pi/2)$. The point $\boldsymbol{x}$ is at distance $g(\boldsymbol{x})$ from the constraint hyperplan $g(\boldsymbol{x}) = 0$.

At iteration $t \in \mathbb{N}$, from a so-called parent point $\boldsymbol{X}_t \in \mathcal{X}_{\text{feasible}}$ and with step-size $\sigma_t \in \mathbb{R}_+^*$ we sample new candidate solutions by adding to $\boldsymbol{X}_t$ a random

vector $\sigma_t \boldsymbol{M}_t^{i,j}$ where $\boldsymbol{M}_t^{i,j}$ is called a random step and $(\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda],j\in\mathbb{N},t\in\mathbb{N}}$ is a i.i.d. sequence of random vectors with distribution H. The i index stands for the λ new samples to be generated, and the j index stands for the unbounded number of samples used by the resampling. We denote $\boldsymbol{M}_t^i$ a feasible step, that is the first element of $(\boldsymbol{M}_t^{i,j})_{j\in\mathbb{N}}$ such that $\boldsymbol{X}_t + \sigma_t \boldsymbol{M}_t^i \in \mathcal{X}_{\text{feasible}}$ (random steps are sampled until a suitable candidate is found). The i^{th} feasible solution $\boldsymbol{Y}_t^i$ is then

$$\boldsymbol{Y}_t^i := \boldsymbol{X}_t + \sigma_t \boldsymbol{M}_t^i \ . \tag{3}$$

Then we denote $\star := \operatorname{argmax}_{i\in[1..\lambda]} f(\boldsymbol{Y}_t^i)$ the index of the feasible solution maximizing the function f, and update the parent point

$$\boldsymbol{X}_{t+1} := \boldsymbol{Y}_t^\star = \boldsymbol{X}_t + \sigma_t \boldsymbol{M}_t^\star \ , \tag{4}$$

where $\boldsymbol{M}_t^\star$ is called the selected step. Then the step-size σ_t, the distribution of the random steps H or other internal parameters may be adapted.

Following [5, 6, 11, 8] we define δ_t as

$$\delta_t := -\frac{g(\boldsymbol{X}_t)}{\sigma_t} \ . \tag{5}$$

3 Distribution of the feasible and selected steps

In this section we link the distributions of the random vectors $\boldsymbol{M}_t^i$ and $\boldsymbol{M}_t^\star$ to the distribution of the random steps $\boldsymbol{M}_t^{i,j}$, and give another way to sample $\boldsymbol{M}_t^i$ and $\boldsymbol{M}_t^\star$ not requiring an unbounded number of samples.

Lemma 1. *Let a $(1,\lambda)$-ES optimize the problem defined in (2) handling constraint through resampling. Take H the distribution of the random step $\boldsymbol{M}_t^{i,j}$, and for $\delta \in \mathbb{R}_+^*$ denote $L_\delta := \{\boldsymbol{x} \in \mathbb{R}^n | g(\boldsymbol{x}) \leq \delta\}$. Providing that H is absolutely continuous and that $H(L_\delta) > 0$ for all $\delta \in \mathbb{R}_+$, the distribution $\tilde{H}_\delta$ of the feasible step and $\tilde{H}_\delta^\star$ the distribution of the selected step when $\delta_t = \delta$ are absolutely continuous, and denoting h, $\tilde{h}_\delta$ and $\tilde{h}_\delta^\star$ the probability density functions of, respectively, the random step, the feasible step $\boldsymbol{M}_t^i$ and the selected step $\boldsymbol{M}_t^\star$ when $\delta_t = \delta$*

$$\tilde{h}_\delta(\boldsymbol{x}) = \frac{h(\boldsymbol{x})\mathbb{1}_{L_\delta}(\boldsymbol{x})}{H(L_\delta)} \ , \tag{6}$$

and

$$\begin{aligned}\tilde{h}_\delta^\star(\boldsymbol{x}) &= \lambda \tilde{h}_\delta(\boldsymbol{x}) \tilde{H}_\delta((-\infty, [\boldsymbol{x}]_1) \times \mathbb{R}^{n-1})^{\lambda-1} \\ &= \lambda \frac{h(\boldsymbol{x})\mathbb{1}_{L_\delta}(\boldsymbol{x}) H((-\infty, [\boldsymbol{x}]_1) \times \mathbb{R}^{n-1} \cap L_\delta)^{\lambda-1}}{H(L_\delta)^\lambda} \ .\end{aligned} \tag{7}$$

Proof. Let $\delta > 0, A \in \mathcal{B}(\mathbb{R}^n)$. Then for $t \in \mathbb{N}, i = 1 \dots \lambda$, using the the fact that $(\boldsymbol{M}_t^{i,j})_{j\in\mathbb{N}}$ is a i.i.d. sequence

$$\begin{aligned}
\tilde{H}_\delta(A) &= \Pr(\boldsymbol{M}_t^i \in A | \delta_t = \delta) \\
&= \sum_{j\in\mathbb{N}} \Pr(\boldsymbol{M}_t^{i,j} \in A \cap L_\delta \text{ and } \forall k < j, \boldsymbol{M}_t^{i,k} \in L_\delta^c | \delta_t = \delta) \\
&= \sum_{j\in\mathbb{N}} \Pr(\boldsymbol{M}_t^{i,j} \in A \cap L_\delta | \delta_t = \delta) \Pr(\forall k < j, \boldsymbol{M}_t^{i,k} \in L_\delta^c | \delta_t = \delta) \\
&= \sum_{j\in\mathbb{N}} H(A \cap L_\delta)(1 - H(L_\delta))^j \\
&= \frac{H(A \cap L_\delta)}{H(L_\delta)} = \int_A \frac{h(\boldsymbol{x}) \mathbb{1}_{L_\delta}(\boldsymbol{x}) d\boldsymbol{x}}{H(L_\delta)} \ ,
\end{aligned}$$

which yield Eq. (6) and that $\tilde{H}_\delta$ admits a density $\tilde{h}_\delta$ and is therefore absolutely continuous.

Since $((\boldsymbol{M}_t^{i,j})_{j\in\mathbb{N}})_{i\in[1..\lambda]}$ is i.i.d., $(\boldsymbol{M}_t^i)_{i\in[1..\lambda]}$ is i.i.d. and

$$\begin{aligned}
\tilde{H}_\delta^\star(A) &= \Pr(\boldsymbol{M}_t^\star \in A | \delta_t = \delta) \\
&= \sum_{i=1}^{\lambda} \Pr(\boldsymbol{M}_t^i \in A \text{ and } \forall j \in [1..\lambda] \backslash \{i\}, [\boldsymbol{M}_t^i]_1 > [\boldsymbol{M}_t^j]_1 | \delta_t = \delta) \\
&= \lambda \Pr(\boldsymbol{M}_t^1 \in A \text{ and } \forall j \in [2..\lambda], [\boldsymbol{M}_t^1]_1 > [\boldsymbol{M}_t^j]_1 | \delta_t = \delta) \\
&= \lambda \int_A \tilde{h}_\delta(\boldsymbol{x}) \Pr(\forall j \in [2..\lambda], [\boldsymbol{M}_t^j]_1 < [\boldsymbol{x}]_1 | \delta_t = \delta) d\boldsymbol{x} \\
&= \int_A \lambda \tilde{h}_\delta(\boldsymbol{x}) \tilde{H}_\delta((-\infty, [\boldsymbol{x}]_1) \times \mathbb{R}^{n-1})^{\lambda-1} d\boldsymbol{x} \ ,
\end{aligned}$$

which shows that $\tilde{H}_\delta^\star$ possess a density, and with (6) yield Eq. (7). □

The vectors $(\boldsymbol{M}_t^i)_{i\in[1..\lambda]}$ and $\boldsymbol{M}_t^\star$ are functions of the vectors $(\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda],j\in\mathbb{N}}$ and of δ_t. In the following Lemma an equivalent way to sample $\boldsymbol{M}_t^i$ and $\boldsymbol{M}_t^\star$ is given which uses a finite number of samples. This method is useful if one wants to avoid dealing with the infinite dimension space implied by the sequence $(\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda,j\in\mathbb{N}}$.

Lemma 2. *Let a $(1, \lambda)$-ES optimize problem (2), handling the constraint through resampling, and take δ_t as defined in (5). Let H denote the distribution of $\boldsymbol{M}_t^{i,j}$ that we assume absolutely continuous, $\nabla g^\perp := -\sin\theta \boldsymbol{e}_1 + \cos\theta \boldsymbol{e}_2$, $\mathbf{Q}$ the rotation matrix of angle θ changing $(\boldsymbol{e}_1, \boldsymbol{e}_2, \dots, \boldsymbol{e}_n)$ into $(\nabla g, \nabla g^\perp, \dots, \boldsymbol{e}_n)$. Take $F_{1,\delta}(x) := \Pr(\boldsymbol{M}_t^i.\nabla g \le x | \delta_t = \delta)$, $F_{2,\delta}(x) := \Pr(\boldsymbol{M}_t^i.\nabla g^\perp \le x | \delta_t = \delta)$ and $F_{k,\delta}(x) := \Pr([\boldsymbol{M}_t^i]_k \le x | \delta_t = \delta)$ for $k \in [3..n]$, the marginal cumulative distribution functions when $\delta_t = \delta$, and C_δ the copula of $(\boldsymbol{M}_t^i.\nabla g, \boldsymbol{M}_t^i.\nabla g^\perp, \dots, \boldsymbol{M}_t^i.\boldsymbol{e}_n)$.*

We define

$$\mathcal{G} : (\delta, (u_i)_{i\in[1..n]}) \in \mathbb{R}_+ \times [0,1]^n \mapsto \mathbf{Q} \begin{pmatrix} F_{1,\delta}^{-1}(u_1) \\ \vdots \\ F_{n,\delta}^{-1}(u_n) \end{pmatrix} \ , \tag{8}$$

$$\mathcal{G}^\star : (\delta, (\boldsymbol{v}_i)_{i\in[1..\lambda]}) \in \mathbb{R}_+ \times [0,1]^{n\lambda} \mapsto \underset{\boldsymbol{G}\in\{\mathcal{G}(\delta,\boldsymbol{v}_i)|i\in[1..\lambda]\}}{\operatorname{argmax}} f(\boldsymbol{G}) \ . \tag{9}$$

Then, if the copula C_δ is constant in regard to δ, for $\mathbf{W}_t = (\boldsymbol{V}_{i,t})_{i\in[1..\lambda]}$ a i.i.d. sequence with $\boldsymbol{V}_{i,t} \sim C_\delta$

$$\mathcal{G}(\delta_t, \boldsymbol{V}_{i,t}) \overset{(d)}{=} \boldsymbol{M}_t^i \ , \tag{10}$$

$$\mathcal{G}^\star(\delta_t, \mathbf{W}_t) \overset{(d)}{=} \boldsymbol{M}_t^\star \ . \tag{11}$$

Proof. Since $\boldsymbol{V}_{i,t} \sim \mathbf{C}_\delta$

$$(\boldsymbol{M}_t^i.\nabla g, \boldsymbol{M}_t^i.\nabla g^\perp, \ldots, \boldsymbol{M}_t^i.\boldsymbol{e}_n) \overset{(d)}{=} (F_{1,\delta}^{-1}(\boldsymbol{V}_{1,t}), F_{2,\delta}^{-1}(\boldsymbol{V}_{2,t}), \ldots, F_{n,\delta}^{-1}(\boldsymbol{V}_{n,t})) \ ,$$

and if the function $\delta \in \mathbb{R}_+ \mapsto C_\delta$ is constant, then the sequence of random vectors $(\boldsymbol{V}_{i,t})_{i\in[1..\lambda],t\in\mathbb{N}}$ is i.i.d.. Finally by definition $\mathbf{Q}^{-1}\boldsymbol{M}_t^i = (\boldsymbol{M}_t^i.\nabla g, \boldsymbol{M}_t^i.\nabla g^\perp, \ldots, \boldsymbol{M}_t^i.\boldsymbol{e}_n)$, which shows Eq. (10). Eq. (11) is a direct consequence of Eq. (10) and the fact that $\boldsymbol{M}_t^\star = \underset{\boldsymbol{G}\in\{\mathcal{G}(\delta,\boldsymbol{v}_i)|i\in[1..\lambda]\}}{\operatorname{argmax}} f(\boldsymbol{G})$ (which holds as f is linear). □

We may now use these results to show the divergence of the algorithm when the step-size is constant, using the theory of Markov chains [15].

