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

doi:10.1162/EVCO_a_00166

Abstract

This paper analyzes a $(1, \lambda)$-Evolution Strategy, a randomized 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* randomized 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 the 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 $h : \mathbb{R} \to \mathbb{R}$.

Manuscript received: November 10, 2014; revised: September 14, 2015; accepted: September 16, 2015.

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 randomized algorithms have been proposed; see Mezura-Montes and Coello (2011) for a survey. For ES, common techniques are resampling (resample a solution until it lies in the feasible domain), repair of solutions that project infeasible points onto the feasible domain (Arnold, 2011b, 2013), penalty methods where infeasible solutions are penalized either by a quantity that depends on the distance to the constraint if this latter one can be computed with adaptive penalty weights (e.g., Hansen et al., 2009; Arnold and Porter, 2015) or by the constraint value itself (e.g., stochastic ranking; Runarsson and Yao, 2000), or methods inspired by multiobjective 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 far from the optimum relative to the step-size, and "solving" this function means diverging. The linear constraint models being close to the constraint relative to the step-size and far from other constraints. Because of the invariance of the algorithm to the composition of the objective function by a strictly increasing map, the linear function can 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, and the constraint being handled through resampling or repairing infeasible solutions (Arnold, 2011a; 2012; 2013). The conclusion is that with step-size adaptation, 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 behavior of the algorithm: typically these models arise from approximations (considering some random variables equal to their expected value, etc.) and assume mathematical properties like the existence of stationary distributions of underlying Markov chains without accompanying proof.

In contrast, our motivation is to study the algorithm without simplifications and to prove rigorously different mathematical properties of the algorithm that allow deducing the exact behavior 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, the 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 and 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 that allow studying the influence of different parameters of the algorithm. We start by illustrating in detail 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 (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 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 for which estimates are provided.

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 analyze the cumulative step-size adaptation case. Finally, in Section 6, we discuss our results and our methodology.

A preliminary version of this paper appeared in conference proceedings (Chotard et al., 2014). The analysis of path length control with cumulation parameter equal to 1 is however fully new, as is the discussion on how to analyze the case with cumulation parameter smaller than 1. Also, Figures 4–11 as well as the convergence of the progress rate in Theorem 1 are new.

1.1 Notation

Throughout this article, we denote by φ the density function of the standard multivariate normal distribution (the dimension being clarified within the context), and by Φ the cumulative distribution function of a standard univariate normal distribution. The standard (unidimensional) normal distribution is denoted by $\mathcal{N}(0, 1)$, the (n-dimensional) multivariate normal distribution with covariance matrix identity by $\mathcal{N}(\mathbf{0}, \mathrm{Id}_n)$, and the ith-order statistic of λ i.i.d. standard normal random variables by $\mathcal{N}_{i:\lambda}$. The uniform distribution on an interval I is denoted by $\mathcal{U}_I$. The set of natural numbers (including 0) is denoted by $\mathbb{N}$, and the set of real numbers by $\mathbb{R}$. We denote by $\mathbb{R}_+$ 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 a topological space $\mathcal{X}$, $\mathcal{B}(\mathcal{X})$ denotes the Borel algebra of $\mathcal{X}$. We denote by μ_{Leb} the Lebesgue measure on $\mathbb{R}$, and for $A \subset \mathbb{R}$, μ_A denotes the trace measure $\mu_A : B \in \mathcal{B}(\mathbb{R}) \mapsto \mu_{\mathrm{Leb}}(A \cap B)$. For two vectors $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$, we denote by $[\mathbf{x}]_i$ the ith-coordinate of $\mathbf{x}$ and by $\mathbf{x}.\mathbf{y}$ the scalar product of $\mathbf{x}$ and $\mathbf{y}$. Take $(a, b) \in \mathbb{N}^2$ with $a \leq b$, we denote by $[a..b]$ the interval of integers between a and b. The gamma function is denoted by Γ. For $\mathbf{X}$ and $\mathbf{Y}$ two random vectors, we write $\mathbf{X} \overset{d}{=} \mathbf{Y}$ 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 write $X_t \overset{a.s.}{\to} X$ if X_t converges almost surely to X, and $X_t \overset{P}{\to} X$ if X_t converges in probability to X. For X a random variable and π a probability measure, we denote by $\mathbf{E}(X)$ the expected value of X, and by $\mathbf{E}_\pi(X)$ the expected value of X when X has distribution π.