4 Divergence of the $(1, \lambda)$-ES with constant step-size

Following the first part of [8], we restrict our attention to the constant step size in the remainder of the paper, that is for all $t \in \mathbb{N}$ we take $\sigma_t = \sigma \in \mathbb{R}_+^*$.

From Eq. (4), by recurrence and dividing by t, we see that

$$\frac{[\boldsymbol{X}_t - \boldsymbol{X}_0]_1}{t} = \frac{\sigma}{t} \sum_{i=0}^{t-1} \boldsymbol{M}_i^\star \ . \tag{12}$$

The latter term suggests the use of a Law of Large Numbers to show the convergence of the left hand side to a constant that we call the divergence rate. The random vectors $(\boldsymbol{M}_t^\star)_{t\in\mathbb{N}}$ are not i.i.d. so in order to apply a Law of Large Numbers on the right hand side of the previous equation we use Markov chain theory, more precisely the fact that $(\boldsymbol{M}_t^\star)_{t\in\mathbb{N}}$ is a function of a $(\delta_t, (\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda],j\in\mathbb{N}})_{t\in\mathbb{N}}$ which is a geometrically ergodic Markov chain. As $(\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda],j\in\mathbb{N},t\in\mathbb{N}}$ is a i.i.d. sequence, it is a Markov chain, and the sequence $(\delta_t)_{t\in\mathbb{N}}$ is also a Markov chain as stated in the following proposition.

Proposition 1. *Let a $(1, \lambda)$-ES with constant step-size optimize problem (2), handling the constraint through resampling, and take δ_t as defined in (5). Then no matter what distribution the i.i.d. sequence $(\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda],(j,t)\in\mathbb{N}^2}$ have, $(\delta_t)_{t\in\mathbb{N}}$ is a homogeneous Markov chain and*

$$\delta_{t+1} = \delta_t - g(\boldsymbol{M}_t^\star) = \delta_t - \cos\theta[\boldsymbol{M}_t^\star]_1 - \sin\theta[\boldsymbol{M}_t^\star]_2 \ . \tag{13}$$

Proof. By definition in (5) and since for all t, $\sigma_t = \sigma$,

$$\begin{aligned}\delta_{t+1} &= -\frac{g(\boldsymbol{X}_{t+1})}{\sigma_{t+1}} \\ &= -\frac{g(\boldsymbol{X}_t) + \sigma g(\boldsymbol{M}_t^\star)}{\sigma} \\ &= \delta_t - g(\boldsymbol{M}_t^\star) \ ,\end{aligned}$$

and as shown in (7) the density of $\boldsymbol{M}_t^\star$ is determined by δ_t. So the distribution of δ_{t+1} is determined by δ_t, hence $(\delta_t)_{t\in\mathbb{N}}$ is a time-homogeneous Markov chain. □

We now show ergodicity of the Markov chain $(\delta_t)_{t\in\mathbb{N}}$, which implies that the t-steps transition kernel (the function $A \mapsto \Pr(\delta_t \in A | \delta_0 = \delta)$ for $A \in \mathcal{B}(\mathbb{R}_+)$) converges towards a stationary measure π, generalizing Propositions 3 and 4 of [8].

Proposition 2. *Let a $(1, \lambda)$-ES with constant step-size optimize problem (2), handling the constraint through resampling. We assume that the distribution of $\boldsymbol{M}_t^{i,j}$ is absolutely continuous with probability density function h, and that h is continuous and strictly positive on $\mathbb{R}^n$. Denote μ_+ the Lebesgue measure on $(\mathbb{R}_+, \mathcal{B}(\mathbb{R}_+))$, and for $\alpha > 0$ take the functions $V : \delta \mapsto \delta$, $V_\alpha : \delta \mapsto \exp(\alpha\delta)$ and $r_1 : \delta \mapsto 1$. Then $(\delta_t)_{t\in\mathbb{N}}$ is μ_+-irreducible, aperiodic and compact sets are small sets for the Markov chain.*

If the following two additional conditions are fulfilled

$$\mathbf{E}(|g(\boldsymbol{M}_t^{i,j})| \ | \ \delta_t = \delta) < \infty \text{ for all } \delta \in \mathbb{R}_+ \ , \text{ and} \tag{14}$$

$$\lim_{\delta\to+\infty} \mathbf{E}(g(\boldsymbol{M}_t^\star)|\delta_t = \delta) \in \mathbb{R}_+^* \ , \tag{15}$$

then $(\delta_t)_{t\in\mathbb{N}}$ is r_1-ergodic and positive Harris recurrent with some invariant measure π.

Furthermore, if

$$\mathbf{E}(\exp(g(\boldsymbol{M}_t^{i,j}))|\delta_t = \delta) < \infty \text{ for all } \delta \in \mathbb{R}_+ \ , \tag{16}$$

then for $\alpha > 0$ small enough, $(\delta_t)_{t\in\mathbb{N}}$ is also $V_\alpha-$geometrically ergodic.

Proof. The probability transition kernel of $(\delta_t)_{t\in\mathbb{N}}$ writes

$$\begin{aligned}
P(\delta, A) &= \int_{\mathbb{R}^n} \mathbb{1}_A(\delta - g(\boldsymbol{x}))\tilde{h}^\star_\delta(\boldsymbol{x})d\boldsymbol{x} \\
&= \int_{\mathbb{R}^n} \mathbb{1}_A(\delta - g(\boldsymbol{x}))\lambda \frac{h(\boldsymbol{x})\mathbb{1}_{L_\delta}(\boldsymbol{x})H((-\infty, [\boldsymbol{x}]_1) \times \mathbb{R}^{n-1} \cap L_\delta)^{\lambda-1}}{H(L_\delta)^\lambda} \\
&= \frac{\lambda}{H(L_\delta)^\lambda} \int_{g^{-1}(A)} h \begin{pmatrix} \delta - [\boldsymbol{u}]_1 \\ -[\boldsymbol{u}]_2 \\ \vdots \\ -[\boldsymbol{u}]_n \end{pmatrix} H((-\infty, \delta - [\boldsymbol{u}]_1) \times \mathbb{R}^{n-1} \cap L_\delta)^{\lambda-1} d\boldsymbol{u} \ ,
\end{aligned}$$

with the substitution of variables $[\boldsymbol{u}]_1 = \delta - [\boldsymbol{x}]_1$ and $[\boldsymbol{u}]_i = -[\boldsymbol{x}]_i$ for $i \in [2..n]$. Denote $L^\star_{\delta,v} := (-\infty, v) \times \mathbb{R}^{n-1} \cap L_\delta$ and $t_\delta : \boldsymbol{u} \mapsto (\delta - [\boldsymbol{u}]_1, -[\boldsymbol{u}]_2, \ldots, -[\boldsymbol{u}]_n)$, take C a compact of $\mathbb{R}_+$, and define ν_C such that for $A \in \mathcal{B}(\mathbb{R}_+)$

$$\nu_C(A) := \lambda \int_{g^{-1}(A)} \inf_{\delta\in C} \frac{h(t_\delta(\boldsymbol{u}))H(L^\star_{\delta,[\boldsymbol{u}]_1})^{\lambda-1}}{H(L_\delta)^\lambda} d\boldsymbol{u} \ .$$

As the density h is supposed to be strictly positive on $\mathbb{R}^n$, for all $\delta \in \mathbb{R}_+$ we have $H(L_\delta) \geq H(L_0) > 0$. Using the fact that H is a finite measure, and is absolutely continuous, applying the dominated convergence theorem shows that the functions $\delta \mapsto H(L_\delta)$ and $\delta \mapsto H((-\infty, \delta - [\boldsymbol{u}]_1) \times \mathbb{R}^{n-1} \cap L_\delta)$ are continuous. Therefore the function $\delta \mapsto h(t_\delta(\boldsymbol{u}))H(L^\star_{\delta,[\boldsymbol{u}]_1})^{\lambda-1}/H(L_\delta)^\lambda$ is continuous and C being a compact, the infimum of this function is reached on C is reached on C. Since this function is strictly positive, if $g^{-1}(A)$ has strictly positive Lebesgue measure then $\nu_C(A) > 0$ which proves that this measure is not trivial. By construction $P(\delta, A) \geq \nu_C(A)$ for all $\delta \in C$, so C is a small set which shows that compact sets are small. Since if $\mu_+(A) > 0$ we have $P(\delta, A) \geq \nu_C(A) > 0$, the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is μ_+-irreducible. Finally, if we take C a compact set of $\mathbb{R}_+$ with strictly positive Lebesgue measure, then it is a small set and $\nu_C(C) > 0$ which means the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is strongly aperiodic.

The function ΔV is defined as $\delta mapsto \mathbf{E}(V(\delta_{t+1})|\delta_t = \delta) - V(\delta)$. We want to show a drift condition (see [15]) on V. Using Eq. (13)

$$\begin{aligned}
\Delta V(\delta) &= \mathbf{E}(\delta - g(\boldsymbol{M}^\star_t)|\delta_t = \delta) - \delta) \\
&= -\mathbf{E}(g(\boldsymbol{M}^\star_t)) \ .
\end{aligned}$$

Therefore using the condition (15), we have that there exists a $\epsilon > 0$ and a $M \in \mathbb{R}_+$ such that $\forall \delta \in (M, +\infty)$, $\Delta V(\delta) \leq -\epsilon$. With condtion (14) implies that the function $\Delta V + \epsilon$ is bounded on the compact $[0, M]$ by a constant $b \in \mathbb{R}$. Hence for all $\delta \in \mathbb{R}_+$

$$\frac{\Delta V(\delta)}{\epsilon} \leq -1 + \frac{b}{\epsilon}\mathbb{1}_{[0,M]}(\delta) \ . \tag{17}$$

For all $x \in \mathbb{R}$ the level set $C_{V,x}$ of the function V, $\{y \in \mathbb{R}_+ | V(y) \leq x\}$, is equal to $[0, x]$ which is a compact set, hence a small set according to what we proved

earlier (and hence petite [15, Proposition 5.5.3]). Therefore V is unbounded off small sets and with (17) and Theorem 9.1.8 of [15], the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is Harris recurrent. The set $[0, M]$ is compact and therefore small and petite, so with (17), if we denote r_1 the constant function $\delta \in \mathbb{R}_+ \mapsto 1$ then with Theorem 14.0.1 of [15] the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is positive and is r_1-ergodic.

We now want to show a drift condition (see [15]) on V_α.

$$\begin{aligned}
\Delta V_\alpha(\delta) &= \mathbf{E}\left(\exp\left(\alpha\delta - \alpha g\left(\boldsymbol{M}_t^\star\right)\right)|\delta_t = \delta\right) - \exp\left(\alpha\delta\right) \\
\frac{\Delta V_\alpha}{V_\alpha}(\delta) &= \mathbf{E}\left(\exp\left(-\alpha g\left(\boldsymbol{M}_t^\star\right)\right)|\delta_t = \delta\right) - 1 \\
&= \int_{\mathbb{R}^n} \lim_{t\to+\infty} \sum_{k=0}^{t} \frac{(-\alpha g(\boldsymbol{x}))^k}{k!} \tilde{h}_\delta^\star(\boldsymbol{x}) d\boldsymbol{x} - 1 \ .
\end{aligned}$$

With Eq. (7) we see that $\tilde{h}_\delta^\star(\boldsymbol{x}) \leq \lambda h(\boldsymbol{x})/H(L_0)^\lambda$, so with our assumption that $\mathbf{E}(\exp\alpha|g(\boldsymbol{M}_t^{i,j})||\delta_t = \delta) < \infty$ for $\alpha > 0$ small enough we have that the function $\delta \mapsto \mathbf{E}(\exp(\alpha|g(\boldsymbol{M}_t^\star)||\delta_t = \delta)$ is bounded for the same α. As $\sum_{k=0}^{t}(-\alpha g(\boldsymbol{x}))^k/k!\tilde{h}_\delta^\star(\boldsymbol{x}) \leq \exp(\alpha|g(\boldsymbol{x})|)\tilde{h}_\delta^\star(\boldsymbol{x})$ which, with condition (16), is integrable so we may apply the theorem of dominated convergence to invert limit and integral:

$$\begin{aligned}
\frac{\Delta V_\alpha}{V_\alpha}(\delta) &= \lim_{t\to+\infty} \sum_{k=0}^{t} \int_{\mathbb{R}^n} \frac{(-\alpha g(\boldsymbol{x}))^k}{k!} \tilde{h}_\delta^\star(\boldsymbol{x}) d\boldsymbol{x} - 1 \\
&= \sum_{k\in\mathbb{N}} (-\alpha)^k \frac{\mathbf{E}\left(g\left(\boldsymbol{M}_t^\star\right)^k|\delta_t = \delta\right)}{k!} - 1
\end{aligned}$$

Since $\tilde{h}_\delta^\star(\boldsymbol{x}) \leq \lambda h(\boldsymbol{x})/H(L_0)^2$, $(-\alpha)^k\mathbf{E}(g(\boldsymbol{M}_t^\star)^k|\delta_t = \delta)/k! \leq (-\alpha)^k\mathbf{E}(g(\boldsymbol{M}_t^{i,j})^k)/k!$ which is integrable with respect to the counting measure so we may apply the dominated convergence theorem with the counting measure to invert limit and serie.

$$\begin{aligned}
\lim_{\delta\to+\infty} \frac{\Delta V_\alpha}{V_\alpha}(\delta) &= \sum_{k\in\mathbb{N}} \lim_{\delta\to+\infty} (-\alpha)^k \frac{\mathbf{E}\left(g\left(\boldsymbol{M}_t^\star\right)^k|\delta_t = \delta\right)}{k!} - 1 \\
&= -\alpha \lim_{\delta\to+\infty} \mathbf{E}\left(g\left(\boldsymbol{M}_t^\star\right)|\delta_t = \delta\right) + o\left(\alpha\right) \ .
\end{aligned}$$

With condition (17) we supposed that $\lim_{\delta\to+\infty}\mathbf{E}(g(\boldsymbol{M}_t^\star)|\delta_t = \delta) > 0$ this implies that for $\alpha > 0$ and small enough, $lim_{\delta\to+\infty}\Delta V_\alpha(\delta)/V_\alpha(\delta) < 0$, hence there exists $M \in \mathbb{R}_+$ and $epsilon > 0$ such that $\forall\delta > M$, $\Delta V_\alpha(\delta) < -\epsilon V_\alpha(\delta)$. Finally as $\Delta V_\alpha - V_\alpha$ is bounded on $[0, M]$ there exists $b \in \mathbb{R}$ such that

$$\Delta V_\alpha(\delta) \leq -\epsilon V_\alpha(\delta) + b\mathbb{1}_{[0,M]}(\delta) \ .$$

According to what we did before in this proof, the compact set $[0, M]$ is small, and hence is petite ([15, Proposition 5.5.3]). So the μ_+-irreducible Markov chain

$(\delta_t)_{t\in\mathbb{N}}$ satisfies the conditions of Theorem 15.0.1 of [15] which with Theorem 14.0.1 of [15] proves that the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is V_α-geometrically ergodic. □

We now use a law of large numbers ([15] Theorem 17.0.1) on the Markov chain $(\delta_t, (\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda],j\in\mathbb{N}})_{t\in\mathbb{N}}$ to obtain an almost sure divergence of the algorithm.

Proposition 3. *Let a $(1,\lambda)$-ES optimize problem (2), handling the constraint through resampling. Assume that the distribution H of the random step $\boldsymbol{M}_t^{i,j}$ is absolutely continuous with continuous and strictly positive density h, that conditions (16) and (15) of Proposition 2 hold, and denote π and μ_M the stationary distribution of respectively $(\delta_t)_{t\in\mathbb{N}}$ and $(\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda],(j,t)\in\mathbb{N}^2}$. Then*

$$\frac{[\boldsymbol{X}_t - \boldsymbol{X}_0]_1}{t} \xrightarrow[t\to+\infty]{a.s.} \sigma\mathbf{E}_{\pi\times\mu_M}([\boldsymbol{M}_t^\star]_1) \ . \tag{18}$$

Furthermore if $\mathbf{E}([\boldsymbol{M}_t^\star]_2) < 0$, then the right hand side of Eq. (18) is strictly positive.

Proof. According to Proposition 2 the sequence $(\delta_t)_{t\in\mathbb{N}}$ is a Harris recurrent positive Markov chain with invariant measure π. As $(\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda],(j,t)\in\mathbb{N}^2}$ is a i.i.d. sequence with distribution μ_M, $(\delta_t, (\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda],j\in\mathbb{N}})_{t\in\mathbb{N}}$ is also a Harris recurrent positive Markov chain. As $[\boldsymbol{M}_t^\star]_1$ is a function of δ_t and $(\boldsymbol{M}_t^{i,j})_{i\in[1..\lambda],j\in\mathbb{N}}$, if $\mathbf{E}_{\pi\times\mu_M}(|[\boldsymbol{M}_t^\star]_1|) < \infty$, according to Theorem 17.0.1 of [15], we may apply a law of large numbers on the right hand side of Eq. (12) to obtain (18).

Using Fubini-Tonelli's theorem $\mathbf{E}_{\pi\times\mu_M}(|[\boldsymbol{M}_t^\star]_1|) = \mathbf{E}_\pi(\mathbf{E}_{\mu_M}(|[\boldsymbol{M}_t^\star]_1||\delta_t = \delta))$. From Eq. (7) for all $\boldsymbol{x} \in \mathbb{R}^n$, $\tilde{h}_\delta^\star(\boldsymbol{x}) \leq \lambda h(\boldsymbol{x})/H(L_0)^2$, so the condition in (16) implies that for all $\delta \in \mathbb{R}_+$, $\mathbf{E}_{\mu_M}(|[\boldsymbol{M}_t^\star]_1||\delta_t = \delta)$ is finite. Furthermore, with condition (15), the function $\delta \in \mathbb{R}_+ \mapsto \mathbf{E}_{\mu_M}(|[\boldsymbol{M}_t^\star]_1||\delta_t = \delta)$ is bounded by some $M \in \mathbb{R}$. Therefore as π is a probability measure, $\mathbf{E}_\pi(\mathbf{E}_{\mu_M}(|[\boldsymbol{M}_t^\star]_1||\delta_t = \delta)) \leq M < \infty$ so we may apply the law of large numbers of Theorem 17.0.1 of [15].

Using the fact that π is an invariant measure, we have $\mathbf{E}_\pi(\delta_t) = \mathbf{E}_\pi(\delta_{t+1})$, so $\mathbf{E}_\pi(\delta_t) = \mathbf{E}_\pi(\delta_t - \sigma g(\boldsymbol{M}_t^\star))$ and hence $\cos\theta\mathbf{E}_\pi([\boldsymbol{M}_t^\star]_1) = -\sin\theta\mathbf{E}_\pi([\boldsymbol{M}_t^\star]_2)$. So using the assumption that $\mathbf{E}([\boldsymbol{M}_t^{i,j}]_2) \leq 0$ then we get the strict positivity of $\mathbf{E}_{\pi\times\mu_M}([\boldsymbol{M}_t^{i,j}]_1)$. □

5 Application to More Specific Distributions

Throughout this section we give cases where the assumptions on the distribution of the random steps H used in Proposition 2 or Proposition 3 are verified.

The following lemma shows an equivalence between a non-identity covariance matrix for H and a different norm and constraint angle θ.

Lemma 3. *Let a $(1,\lambda)$-ES optimize problem (2), handling the constraint with resampling. Assume that the distribution H of the random step $\boldsymbol{M}_t^{i,j}$ has positive definite covariance matrix $\mathbf{C}$ with eigenvalues $(\alpha_i^2)_{i\in[1..n]}$ and take $\boldsymbol{B} =$*

$(b_{i,j})_{(i,j)\in[1..n]^2}$ *such that* $\boldsymbol{B}\mathbf{C}\boldsymbol{B}^{-1}$ *is diagonal. Denote* $\mathcal{A}_{H,g,\boldsymbol{X}_0}$ *the sequence of parent points* $(\boldsymbol{X}_t)_{t\in\mathbb{N}}$ *of the algorithm with distribution* H *for the random steps* $\boldsymbol{M}_t^{i,j}$*, constraint angle* θ *and initial parent* $\boldsymbol{X}_0$*. Then for all* $k \in [1..n]$

$$\beta_k \left[\mathcal{A}_{H,\theta,\boldsymbol{X}_0}\right]_k \overset{(d)}{=} \left[\mathcal{A}_{\mathbf{C}^{-1/2}H,\theta',\boldsymbol{X}'_0}\right]_k \quad , \tag{19}$$

where $\beta_k = \sqrt{\sum_{j=1}^n \frac{b_{j,i}^2}{\alpha_i^2}}$*,* $\theta' = \arccos(\frac{\beta_1 cos\theta}{\beta_g})$ *with* $\beta_g = \sqrt{\beta_1^2\cos^2\theta + \beta_2^2\sin^2\theta}$*, and* $[\boldsymbol{X}'_0]_k = \beta_k[\boldsymbol{X}_0]_k$ *for all* $k \in [1..n]$*.*

Proof. Take $(\bar{\boldsymbol{e}}_k)_{k\in[1..n]}$ the image of $(\boldsymbol{e}_k)_{k\in[1..n]}$ by $\boldsymbol{B}^{-1}$. We define a new norm $\|\cdot\|_-$ such that $\|\bar{\boldsymbol{e}}_k\|_- = 1/\alpha_k$. We define two orthonormal basis $(\boldsymbol{e}'_k)_{k\in[1..n]}$ and $(\bar{\boldsymbol{e}}'_k)_{k\in[1..n]}$ for $(\mathbb{R}^n, \|\cdot\|_-)$ by taking $\boldsymbol{e}'_k = \boldsymbol{e}_k/\|\boldsymbol{e}_k\|_-$ and $\bar{\boldsymbol{e}}'_k = \bar{\boldsymbol{e}}_k/\|\bar{\boldsymbol{e}}_k\|_- = \alpha_k\bar{\boldsymbol{e}}_k$. As $\mathrm{Var}(\boldsymbol{M}_t^{i,j}.\bar{\boldsymbol{e}}_k) = \alpha_k^2$, $\mathrm{Var}(\boldsymbol{M}_t^{i,j}.\bar{\boldsymbol{e}}'_k) = 1$ so in $(\mathbb{R}^n, \|\cdot\|_-)$ the covariance matrix of $\boldsymbol{M}_t^{i,j}$ is the identity.