2 Problem Statement and Algorithm Definition

2.1 (1, λ)-ES with Resampling

In this paper we study the behavior of a $(1, \lambda)$-Evolution Strategy *maximizing* a function $f\colon \mathbb{R}^n \to \mathbb{R}$, $\lambda \geq 2$, $n \geq 2$, 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 infeasible 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 algorithm 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 ith candidate solution is denoted by $\tilde{\mathbf{Y}}_t^i$, and the realization of the multivariate normal distribution giving $\tilde{\mathbf{Y}}_t^i$ is $\tilde{\mathbf{N}}_t^i$, that is,

$$\tilde{\mathbf{Y}}_t^i = \mathbf{X}_t + \sigma_t \tilde{\mathbf{N}}_t^i. \tag{2}$$

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.

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, that is,

$$\mathbf{X}_{t+1} = \tilde{\mathbf{Y}}_t^\star = \mathbf{X}_t + \sigma_t \tilde{\mathbf{N}}_t^\star. \tag{3}$$

The step-size and other internal parameters are then adapted. We denote for the moment in a nonspecific 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 define later specific rules for this adaptation.

2.2 Linear Fitness Function with Linear Constraint

In this article 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 by $\mathbf{n} := -\nabla g$ 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 as 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

$$\begin{aligned} &\text{maximize} \quad f(\mathbf{x}) = [\mathbf{x}]_1 \quad \text{subject to} \\ &g(\mathbf{x}) = -[\mathbf{x}]_1 \cos\theta - [\mathbf{x}]_2 \sin\theta > 0. \end{aligned} \tag{7}$$

Although $\tilde{\mathbf{N}}_t^i$ and $\tilde{\mathbf{N}}_t^\star$ are in $\mathbb{R}^n$, because of 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 notation, the vector $\tilde{\mathbf{N}}_t^i$ will denote the two-dimensional vector $([\tilde{\mathbf{N}}_t^i]_1, [\tilde{\mathbf{N}}_t^i]_2)$, likewise $\tilde{\mathbf{N}}_t^\star$ will denote the two-dimensional vector $([\tilde{\mathbf{N}}_t^\star]_1, [\tilde{\mathbf{N}}_t^\star]_2)$, and $\mathbf{n}$ will denote the two-dimensional vector $(\cos\theta, \sin\theta)$. The coordinate system will also be used as $(\mathbf{e}_1, \mathbf{e}_2)$ only.

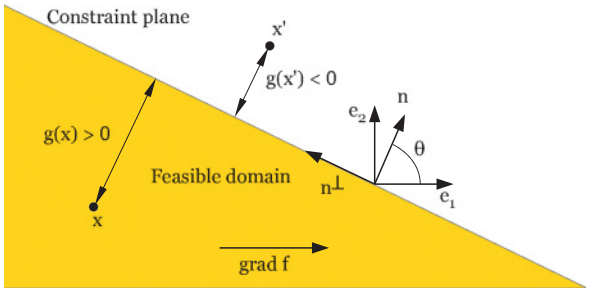

Figure 1: Linear function with a linear constraint in the plane generated by ∇f and $\mathbf{n}$, a normal vector to the constraint hyperplane with angle $\theta \in (0, \pi/2)$ with ∇f. The point $\mathbf{x}$ is at distance $g(\mathbf{x})$ from the constraint.

Following Arnold (2011a; 2012) and 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 ith 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 $\boldsymbol{n}$ as in Eq. (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*

$$p_\delta(\boldsymbol{x}) = \frac{\varphi(\boldsymbol{x})\mathbf{1}_{\mathbb{R}_+^*}(\delta - \boldsymbol{x}.\boldsymbol{n})}{\Phi(\delta)}. \tag{9}$$

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

$$p_{1,\delta}(x) = \varphi(x) \frac{\Phi\left(\frac{\delta - x\cos\theta}{\sin\theta}\right)}{\Phi(\delta)} \tag{10}$$