Take h the function that to $\boldsymbol{x} \in \mathbb{R}^n$ maps its image in the new orthonormal basis $(\boldsymbol{e}'_k)_{k\in[1..n]}$. As $\boldsymbol{e}'_k = \boldsymbol{e}_k/\|\boldsymbol{e}_k\|_-$, $h(\boldsymbol{x}) = (\|\boldsymbol{e}_k\|_-[\boldsymbol{x}]_k)_{k\in[1..n]}$, where $\|\boldsymbol{e}_k\|_- = \|\sum_{i=1}^n b_{i,k}\bar{\boldsymbol{e}}_k\|_- = \sqrt{\sum_{i=1}^n b_{i,k}^2/\alpha_k^2} = \beta_k$. As we changed the norm, the angle between ∇f and ∇g is also different in the new space. Indeed $\cos\theta' = h(\nabla g).h(\nabla f)/(\|h(\nabla g)\|_-\|h(\nabla f)\|_-) = \beta_1^2\cos\theta/(\sqrt{\beta_1^2\cos^2\theta + \beta_2^2\sin^2\theta}\beta_1)$ which equals $\beta_1\cos\theta/\beta_g$.

If we take $\boldsymbol{N}_t^{i,j} \sim \mathbf{C}^{-1/2}H$ then it has the same distribution as $h(\boldsymbol{M}_t^{i,j})$. Take $\boldsymbol{X}'_t = h(\boldsymbol{X}_t)$ then for a constraint angle $\theta' = \arccos(\beta_1\cos\theta/\beta_g)$ and a normalized distance to the constraint $\delta_t = \boldsymbol{X}'_t.h(\nabla g)/\sigma_t$ the ressampling is the same for $\boldsymbol{N}_t^{i,j}$ and $h(\boldsymbol{M}_t^{i,j})$ so $\boldsymbol{N}_t^i \overset{(d)}{=} h(\boldsymbol{M}_t^i)$. Finally the rankings induced by ∇f or $h(\nabla f)$ are the same so the selection in the same, hence $\boldsymbol{N}_t^\star \overset{(d)}{=} h(\boldsymbol{M}_t^\star)$, and therefore $\boldsymbol{X}'_{t+1} \overset{(d)}{=} h(\boldsymbol{X}_{t+1})$. □

Although Eq. (18) shows divergence of the algorithm, it is important that it diverges in the right direction, i.e. that the right hand side of Eq. (18) has a positive sign. This is achieved when the distribution of the random steps is isotropic, as stated in the following proposition.

Proposition 4. *Let a* $(1,\lambda)$*-ES optimize problem* (2) *with constant step-size, handling the constraint with resampling. Suppose that the Markov chain* $(\delta_t)_{t\in\mathbb{N}}$ *is positive Harris, that the distribution* H *of the random step* $\boldsymbol{M}_t^{i,j}$ *is absolutely continuous with strictly positive density* h*, and take* $\mathbf{C}$ *its covariance matrix. If the distribution* $\mathbf{C}^{-1/2}H$ *is isotropic then* $\mathbf{E}_{\pi\times\mu_M}([\boldsymbol{M}_t^\star]_1) > 0$*.*

Proof. First if $\mathbf{C} = \mathbf{I}_n$, using the same method than in the proof of Lemma 1

$$h^\star_{\delta,2}(y) = \lambda\int_{\mathbb{R}}\dots\int_{\mathbb{R}}\tilde{h}_\delta(u_1, y, u_3, \dots, u_n)\Pr(u_1 \geq [\boldsymbol{M}_t^i]_1)^{\lambda-1}du_1\prod_{k=3}^n du_k \quad .$$

Using Eq.(6) and the fact that the condition $\boldsymbol{x} \in L_\delta$ is equivalent to $[\boldsymbol{x}]_1 \leq (\delta - [\boldsymbol{x}]_2 \sin\theta)/\cos\theta$ we obtain

$$h^\star_{\delta,2}(y) = \lambda \int_{\mathbb{R}} \dots \int_{-\infty}^{\frac{\delta - y\sin\theta}{\cos\theta}} \frac{h(u_1, y, u_3, \dots, u_n)}{H(L_\delta)} \Pr(u_1 \geq [\boldsymbol{M}^i_t]_1)^{\lambda-1} du_1 \prod_{k=3}^{n} du_k \ .$$

If the distribution of the random steps steps is isotropic then $h(u_1, y, u_3, \dots, u_n) = h(u_1, -y, u_3, \dots, u_n)$, and as the density h is supposed strictly positive, for $y > 0$ and all $\delta \in$, $h^\star_{\delta,2}(y) - h^\star_{\delta,2}(-y) < 0$ so $\mathbf{E}([\boldsymbol{M}^\star_t]_2 | \delta_t = \delta) < 0$. If the Markov chain is Harris recurrent and positive then this imply that $\mathbf{E}_\pi([\boldsymbol{M}^\star_t]_2) < 0$ and using the reasoning in the proof of Proposition 3 $\mathbf{E}_\pi([\boldsymbol{M}^\star_t]_1) > 0$.

For any covariance matrix $\mathbf{C}$ this result is generalized with the use of Lemma 3. □

Lemma 3 and Proposition 4 imply the following result to hold for multivariate normal distributions.

Proposition 5. *Let a $(1, \lambda)$-ES optimize problem (2) with constant step-size, handling the constraint with resampling. If H is a multivariate normal distribution with mean $\mathbf{0}$, then $(\delta_t)_{t\in\mathbb{N}}$ is a geometrically ergodic positive Harris Markov chain, Eq. (18) holds and its right hand side is strictly positive.*

Proof. Suppose $\boldsymbol{M}^{i,j}_t \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_n)$. Then H is absolutely continuous and h is strictly positive. The function $\boldsymbol{x} \mapsto \exp(g(\boldsymbol{x}))\exp(-\|\boldsymbol{x}\|^2/2)/\sqrt{2\pi}$ is integrable, so Eq. (16) is satisfied. Furthermore, when $\delta \to +\infty$ the constraint disappear so $\boldsymbol{M}^{i,j}_t$ behaves like $(\mathcal{N}_{\lambda:\lambda}, \mathcal{N}(0,1), \dots, \mathcal{N}(0,1))$ where $\mathcal{N}_{\lambda:\lambda}$ is the last order statistic of λ i.i.d. standard normal variables, so using that $\mathbf{E}(\mathcal{N}_{\lambda:\lambda}) > 0$ and $\mathbf{E}(\mathcal{N}(0,1)) = 0$, with multiple uses of the dominated convergence theorem we obtain condition (15) so with Proposition 2 the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is geometrically ergodic and positive Harris.

Finally H being isotropic the conditions of Proposition 4 are fulfilled, and therefore so are every condition of Proposition 3 which shows what we wanted. □

To obtain sufficient conditions for the density of the random steps to be strictly positive, it is advantageous to decompose that distribution into its marginals and the copula combining them. We pay a particular attention to *Archimedean copulas,* i.e., copulas defined

$$(\forall \boldsymbol{u} \in [0,1]^n)\ C_\psi(\boldsymbol{u}) = \psi(\psi^{-1}([\boldsymbol{u}]_1) + \dots + \psi^{-1}([\boldsymbol{u}]_n)), \tag{20}$$

where $\psi : [0, +\infty] \to [0, 1]$ is an *Archimedean generator*, i.e., $\psi(0) = 1, \psi(+\infty) = \lim_{t\to+\infty} \psi(t) = 0$, ψ is continuous and strictly decreasing on $[0, \inf\{t : \psi(t) = 0\})$, and ψ^{-1} denotes the generalized inverse of ψ,

$$(\forall u \in [0,1])\ \psi^{-1}(u) = \inf\{t \in [0, +\infty] : \psi(t) = u\}. \tag{21}$$

The reason for our interest is that Archimedean copulas are invariant with respect to permutations of variables, i.e.,

$$(\forall \boldsymbol{u} \in [0,1]^n)\ C_\psi(\mathbf{Q}\boldsymbol{u}) = C_\psi(\boldsymbol{u}). \tag{22}$$

holds for any permutation matrix $\mathbf{Q} \in \mathbb{R}^{n,n}$. This can be seen as a weak form of isotropy because in the case of isotropy, (20) holds for any rotation matrix, and a permutation matrix is a specific rotation matrix.

Proposition 6. *Let H be the distribution of the two first dimensions of the random step $\boldsymbol{M}_t^{i,j}$, H_1 and H_2 be its marginals, and C be the copula relating H to H_1 and H_2. Then the following holds:*

1. *Sufficient for H to have a continuous strictly positive density is the simultaneous validity of the following two conditions.*
   (i) *H_1 and H_2 have continuous strictly positive densities h_1 and h_2, respectively.*
   (ii) *C has a continuous strictly positive density c.*
   Moreover, if (i) and (ii) are valid, then

$$(\forall \boldsymbol{x} \in \mathbb{R}^2)\ h(\boldsymbol{x}) = c(H_1([\boldsymbol{x}]_1), H_2([\boldsymbol{x}]_2))h_1([\boldsymbol{x}]_1)h_2([\boldsymbol{x}]_2). \tag{23}$$

2. *If C is Archimedean with generator ψ, then it is sufficient to replace (ii) with*
   (ii') ψ is at least 4-monotone, i.e., ψ is continuous on $[0,+\infty]$, ψ'' is decreasing and convex on $\mathbb{R}_+$, and $(\forall t \in \mathbb{R}_+)\ (-1)^k\psi^{(k)}(t) \geq 0, k = 0,1,2$.
   In this case, if (i) and (ii') are valid, then

$$(\forall \boldsymbol{x} \in \mathbb{R}^2)\ h(\boldsymbol{x}) = \frac{\psi''(\psi^{-1}(H_1([\boldsymbol{x}]_1)) + \psi^{-1}(H_2([\boldsymbol{x}]_2)))}{\psi'(\psi^{-1}(H_1([\boldsymbol{x}]_1)) + \psi^{-1}(H_2([\boldsymbol{x}]_2)))} h_1([\boldsymbol{x}]_1)h_2([\boldsymbol{x}]_2). \tag{24}$$

Proof. The continuity and strict positivity of the density of H is a straightforward consequence of the conditions (i) and (ii), respectively (ii'). In addition, the assumption that ψ is at least 4-monotone implies that it is also 2-monotone, which is for the function C_ψ in (20) with $n = 2$ a necessary and sufficient condition to be indeed a copula [16]. To prove (23), the relationships

$$(\forall \boldsymbol{x} \in \mathbb{R}^2)\ h(\boldsymbol{x}) = \frac{\partial^2 H}{\partial[\boldsymbol{x}]_1 \partial[\boldsymbol{x}]_2}(\boldsymbol{x}), h_1([\boldsymbol{x}]_1) = \frac{H_1}{d[\boldsymbol{x}]_1}([\boldsymbol{x}]_1), h_2([\boldsymbol{x}]_2) = \frac{H_2}{d[\boldsymbol{x}]_2}([x]_2), \tag{25}$$

are combined with the Sklar's theorem ([17], cf. also [18])

$$(\forall \boldsymbol{x} \in \mathbb{R}^2)\ H(\boldsymbol{x}) = C(H_1([\boldsymbol{x}]_1), H_2([\boldsymbol{x}]_2)) \tag{26}$$

and with

$$c(\boldsymbol{u}) = \frac{\partial^2 C}{\partial[\boldsymbol{u}]_1 \partial[\boldsymbol{u}]_2}(\boldsymbol{u}). \tag{27}$$

For Archimedean copulas, combining (27) with (20) turns (23) into (24). □

6 Discussion

The paper presents a generalization of recent results of the first author [8] concerning linear optimization by a $(1, \lambda)$-ES in the constant step size case. The generalization consists in replacing the assumption of normality of random steps involved in the evolution strategy by substantially more general distributional assumptions. This generalization shows that isotropic distributions solve the linear problem. Also, although the conditions for the ergodicity of the studied Markov chain accept some heavy-tail distributions, an exponentially vanishing tail allow for geometric ergodicity, which imply a faster convergence to its stationary distribution, and faster convergence of Monte Carlo simulations. In our opinion, these conditions increase the insight into the role that different kinds of distributions play in evolutionary computation, and enlarges the spectrum of possibilities for designing evolutionary algorithms with solid theoretical fundamentals. At the same time, applying the decomposition of a multidimensional distribution into its marginals and the copula combining them, the paper attempts to bring a small contribution to the research into applicability of copulas in evolutionary computation, complementing the more common application of copulas to the Estimation of Distribution Algorithms [12, 14, 13].

Needless to say, more realistic than the constant step size case, but also more difficult to investigate, is the varying step size case. The most important results in [8] actually concern that case. A generalization of those results for non-Gaussian distributions of random steps for cumulative step-size adaptation ([9]) is especially difficult as the evolution path is tailored for Gaussian steps, and some careful tweaking would have to be applied. The σ self-adaptation evolution strategy ([19]), studied in [6] for the same problem, appears easier, and would be our direction for future research.

Acknowledgment

The research reported in this paper has been supported by grant ANR-2010-COSI-002 (SIMINOLE) of the French National Research Agency, and Czech Science Foundation (GAČR) grant 13-17187S.

References

1. X. Yao and Y. Liu, "Fast evolution strategies," in *Evolutionary Programming VI*, pp. 149–161, Springer, 1997.
2. T. Schaul, "Benchmarking Separable Natural Evolution Strategies on the Noiseless and Noisy Black-box Optimization Testbeds," in *Black-box Optimization Benchmarking Workshop, Genetic and Evolutionary Computation Conference*, (Philadelphia, PA), 2012.
3. T. Schaul, T. Glasmachers, and J. Schmidhuber, "High dimensions and heavy tails for natural evolution strategies," in *Genetic and Evolutionary Computation Conference (GECCO)*, 2011.

4. N. Hansen, F. Gemperle, A. Auger, and P. Koumoutsakos, "When do heavy-tail distributions help?," in *Parallel Problem Solving from Nature PPSN IX* (T. P. Runarsson *et al.*, eds.), vol. 4193 of *Lecture Notes in Computer Science*, pp. 62–71, Springer, 2006.
5. D. Arnold, "On the behaviour of the (1,λ)-ES for a simple constrained problem," in *Foundations of Genetic Algorithms - FOGA 11*, pp. 15–24, ACM, 2011.
6. D. Arnold, "On the behaviour of the $(1, \lambda)$-σSA-ES for a constrained linear problem," in *Parallel Problem Solving from Nature - PPSN XII*, pp. 82–91, Springer, 2012.
7. A. Chotard, A. Auger, and N. Hansen, "Cumulative step-size adaptation on linear functions," in *Parallel Problem Solving from Nature - PPSN XII*, pp. 72–81, Springer, september 2012.
8. A. Chotard, A. Auger, and N. Hansen, "Markov chain analysis of evolution strategies on a linear constraint optimization problem," in *IEEE Congress on Evolutionary Computation (CEC)*, 2014.
9. N. Hansen and A. Ostermeier, "Completely derandomized self-adaptation in evolution strategies," *Evolutionary Computation*, vol. 9, no. 2, pp. 159–195, 2001.
10. C. A. Coello Coello, "Constraint-handling techniques used with evolutionary algorithms," in *Proceedings of the 2008 GECCO conference companion on Genetic and evolutionary computation*, GECCO '08, (New York, NY, USA), pp. 2445–2466, ACM, 2008.
11. D. Arnold and D. Brauer, "On the behaviour of the $(1 + 1)$-ES for a simple constrained problem," in *Parallel Problem Solving from Nature - PPSN X* (I. G. R. et al., ed.), pp. 1–10, Springer, 2008.
12. A. Cuesta-Infante, R. Santana, J. Hidalgo, C. Bielza, and P. Larrañaga, "Bivariate empirical and n-variate archimedean copulas in estimation of distribution algorithms," in *IEEE Congress on Evolutionary Computation*, pp. 1–8, 2010.
13. L. Wang, X. Guo, J. Zeng, and Y. Hong, "Copula estimation of distribution algorithms based on exchangeable archimedean copula," *International Journal of Computer Applications in Technology*, vol. 43, pp. 13–20, 2012.
14. R. Salinas-Gutiérrez, A. Hernández-Aguirre, and E. R. Villa-Diharce, "Using copulas in estimation of distribution algorithms," in *MICAI 2009: Advances in Artificial Intelligence*, pp. 658–668, Springer, 2009.
15. S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*. Cambridge University Press, second ed., 1993.
16. A. McNeil and J. Nešlehová, "Multivariate Archimedean copulas, d-monotone functions and l_1-norm symmetric distributions," *The Annals of Statistics*, vol. 37, pp. 3059–3097, 2009.
17. A. Sklar, "Fonctions de répartition à n dimensions et leurs marges," *Publications de l'Institut de Statistique de l'Université de Paris*, vol. 8, pp. 229–231, 1959.
18. R. Nelsen, *An Introduction to Copulas*. 2006.
19. H.-G. Beyer, "Toward a theory of evolution strategies: Self-adaptation," *Evolutionary Computation*, vol. 3, no. 3, pp. 311–347, 1995.

Chapter 5

Summary, Discussion and Perspectives

The context of this thesis is the study of Evolution Strategies (ESs) using tools from the theory of Markov chains. This work is composed of two parts. The first part focuses on adapting specific techniques from the theory of Markov chains to a general non-linear state space model that encompasses in particular the models that appear in the ES context, allowing us to easily prove some Markov chain properties that we could not prove before. In the second part, we study the behaviour of ESs on the linear function with and without a linear constraint. In particular, log-linear divergence or convergence of the ESs is shown.
In Section 5.1 we give a summary of the contributions of this thesis. Then in Section 5.2 we propose different possible extensions to our contributions.

5.1 Summary and Discussion

5.1.1 Sufficient conditions for the φ-irreducibility, aperiodicity and T-chain property of a general Markov chain

In Chapter 3 we showed that we can adapt the results of [98, Chapter 7] to a more general model

$$\Phi_{t+1} = F(\Phi_t, \alpha(\Phi_t, \boldsymbol{U}_{t+1})) \tag{5.1}$$

where $F : X \times O \to X$ is a measurable function that we call a transition function, $\alpha : X \times \Omega \to O$ is a (typically discontinuous) measurable function and $(\boldsymbol{U}_t)_{t\in\mathbb{N}^*}$ are i.i.d. random variables valued in Ω, but the random elements $\boldsymbol{W}_{t+1} = \alpha(\Phi_t, \boldsymbol{U}_{t+1})$ are not necessarily i.i.d., and X, Ω and O are open sets of respectively $\mathbb{R}^n$, R^p and $\mathbb{R}^m$. We derive for this model easily verifiable conditions to show that a Markov chain is a φ-irreducible aperiodic T-chain and that compact sets are small sets for the Markov chain. These conditions are

- the transition function F is C^1
- for all $\boldsymbol{x} \in X$ the random variable $\alpha(\boldsymbol{x}, \boldsymbol{U}_1)$ admits a density $p_{\boldsymbol{x}}$,
- the function $(\boldsymbol{x}, \boldsymbol{w}) \mapsto p_{\boldsymbol{x}}(\boldsymbol{w})$ is lower semi-continuous,

- there exists $\boldsymbol{x}^* \in X$ a strongly globally attracting state, $k \in \mathbb{N}^*$ and $\boldsymbol{w}^* \in O_{\boldsymbol{x}^*,k}$ such that $F^k(\boldsymbol{x}^*, \cdot)$ is a submersion at $\boldsymbol{w}^*$.

The set $O_{\boldsymbol{x}^*,k}$ is the support of the conditional density of $(\boldsymbol{W}_t)_{t\in[1..k]}$ knowing that $\Phi_0 = \boldsymbol{x}^*$; F^k is the k-steps transition function inductively defined by $F^1 := F$ and $F^{t+1}(\boldsymbol{x}, \boldsymbol{w}_1, \ldots, \boldsymbol{w}_{t+1}) := F^t(F(\boldsymbol{x}, \boldsymbol{w}_1), \boldsymbol{w}_2, \ldots, \boldsymbol{w}_{t+1})$; and the concept of strongly globally attracting states is introduced in Chapter 3, namely that $\boldsymbol{x}^* \in X$ is called a strongly globally attracting state if for all $\boldsymbol{y} \in X$ and $\epsilon > 0$ there exists $t_{\boldsymbol{y},\epsilon} \in \mathbb{N}^*$ such that for all $t \geq t_{\boldsymbol{y},\epsilon}$ there exists a $\boldsymbol{w} \in O_{\boldsymbol{y},t}$ such that $F^t(\boldsymbol{y}, \boldsymbol{w}) \in B(\boldsymbol{x}^*, \epsilon)$. We then used these results to show the φ-irreducibility, aperiodicity, T-chain property and that compact sets are small sets for Markov chains underlying the so-called xNES algorithm [59] with identity covariance matrix on scaling invariant functions, or in the CSA algorithm on a linear constrained problem with the cumulation parameter c_σ equal to 1, which were problems we could not solve before these results.

5.1.2 Analysis of Evolution Strategies using the theory of Markov chains

In Section 4.2 we presented an analysis of the $(1, \lambda)$-CSA-ES on a linear function. The analysis shows the geometric ergodicity of an underlying Markov chain, from which it is deduced that the step-size of the $(1, \lambda)$-CSA-ES diverges log-linearly almost surely for $\lambda \geq 3$, or for $\lambda = 2$ and with the cumulation parameter $c_\sigma < 1$. When $\lambda = 2$ and $c_\sigma = 1$, the sequence $(\ln(\sigma_t))_{t\in\mathbb{N}}$ is an unbiased random walk. It was also shown in the simpler case of $c_\sigma = 1$ that the sequence of $|f|$-value of the mean of the sampling distribution, $(|f(\boldsymbol{X}_t)|)_{t\in\mathbb{N}}$, diverges log-linearly almost surely when $\lambda \geq 3$ at the same rate than the step-size. An expression of the divergence rate is derived, which explicitly gives the influence of the dimension of the search space and of the cumulation parameter c_σ on the divergence rate. The geometric ergodicity also shows the convergence of Monte Carlo simulations to estimate the divergence rate of the algorithm (and the fact that the ergodicity is geometric ensures a fast convergence), justifying the use of these simulations.
A study of the variance of $\ln(\sigma_{t+1}/\sigma_t)$ is also conducted and an expression of the variance of $\ln(\sigma_{t+1}/\sigma_t)$ is derived. For a cumulation parameter c_σ equal to $1/n^\alpha$ (where n is the dimension of the search problem), the standard deviation of $\ln(\sigma_{t+1}/\sigma_t)$ is about $\sqrt{(n^{2\alpha}+n)/n^{3\alpha}}$ times larger than its expected value. This indicates that keeping $c_\sigma < 1/n^{1/3}$ ensures that the standard deviation of $\ln(\sigma_{t+1}/\sigma_t)$ becomes negligible compared to its expected value when the dimension goes to infinity, which implies the stability of the algorithm with respect to the dimension.
In Section 4.3 we present two analyses of $(1, \lambda)$-ESs on a linear function $f : \mathbb{R}^n \to \mathbb{R}$ with a linear constraint $g : \mathbb{R}^n \to \mathbb{R}$. W.l.o.g. we can assume the problem to be

$$\begin{aligned} &\text{maximize } f(\boldsymbol{x}) \text{ for } \boldsymbol{x} \in \mathbb{R}^n \\ &\text{subject to } g(\boldsymbol{x}) \geq 0 \quad . \end{aligned}$$