(see Arnold, 2011a, Eq. 4), and we denote by $F_{1,\delta}$ 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 $\boldsymbol{n}$ as in Eq. (6). Let the feasible step $\tilde{\boldsymbol{N}}_t^i$ be the random vector described in Lemma 1 and $\boldsymbol{Q}$ be the two-dimensional rotation matrix of angle θ. Then*

$$\tilde{\boldsymbol{N}}_t^i \overset{d}{=} \tilde{F}_{\delta_t}^{-1}(U_t^i)\boldsymbol{n} + \mathcal{N}_t^i \boldsymbol{n}^\perp = \boldsymbol{Q}^{-1} \begin{pmatrix} \tilde{F}_{\delta_t}^{-1}(U_t^i) \\ \mathcal{N}_t^i, \end{pmatrix}, \tag{11}$$

where $\tilde{F}_{\delta_t}^{-1}$ denotes the generalized inverse[1] *of the cumulative distribution of $\tilde{\mathbf{N}}_t^i.\boldsymbol{n}$, $U_t^i \sim \mathcal{U}_{[0,1]}$, $\mathcal{N}_t^i \sim \mathcal{N}(0,1)$ 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 it has the density described in the following lemma.

LEMMA 3: *Let a $(1, \lambda)$-ES with resampling optimize problem (7). Then the distribution of the selected step $\tilde{\boldsymbol{N}}_t^\star$ only depends on the normalized distance to the constraint δ_t and its density, given that δ_t equals δ, reads*

$$\begin{aligned} p_\delta^\star(\boldsymbol{x}) &= \lambda p_\delta(\boldsymbol{x})\, F_{1,\delta}([\boldsymbol{x}]_1)^{\lambda-1}, \\ &= \lambda \frac{\varphi(\boldsymbol{x})\mathbf{1}_{\mathbb{R}_+^*}(\delta - \boldsymbol{x}.\boldsymbol{n})}{\Phi(\delta)} \left(\int_{-\infty}^{[x]_1} \varphi(u) \frac{\Phi(\frac{\delta - u\cos\theta}{\sin\theta})}{\Phi(\delta)} du \right)^{\lambda-1}, \end{aligned} \tag{12}$$

[1] The generalized inverse of $\tilde{F}_\delta$ is $\tilde{F}_\delta^{-1}(y) := inf_{x\in\mathbb{R}}\{\tilde{F}_\delta(x) \geq y\}$.

where p_δ is the density of $\tilde{\mathbf{N}}_t^i$ given that $\delta_t = \delta$ given in Eq. (9) and $F_{1,\delta}$ the cumulative distribution function of $[\tilde{\mathbf{N}}_t^i]_1$ 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 $(\tilde{\mathbf{N}}_t^i)_{i\in[1..\lambda]}$ correspond to the order statistic on $([\tilde{\mathbf{N}}_t^i]_1)_{i\in[1..\lambda]}$. If we look at the joint cumulative distribution $F_\delta^\star$ of $\tilde{\mathbf{N}}_t^\star$

$$\begin{aligned} F_\delta^\star(x, y) &= \Pr\left([\tilde{\mathbf{N}}_t^\star]_1 \le x, [\tilde{\mathbf{N}}_t^\star]_2 \le y\right) \\ &= \sum_{i=1}^{\lambda} \Pr\left(\tilde{\mathbf{N}}_t^i \le \begin{pmatrix} x \\ y \end{pmatrix}, [\tilde{\mathbf{N}}_t^j]_1 < [\tilde{\mathbf{N}}_t^i]_1 \text{ for } j \ne i\right) \end{aligned}$$

by summing disjoints events. The vectors $(\tilde{\mathbf{N}}_t^i)_{i\in[1..\lambda]}$ being independent and identically distributed,

$$\begin{aligned} F_\delta^\star(x, y) &= \lambda \Pr\left(\tilde{\mathbf{N}}_t^1 \le \begin{pmatrix} x \\ y \end{pmatrix}, [\tilde{\mathbf{N}}_t^j]_1 < [\tilde{\mathbf{N}}_t^1]_1 \text{ for } j \ne 1\right) \\ &= \lambda \int_{-\infty}^{x}\int_{-\infty}^{y} p_\delta(u, v) \prod_{j=2}^{\lambda} \Pr([\tilde{\mathbf{N}}_t^j]_1 < u) \mathrm{d}v\mathrm{d}u \\ &= \lambda \int_{-\infty}^{x}\int_{-\infty}^{y} p_\delta(u, v) F_{1,\delta}(u)^{\lambda-1} \mathrm{d}v\mathrm{d}u. \end{aligned}$$

Deriving $F_\delta^\star$ on x and y yields the density of $\tilde{\mathbf{N}}_t^\star$ of Eq. (12). □

We may now obtain the marginal of $[\tilde{\mathbf{N}}_t^\star]_1$ and $[\tilde{\mathbf{N}}_t^\star]_2$.