The angle $\theta := (\nabla f, \nabla g)$ is an important characteristic of this problem, and the two studies of Section 4.3 are restricted to $\theta \in (0, \pi/2)$.
The first analysis on this problem, which is presented in 4.3.1, is of a $(1, \lambda)$-ES where two updates of the step-size are considered: one for which the step-size is kept constant, and

one where the step-size is adapted through cumulative step-size adaptation (see 2.3.8). This study was inspired by [14] which showed that the $(1,\lambda)$-CSA-ES fails on this linear constrained problem for too low values of θ, assuming the existence of an invariant measure for the sequence $(\delta_t)_{t\in\mathbb{N}}$, where δ_t is the signed distance from the mean of the sampling distribution to the constraint, normalized by the step-size, i.e. $\delta_t := g(\boldsymbol{X}_t)/\sigma_t$. In 4.3.1 for the $(1,\lambda)$-ES with constant step-size, $(\delta_t)_{t\in\mathbb{N}}$ is shown to be a geometrically ergodic φ-irreducible aperiodic Markov chain for which compact sets are small sets, from which the almost sure divergence of the algorithm, as detailed in (4.12), is deduced. Then for the $(1,\lambda)$-CSA-ES, the sequence $(\delta_t, \boldsymbol{p}_t^\sigma)_{t\in\mathbb{N}}$ where $\boldsymbol{p}_t^\sigma$ is the evolution path defined in (2.12) is shown to be a Markov chain, and in the simplified case where the cumulation parameter c_σ equals 1, $(\delta_t)_{t\in\mathbb{N}}$ is shown to be a geometrically ergodic φ-irreducible aperiodic Markov chain for which compact sets are small sets, from which the almost sure log-linear convergence or divergence of the step-size at a rate r is deduced. The sign of r indicates whether convergence or divergence takes place, and r is estimated through the use of Monte Carlo simulations. These simulations, justified by the geometric ergodicity of the Markov chain $(\delta_t)_{t\in\mathbb{N}}$, investigate the dependence of r with respect to different parameters, such as the constraint angle θ, the cumulation parameter c_σ or the population size λ, and show that for a large enough population size or low enough cumulation parameter, r is positive and so the step-size of the $(1,\lambda)$-CSA-ES successfully diverges log-linearly on this problem, but conversely for a low enough value of the constraint angle θ, r is negative and so the step-size of the $(1,\lambda)$-CSA-ES then converges log-linearly, thus failing on this problem.

The second analysis on the linear function with a linear constraint, presented in 4.3.2, investigates a $(1,\lambda)$-ES with constant step-size and a not necessarily Gaussian sampling distribution. The analysis establishes that if the sampling distribution is absolutely continuous and supported on $\mathbb{R}^n$ then the sequence $(\delta_t)_{t\in\mathbb{N}}$ is a φ-irreducible aperiodic Markov chain for which compact sets are small sets. From this, sufficient conditions are derived to ensure that the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is positive, Harris recurrent and V-geometrically ergodic for a specific function V. The Harris recurrence and positivity of the Markov chain is then used to apply a law of large numbers and deduce the divergence of the algorithm under these conditions. The effect of the covariance of the sampling distribution on the problem is then investigated, and it is shown that changing the covariance matrix is equivalent to changing the norm on the space, which in turn implies a change of the constraint angle θ. This effect gives useful insight on the results presented in 4.3.1, as it has been shown in 4.3.1 that a too low value of the constraint angle implies the log-linear convergence of the step-size for the $(1,\lambda)$-CSA-ES, therefore failing to solve the problem. Changing the covariance matrix can therefore trigger the success of the $(1,\lambda)$-CSA-ES on this problem. Finally, sufficient conditions on the marginals of the sampling distribution and the copula combining them are given to get the absolute continuity of the sampling distribution.

The results of Chapter 4 are important relatively to [2] and [24]. In [2] an IGO-flow (see 2.5.3) which can be related to a continuous-time ES is shown to locally converge to the critical points with positive definite Hessian of any C^2 function with Lebesgue negligible level sets, under assumptions including that the step-size of the algorithm diverges log-linearly on the

linear function. In [24] the $(1+1)$-ES using the so called one-fifth success rule [117] is shown to converge log-linearly on positively homogeneous functions (see (2.34) for a definition of positively homogeneous functions), under the assumption that $\mathbf{E}(\sigma_t/\sigma_{t+1}) < 1$ on the linear function, which is related to the log-linear divergence of the step-size on the linear function. We showed in Section 4.2 that for the $(1,\lambda)$-CSA-ES the step-size diverges log-linearly on the linear function; and in 4.3.1 that for too low constraint angle it does not; but with 4.3.2, adaptation of the covariance matrix can allow the step-size of a $(1,\lambda)$-ES with CSA step-size adaptation to successfully diverge even for low constraint angles. Therefore, although our analyses of ESs are restricted to linear problems, they relate to the convergence of ESs on C^2 and positively homogeneous functions.

5.2 Perspectives

The results presented in Chapter 3 present many different extensions:

- The techniques developed can be applied to prove φ-irreducibility, aperiodicity, T-chain property and that compact sets are small sets on many other problems. Particular problems of interest to us would be ESs adapting the covariance matrix, or using an evolution path (see 2.3.8).

- The transition function F from our model described in (5.1) is assumed in most of the results of Chapter 3 to be C^1. However, for an ES using the cumulative step-size adaptation described in (2.13), due to the square root in $\|\boldsymbol{p}_t^\sigma\| = \sqrt{\sum_{i=1}^n [\boldsymbol{p}_t^\sigma]_i^2}$ the transition function involved is not differentiable when $\boldsymbol{p}_t^\sigma = \mathbf{0}$. Although this can be alleviated by studying the slightly different version of CSA described in (4.11), as has been done in Chapter 4, it would be useful to extend the results of Chapter 3 to transition functions that are not C^1 everywhere.

- The distribution of the random elements $\alpha(\boldsymbol{x}, \boldsymbol{U}_{t+1})$ described in (5.1) is assumed in our model to be absolutely continuous. However, in elitist ESs such as the $(1+1)$-ES, there is a positive probability that the mean of the sampling distribution of the ES does not change over an iteration. Therefore, the distribution of $\alpha(\boldsymbol{x}, \boldsymbol{U}_{t+1})$ in this context has a singularity and does not fit the model of Chapter 3. Extending the results of Chapter 3 to a model where the distribution of the random elements $\alpha(\boldsymbol{x}, \boldsymbol{U}_{t+1})$ admits singularities would then allow us to apply them to elitist ESs.

- In [98, Chapter 7], the context described in 1.2.6 is that the Markov chain $(\Phi_t)_{t\in\mathbb{N}}$ is defined via Φ_0 following some initial distribution and $\Phi_{t+1} = F(\Phi_t, \boldsymbol{U}_{t+1})$; where the transition function $F : X \times \Omega \to X$ is supposed C^∞, and $(\boldsymbol{U}_t)_{t\in\mathbb{N}}$ is a sequence of i.i.d. random elements valued in Ω and admitting a density p. In this context it is shown that if the set $O_{\boldsymbol{w}} := \{\boldsymbol{u} \in O | p_{\boldsymbol{x}}(\boldsymbol{u}) > 0\}$ is connected, if there exists $\boldsymbol{x}^* \in X$ a globally attracting state, and if the control model $CM(F)$ is forward accessible (see 1.2.6), then the aperiodicity is proven to be implied by the connexity of the set $\overline{A_+(\boldsymbol{x}^*)}$. In our context, we gave sufficient conditions (including the existence of a strongly globally attracting

state) to prove aperiodicity. It would be interesting to investigate if the existence of a strongly globally attracting state is a necessary condition for aperiodicity, and to see if we could use in our context the condition of connexity of $\overline{A_+(\boldsymbol{x}^*)}$ to prove aperiodicity.

The techniques of Chapter 3 could be used to investigate the log-linear convergence of different ESs on scale-invariant functions (see (2.33) for a definition of scale-invariant functions). However, new techniques need to be developed to fully investigate the $(1,\lambda)$-CSA-ES on scale-invariant functions, when the Markov chain of interest is $(\boldsymbol{Z}_t, \boldsymbol{p}_t^\sigma)_{t\in\mathbb{N}}$ where $\boldsymbol{Z}_t = \boldsymbol{X}_t/\sigma_t$ and $\boldsymbol{p}_t^\sigma$ is the evolution path defined in (2.12). Indeed, as explained in 2.5.2, the drift function $V\colon (\boldsymbol{z},\boldsymbol{p}) \mapsto \|\boldsymbol{z}\|^\alpha + \|\boldsymbol{p}\|^\beta$, which generalizes the drift function usually considered for cases without the evolution path, cannot be used in this case with an evolution path to show a negative drift: a value close to 0 of $\|\boldsymbol{p}_t^\sigma\|$ combined with a high value of $\|\boldsymbol{Z}_t\|$ will result, due to (4.7), in a positive drift ΔV. To counteract this effect, in future research we will investigate drift functions that measure the mean drift after several iterations of the algorithm, which leave some iterations for the norm of the evolution path to increase, and then for the norm of $\boldsymbol{Z}_t$ to decrease.

Bibliography

[1] Akiko N Aizawa and Benjamin W Wah. Scheduling of genetic algorithms in a noisy environment. *Evolutionary Computation*, 2(2):97–122, 1994.

[2] Youhei Akimoto, Anne Auger, and Nikolaus Hansen. Convergence of the continuous time trajectories of isotropic evolution strategies on monotonic\ mathcal c^ 2-composite functions. In *Parallel Problem Solving from Nature-PPSN XII*, pages 42–51. Springer, 2012.

[3] Youhei Akimoto, Yuichi Nagata, Isao Ono, and Shigenobu Kobayashi. Bidirectional relation between cma evolution strategies and natural evolution strategies. In *Parallel Problem Solving from Nature, PPSN XI*, pages 154–163. Springer, 2010.

[4] Shun-Ichi Amari. Natural gradient works efficiently in learning. *Neural computation*, 10(2):251–276, 1998.

[5] Dirk V Arnold. *Noisy Optimization with Evolution Strategies*. Kluwer Academic Publishers, 2002.

[6] Dirk V Arnold. Resampling versus repair in evolution strategies applied to a constrained linear problem. *Evolutionary computation*, 21(3):389–411, 2013.

[7] Dirk V Arnold and H-G Beyer. Performance analysis of evolutionary optimization with cumulative step length adaptation. *IEEE Transactions on Automatic Control*, 49(4):617–622, 2004.

[8] Dirk V Arnold and Hans-Georg Beyer. Investigation of the (μ, λ)-es in the presence of noise. In *Evolutionary Computation, 2001. Proceedings of the 2001 Congress on*, volume 1, pages 332–339. IEEE, 2001.

[9] Dirk V Arnold and Hans-Georg Beyer. Local performance of the (μ/μi, λ)-es in a noisy environment. *Foundations of Genetic Algorithms*, 6:127–141, 2001.

[10] Dirk V Arnold and Hans-Georg Beyer. Local performance of the (1+ 1)-es in a noisy environment. *Evolutionary Computation, IEEE Transactions on*, 6(1):30–41, 2002.

[11] Dirk V Arnold and Hans-Georg Beyer. *On the effects of outliers on evolutionary optimization*. Springer, 2003.

[12] Dirk V Arnold and Hans-Georg Beyer. A general noise model and its effects on evolution strategy performance. *Evolutionary Computation, IEEE Transactions on*, 10(4):380–391, 2006.

[13] Dirk V Arnold and Hans-Georg Beyer. On the behaviour of evolution strategies optimising cigar functions. *Evolutionary computation*, 18(4):661–682, 2010.

[14] D.V. Arnold. On the behaviour of the $(1,\lambda)$-ES for a simple constrained problem. In *Foundations of Genetic Algorithms - FOGA 11*, pages 15–24. ACM, 2011.

[15] D.V. Arnold. On the behaviour of the $(1,\lambda)$-σSA-ES for a constrained linear problem. In *Parallel Problem Solving from Nature - PPSN XII*, pages 82–91. Springer, 2012.

[16] D.V. Arnold and H.G. Beyer. Random dynamics optimum tracking with evolution strategies. In *Parallel Problem Solving from Nature - PPSN VII*, pages 3–12. Springer, 2002.

[17] Sandra Astete-Morales, Marie-Liesse Cauwet, and Olivier Teytaud. Evolution strategies with additive noise: A convergence rate lower bound. In *Foundations of Genetic Algorithms*, page 9, 2015.

[18] A. Auger. Convergence results for the $(1,\lambda)$-SA-ES using the theory of φ-irreducible markov chains. *Theoretical Computer Science*, 334(1–3):35–69, 2005.

[19] A. Auger and N. Hansen. Reconsidering the progress rate theory for evolution strategies in finite dimensions. In ACM Press, editor, *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2006)*, pages 445–452, 2006.

[20] A. Auger and N. Hansen. Theory of evolution strategies: a new perspective. In A. Auger and B. Doerr, editors, *Theory of Randomized Search Heuristics: Foundations and Recent Developments*, chapter 10, pages 289–325. World Scientific Publishing, 2011.

[21] A Auger, N Hansen, JM Perez Zerpa, R Ros, and M Schoenauer. Empirical comparisons of several derivative free optimization algorithms. In *Acte du 9ieme colloque national en calcul des structures*, volume 1. Citeseer, 2009.

[22] Anne Auger. Analysis of stochastic continuous comparison-based black-box optimization, 2015.

[23] Anne Auger and Nikolaus Hansen. A restart cma evolution strategy with increasing population size. In *Evolutionary Computation, 2005. The 2005 IEEE Congress on*, volume 2, pages 1769–1776. IEEE, 2005.

[24] Anne Auger and Nikolaus Hansen. Linear convergence on positively homogeneous functions of a comparison based step-size adaptive randomized search: the (1+1) ES with generalized one-fifth success rule. *CoRR*, abs/1310.8397, 2013.

[25] Anne Auger and Nikolaus Hansen. On proving linear convergence of comparison-based step-size adaptive randomized search on scaling-invariant functions via stability of markov chains. *CoRR*, abs/1310.7697, 2013.

[26] Thomas Bäck. *Evolutionary algorithms in theory and practice: evolution strategies, evolutionary programming, genetic algorithms*. Oxford university press, 1996.

[27] Shumeet Baluja. Population-based incremental learning. a method for integrating genetic search based function optimization and competitive learning. Technical report, DTIC Document, 1994.

[28] T Bäck, F Hoffmeister, and HP Schwefel. A survey of evolution strategies. In *Proceedings of the Fourth International Conference on Genetic Algorithms*, 1991.

[29] Dimitri P Bertsekas. *Constrained optimization and Lagrange multiplier methods*. Academic press, 2014.

[30] H.-G. Beyer. *The theory of evolution strategies*. Natural computing series. Springer, Berlin, 2001.

[31] Hans-Georg Beyer. Toward a theory of evolution strategies: On the benefits of sex—the $(\mu/\mu, \lambda)$ theory. *Evolutionary Computation*, 3(1):81–111, 1995.

[32] Hans-Georg Beyer. Toward a theory of evolution strategies: Self-adaptation. *Evolutionary Computation*, 3(3):311–347, 1995.

[33] Alexis Bienvenüe and Olivier François. Global convergence for evolution strategies in spherical problems: some simple proofs and difficulties. *Theor. Comput. Sci.*, 306:269–289, September 2003.

[34] Ihor O Bohachevsky, Mark E Johnson, and Myron L Stein. Generalized simulated annealing for function optimization. *Technometrics*, 28(3):209–217, 1986.

[35] Mohammad Reza Bonyadi and Zbigniew Michalewicz. A locally convergent rotationally invariant particle swarm optimization algorithm. *Swarm Intelligence*, 8(3):159–198, 2014.

[36] Jürgen Branke, Christian Schmidt, and Hartmut Schmec. Efficient fitness estimation in noisy environments. In *Proceedings of genetic and evolutionary computation*, 2001.

[37] Charles George Broyden, John E Dennis, and Jorge J Moré. On the local and superlinear convergence of quasi-newton methods. *IMA Journal of Applied Mathematics*, 12(3):223–245, 1973.

[38] Richard H Byrd, Peihuang Lu, Jorge Nocedal, and Ciyou Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

[39] Erick Cantú-Paz. Adaptive sampling for noisy problems. In *Genetic and Evolutionary Computation–GECCO 2004*, pages 947–958. Springer, 2004.

[40] Rachid Chelouah and Patrick Siarry. A continuous genetic algorithm designed for the global optimization of multimodal functions. *Journal of Heuristics*, 6(2):191–213, 2000.

[41] Siddhartha Chib and Edward Greenberg. Understanding the metropolis-hastings algorithm. *The american statistician*, 49(4):327–335, 1995.

[42] A. Chotard and A. Auger. Verifiable conditions for irreducibility, aperiodicity and weak feller property of a general markov chain. Submitted to Bernouilli, 2015.

[43] A. Chotard, A. Auger, and N. Hansen. Cumulative step-size adaptation on linear functions: Technical report. Technical report, Inria, 2012.

[44] A. Chotard, A. Auger, and N. Hansen. Markov chain analysis of evolution strategies on a linear constraint optimization problem. In *Evolutionary Computation (CEC), 2014 IEEE Congress on*, pages 159–166, July 2014.

[45] A. Chotard, A. Auger, and N. Hansen. Markov chain analysis of cumulative step-size adaptation on a linear constraint problem. *Evol. Comput.*, 2015.

[46] Alexandre Chotard, Anne Auger, and Nikolaus Hansen. Cumulative step-size adaptation on linear functions. In *Parallel Problem Solving from Nature - PPSN XII*, pages 72–81. Springer, september 2012.

[47] Alexandre Chotard and Martin Holena. A generalized markov-chain modelling approach to $(1,\lambda)$-es linear optimization. In Thomas Bartz-Beielstein, Jürgen Branke, Bogdan Filipič, and Jim Smith, editors, *Parallel Problem Solving from Nature – PPSN XIII*, volume 8672 of *Lecture Notes in Computer Science*, pages 902–911. Springer International Publishing, 2014.

[48] Alexandre Chotard and Martin Holena. A generalized markov-chain modelling approach to $(1,\lambda)$-es linear optimization: Technical report. Technical report, Inria, 2014.

[49] Maurice Clerc and James Kennedy. The particle swarm-explosion, stability, and convergence in a multidimensional complex space. *Evolutionary Computation, IEEE Transactions on*, 6(1):58–73, 2002.

[50] Carlos A Coello Coello. Theoretical and numerical constraint-handling techniques used with evolutionary algorithms: a survey of the state of the art. *Computer methods in applied mechanics and engineering*, 191(11):1245–1287, 2002.

[51] Carlos Artemio Coello Coello. Constraint-handling techniques used with evolutionary algorithms. In *Proceedings of the 14th annual conference companion on Genetic and evolutionary computation*, pages 849–872. ACM, 2012.

[52] Andrew R Conn, Nicholas IM Gould, and Ph L Toint. *Trust region methods*, volume 1. Siam, 2000.

[53] Anton Dekkers and Emile Aarts. Global optimization and simulated annealing. *Mathematical programming*, 50(1-3):367–393, 1991.

[54] Peter Deuflhard. *Newton methods for nonlinear problems: affine invariance and adaptive algorithms*, volume 35. Springer Science & Business Media, 2011.

[55] Johannes M Dieterich and Bernd Hartke. Empirical review of standard benchmark functions using evolutionary global optimization. *arXiv preprint arXiv:1207.4318*, 2012.

[56] Benjamin Doerr, Edda Happ, and Christian Klein. Crossover can provably be useful in evolutionary computation. In *Proceedings of the 10th annual conference on Genetic and evolutionary computation*, pages 539–546. ACM, 2008.

[57] Roger Gämperle, Sibylle D Müller, and Petros Koumoutsakos. A parameter study for differential evolution. *Advances in intelligent systems, fuzzy systems, evolutionary computation*, 10:293–298, 2002.

[58] Sylvain Gelly, Sylvie Ruette, and Olivier Teytaud. Comparison-based algorithms are robust and randomized algorithms are anytime. *Evol. Comput.*, 15(4):411–434, December 2007.

[59] Tobias Glasmachers, Tom Schaul, Sun Yi, Daan Wierstra, and Jürgen Schmidhuber. Exponential natural evolution strategies. In *Proceedings of the 12th annual conference on Genetic and evolutionary computation*, pages 393–400. ACM, 2010.

[60] David E Goldberg. *Genetic algorithms*. Pearson Education India, 2006.

[61] D Goldfarb and Ph L Toint. Optimal estimation of jacobian and hessian matrices that arise in finite difference calculations. *Mathematics of Computation*, 43(167):69–88, 1984.

[62] Ulrich Hammel and Thomas Bäck. Evolution strategies on noisy functions how to improve convergence properties. In *Parallel Problem Solving from Nature—PPSN III*, pages 159–168. Springer, 1994.

[63] N. Hansen. An analysis of mutative σ-self-adaptation on linear fitness functions. *Evolutionary Computation*, 14(3):255–275, 2006.

[64] N. Hansen, F. Gemperle, A. Auger, and P. Koumoutsakos. When do heavy-tail distributions help? In T. P. Runarsson et al., editors, *Parallel Problem Solving from Nature PPSN IX*, volume 4193 of *Lecture Notes in Computer Science*, pages 62–71. Springer, 2006.

[65] N. Hansen, S.P.N. Niederberger, L. Guzzella, and P. Koumoutsakos. A method for handling uncertainty in evolutionary optimization with an application to feedback control of combustion. *IEEE Transactions on Evolutionary Computation*, 13(1):180–197, 2009.

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

[67] Nikolaus Hansen, Dirk V Arnold, and Anne Auger. Evolution strategies. In *Springer Handbook of Computational Intelligence*, pages 871–898. Springer, 2015.

[68] Nikolaus Hansen, Asma Atamna, and Anne Auger. How to Assess Step-Size Adaptation Mechanisms in Randomised Search. In T. Bartz-Beielstein et al, editor, *Parallel Problem Solving from Nature, PPSN XIII*, volume 8672 of *LNCS*, pages 60–69, Ljubljana, Slovenia, September 2014. Springer.

[69] Nikolaus Hansen and Anne Auger. Principled Design of Continuous Stochastic Search: From Theory to Practice. In Yossi Borenstein and Alberto Moraglio, editors, *Theory and Principled Methods for the Design of Metaheuristics*, Natural Computing Series, pages 145–180. Springer, 2014.

[70] Nikolaus Hansen, Raymond Ros, Nikolas Mauny, Marc Schoenauer, and Anne Auger. 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.

[71] W Keith Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, 57(1):97–109, 1970.

[72] Onésimo Herná-Lerma and Jean Bernard Lasserre. Markov chains and ergodic theorems. In *Markov Chains and Invariant Probabilities*, pages 21–39. Springer, 2003.

[73] John H Holland. Adaptation in natural and artificial system: an introduction with application to biology, control and artificial intelligence. *Ann Arbor, University of Michigan Press*, 1975.

[74] Jens Jägersküpper. *Analysis of a simple evolutionary algorithm for minimization in Euclidean spaces*. Springer, 2003.

[75] Jens Jägersküpper. Rigorous runtime analysis of the (1+ 1) es: 1/5-rule and ellipsoidal fitness landscapes. In *Foundations of Genetic Algorithms*, pages 260–281. Springer, 2005.

[76] Jens Jägersküpper. Probabilistic runtime analysis of (1+< over>, λ), es using isotropic mutations. In *Proceedings of the 8th annual conference on Genetic and evolutionary computation*, pages 461–468. ACM, 2006.

[77] Jens JäGersküPper. Lower bounds for randomized direct search with isotropic sampling. *Operations research letters*, 36(3):327–332, 2008.

[78] Mohamed Jebalia and Anne Auger. Log-linear convergence of the scale-invariant (μ/μ w, λ)-es and optimal μ for intermediate recombination for large population sizes. In *Parallel Problem Solving from Nature, PPSN XI*, pages 52–62. Springer, 2010.

[79] Mohamed Jebalia, Anne Auger, and Nikolaus Hansen. Log-linear convergence and divergence of the scale-invariant (1+ 1)-es in noisy environments. *Algorithmica*, 59(3):425–460, 2011.

[80] Yaochu Jin and Jürgen Branke. Evolutionary optimization in uncertain environments-a survey. *Evolutionary Computation, IEEE Transactions on*, 9(3):303–317, 2005.

[81] Terry Jones. Crossover, macromutation, and population-based search. In *Proceedings of the Sixth International Conference on Genetic Algorithms*, pages 73–80. Citeseer, 1995.

[82] William Karush. *Minima of functions of several variables with inequalities as side constraints*. PhD thesis, Master's thesis, Dept. of Mathematics, Univ. of Chicago, 1939.

[83] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Neural Networks, 1995. Proceedings., IEEE International Conference*, volume 4, pages 1942–1948, 1995.

[84] Scott Kirkpatrick, C Daniel Gelatt, Mario P Vecchi, et al. Optimization by simulated annealing. *science*, 220(4598):671–680, 1983.

[85] Slawomir Koziel and Zbigniew Michalewicz. Evolutionary algorithms, homomorphous mappings, and constrained parameter optimization. *Evolutionary computation*, 7(1):19–44, 1999.

[86] HW Kuhn and AW Tucker. Nonlinear programming. sid 481–492 i proc. of the second berkeley symposium on mathematical statistics and probability, 1951.

[87] Jouni Lampinen and Ivan Zelinka. On Stagnation Of The Differential Evolution Algorithm. In *Proceedings of MENDEL 2000, 6th International Mendel Conference on Soft Computing*, pages 76–83, 2000.

[88] Pedro Larranaga and Jose A Lozano. *Estimation of distribution algorithms: A new tool for evolutionary computation*, volume 2. Springer Science & Business Media, 2002.

[89] Randall J LeVeque. *Finite difference methods for ordinary and partial differential equations: steady-state and time-dependent problems*, volume 98. Siam, 2007.

[90] Dong C Liu and Jorge Nocedal. On the limited memory bfgs method for large scale optimization. *Mathematical programming*, 45(1-3):503–528, 1989.

[91] M Locatelli. Simulated annealing algorithms for continuous global optimization: convergence conditions. *Journal of Optimization Theory and applications*, 104(1):121–133, 2000.

[92] Ilya Loshchilov. A computationally efficient limited memory CMA-ES for large scale optimization. *CoRR*, abs/1404.5520, 2014.

[93] David G Luenberger. *Introduction to linear and nonlinear programming*, volume 28. Addison-Wesley Reading, MA, 1973.

[94] Rafael Martí. Multi-start methods. In *Handbook of metaheuristics*, pages 355–368. Springer, 2003.

[95] Ken IM McKinnon. Convergence of the nelder–mead simplex method to a nonstationary point. *SIAM Journal on Optimization*, 9(1):148–158, 1998.

[96] N Metropolis, A Rosenbluth, M Rosenbluth, A Teller, and E Teller. Simulated annealing. *Journal of Chemical Physics*, 21:1087–1092, 1953.

[97] Nicholas Metropolis, Arianna W Rosenbluth, Marshall N Rosenbluth, Augusta H Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The journal of chemical physics*, 21(6):1087–1092, 1953.

[98] S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, second edition, 1993.

[99] Efrén Mezura-Montes and Carlos A Coello Coello. Constrained optimization via multiobjective evolutionary algorithms. In *Multiobjective problem solving from nature*, pages 53–75. Springer, 2008.

[100] Efrén Mezura-Montes and Carlos A Coello Coello. Constraint-handling in nature-inspired numerical optimization: past, present and future. *Swarm and Evolutionary Computation*, 1(4):173–194, 2011.

[101] Zbigniew Michalewicz. *Genetic algorithms+ data structures= evolution programs*. Springer Science & Business Media, 2013.

[102] Zbigniew Michalewicz and Girish Nazhiyath. Genocop iii: A co-evolutionary algorithm for numerical optimization problems with nonlinear constraints. In *Evolutionary Computation, 1995., IEEE International Conference on*, volume 2, pages 647–651. IEEE, 1995.

[103] Zbigniew Michalewicz and Marc Schoenauer. Evolutionary algorithms for constrained parameter optimization problems. *Evolutionary computation*, 4(1):1–32, 1996.

[104] Melanie Mitchell. *An introduction to genetic algorithms*. MIT press, 1998.

[105] Christopher K Monson and Kevin D Seppi. Linear equality constraints and homomorphous mappings in pso. In *Evolutionary Computation, 2005. The 2005 IEEE Congress on*, volume 1, pages 73–80. IEEE, 2005.

[106] Heinz Mühlenbein, M Schomisch, and Joachim Born. The parallel genetic algorithm as function optimizer. *Parallel computing*, 17(6):619–632, 1991.

[107] Marco Muselli. A theoretical approach to restart in global optimization. *Journal of Global Optimization*, 10(1):1–16, 1997.

[108] John A Nelder and Roger Mead. A simplex method for function minimization. *The computer journal*, 7(4):308–313, 1965.

[109] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.

[110] Jorge Nocedal and Stephen J Wright. Conjugate gradient methods. *Numerical Optimization*, pages 101–134, 2006.

[111] Y. Ollivier, L. Arnold, A. Auger, and N. Hansen. Information-Geometric Optimization Algorithms: A Unifying Picture via Invariance Principles. *ArXiv e-prints*, June 2011.

[112] Michael JD Powell. The newuoa software for unconstrained optimization without derivatives. In *Large-scale nonlinear optimization*, pages 255–297. Springer, 2006.

[113] Michael JD Powell. The bobyqa algorithm for bound constrained optimization without derivatives. *Department of Applied Mathematics and Theoretical Physics. Department of Applied Mathematics and Theoretical Physics, Cambrigde, England: sn*, 2009.

[114] Kenneth Price, Rainer M Storn, and Jouni A Lampinen. *Differential evolution: a practical approach to global optimization*. Springer Science & Business Media, 2006.

[115] T.A. Jeeves R. Hooke. Direct search solution of numerical and statistical problems. *Journal of the Association for Computing Machinery (ACM)*, 8:212–239, 1961.

[116] I. Rechenberg. Cybernetic solution path of an experimental problem. *Royal Aircraft Establishment Library Translation No. 1122*, 1122, 1965.

[117] Ingo Rechenberg. Evolution strategy: Optimization of technical systems by means of biological evolution. *Fromman-Holzboog, Stuttgart*, 104, 1973.

[118] Daniel Revuz. *Markov chains*. Elsevier, 2008.

[119] Raymond Ros. Comparison of newuoa with different numbers of interpolation points on the bbob noiseless testbed. In *Proceedings of the 12th annual conference companion on Genetic and evolutionary computation*, pages 1487–1494. ACM, 2010.

[120] Raymond Ros. Comparison of newuoa with different numbers of interpolation points on the bbob noisy testbed. In *Proceedings of the 12th annual conference companion on Genetic and evolutionary computation*, pages 1495–1502. ACM, 2010.

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

[122] Günter Rudolph. Self-adaptive mutations may lead to premature convergence. *Evolutionary Computation, IEEE Transactions on*, 5(4):410–414, 2001.

[123] Victor S Ryaben'kii and Semyon V Tsynkov. *A theoretical introduction to numerical analysis*. CRC Press, 2006.

[124] Sancho Salcedo-Sanz. A survey of repair methods used as constraint handling techniques in evolutionary algorithms. *Computer science review*, 3(3):175–192, 2009.

[125] Yasuhito Sano and Hajime Kita. Optimization of noisy fitness functions by means of genetic algorithms using history of search with test of estimation. In *Evolutionary Computation, 2002. CEC'02. Proceedings of the 2002 Congress on*, volume 1, pages 360–365. IEEE, 2002.

[126] Hans-Paul Schwefel. Kybernetische evolution als strategie der experimentellen forschung in der strömungstechnik. *Master's thesis, Technical University of Berlin*, 1965.

[127] Hans-Paul Schwefel. Adaptive mechanismen in der biologischen evolution und ihr einfluss auf die evolutionsgeschwindigkeit. *Interner Bericht der Arbeitsgruppe Bionik und Evolutionstechnik am Institut für Mess-und Regelungstechnik Re*, 215(3), 1974.

[128] Hans-Paul Schwefel. *Numerical optimization of computer models*. John Wiley & Sons, Inc., 1981.

[129] Hans-Paul Schwefel. *Collective phenomena in evolutionary systems*. Universität Dortmund. Abteilung Informatik, 1987.

[130] Hans-Paul Schwefel. Evolution and optimum seeking. sixth-generation computer technology series, 1995.

[131] Bernhard Sendhoff, Hans-Georg Beyer, and Markus Olhofer. The influence of stochastic quality functions on evolutionary search. *Recent advances in simulated evolution and learning*, 2:152–172, 2004.

[132] Yuhui Shi and Russell C Eberhart. Empirical study of particle swarm optimization. In *Evolutionary Computation, 1999. CEC 99. Proceedings of the 1999 Congress on*, volume 3. IEEE, 1999.

[133] Alice E Smith, David W Coit, Thomas Baeck, David Fogel, and Zbigniew Michalewicz. Penalty functions. *Evolutionary computation*, 2:41–48, 2000.

[134] Rainer Storn and Kenneth Price. Differential evolution–a simple and efficient heuristic for global optimization over continuous spaces. *Journal of global optimization*, 11(4):341–359, 1997.

[135] Olivier Teytaud and Sylvain Gelly. General lower bounds for evolutionary algorithms. In *Parallel Problem Solving from Nature-PPSN IX*, pages 21–31. Springer, 2006.

[136] Virginia Torczon. On the convergence of pattern search algorithms. *SIAM Journal on optimization*, 7(1):1–25, 1997.

[137] W Townsend. The single machine problem with quadratic penalty function of completion times: a branch-and-bound solution. *Management Science*, 24(5):530–534, 1978.

[138] Daan Wierstra, Tom Schaul, Tobias Glasmachers, Yi Sun, Jan Peters, and Jürgen Schmidhuber. Natural evolution strategies. *The Journal of Machine Learning Research*, 15(1):949–980, 2014.

[139] Daan Wierstra, Tom Schaul, Jan Peters, and Juergen Schmidhuber. Natural evolution strategies. In *Evolutionary Computation, 2008. CEC 2008.(IEEE World Congress on Computational Intelligence). IEEE Congress on*, pages 3381–3387. IEEE, 2008.

[140] Philip Wolfe. Convergence conditions for ascent methods. *SIAM review*, 11(2):226–235, 1969.

[141] David H Wolpert and William G Macready. No free lunch theorems for optimization. *Evolutionary Computation, IEEE Transactions on*, 1(1):67–82, 1997.

[142] Margaret H Wright. Direct search methods: Once scorned, now respectable. *Pitman Research Notes in Mathematics Series*, pages 191–208, 1996.

[143] Xinjie Yu and Mitsuo Gen. *Introduction to evolutionary algorithms*. Springer Science & Business Media, 2010.

[144] Ya-xiang Yuan. A review of trust region algorithms for optimization. In *ICIAM*, volume 99, pages 271–282, 2000.

[145] Zelda B Zabinsky and Robert L Smith. Pure adaptive search in global optimization. *Mathematical Programming*, 53(1-3):323–338, 1992.

[146] Anatoly Zhigljavsky and Antanas Žilinskas. *Stochastic global optimization*, volume 9. Springer Science & Business Media, 2007.