COROLLARY 1: *Let a $(1, \lambda)$-ES with resampling optimize problem (7). Then the marginal distribution of $[\tilde{\mathbf{N}}_t^\star]_1$ only depends on δ_t, and its density given that δ_t equals δ reads*

$$\begin{aligned} p_{1,\delta}^\star(x) &= \lambda p_{1,\delta}(x) F_{1,\delta}(x)^{\lambda-1}, \\ &= \lambda\varphi(x) \frac{\Phi\left(\frac{\delta - x\cos\theta}{\sin\theta}\right)}{\Phi(\delta)} F_{1,\delta}(x)^{\lambda-1}, \end{aligned} \tag{13}$$

and the same holds for $[\tilde{\mathbf{N}}_t^\star]_2$, whose marginal density reads

$$p_{2,\delta}^\star(y) = \lambda \frac{\varphi(y)}{\Phi(\delta)} \int_{-\infty}^{\frac{\delta - y\sin\theta}{\cos\theta}} \varphi(u) F_{1,\delta}(u)^{\lambda-1} du. \tag{14}$$

PROOF: Integrating Eq. (12) directly yields Eq. (13).

The conditional density function of $[\tilde{\mathbf{N}}_t^\star]_2$ is

$$p_{2,\delta}^\star(y | [\tilde{\mathbf{N}}_t^\star]_1 = x) = \frac{p_\delta^\star((x, y))}{p_{1,\delta}^\star(x)}.$$

As $p_{2,\delta}^\star(y) = \int_{\mathbb{R}} p_{2,\delta}^\star(y | [\tilde{\mathbf{N}}_t^\star]_1 = x) p_{1,\delta}^\star(x) \mathrm{d}x$, using the previous equation with Eq. (12) gives that $p_{2,\delta}^\star(y) = \int_{\mathbb{R}} \lambda p_\delta((x, y)) F_{1,\delta}(x)^{\lambda-1} \mathrm{d}x$, which with Eq. (9) gives

$$p_{2,\delta}^\star(y) = \lambda \frac{\varphi(y)}{\Phi(\delta)} \int_{\mathbb{R}} \varphi(x) \mathbf{1}_{\mathbb{R}_+^*}\left(\delta - \begin{pmatrix} x \\ y \end{pmatrix}.\mathbf{n}\right) F_{1,\delta}(x)^{\lambda-1} \mathrm{d}x.$$

The condition $\delta - x\cos\theta - y\sin\theta > 0$ is equivalent to $x < (\delta - y\sin\theta)/\cos\theta$; hence Eq. (14) holds. □

We will need an expression of the random vector $\tilde{\mathbf{N}}_t^\star$ as a function of δ_t and a random vector composed of a *finite* number of i.i.d. random variables. Using the notation of Lemma 2, we define the function $\tilde{\mathcal{G}} : \mathbb{R}_+^* \times ([0,1] \times \mathbb{R}) \to \mathbb{R}^2$ as

$$\tilde{\mathcal{G}}(\delta, \mathbf{w}) = \mathbf{Q}^{-1} \begin{pmatrix} \tilde{F}_\delta^{-1}([\mathbf{w}]_1) \\ [\mathbf{w}]_2 \end{pmatrix}. \tag{15}$$

According to Lemma 2, given that $U \sim \mathcal{U}_{[0,1]}$ and $\mathcal{N} \sim \mathcal{N}(0,1)$, $(\tilde{F}_\delta^{-1}(U), \mathcal{N})$ (resp. $\tilde{\mathcal{G}}(\delta, (U, \mathcal{N}))$) is distributed as the resampled step $\tilde{\mathbf{N}}_t^i$ 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

$$\mathcal{G}(\delta, (\mathbf{w}_i)_{i\in[1..\lambda]}) = \underset{\mathbf{N}\in\{\tilde{\mathcal{G}}(\delta,\mathbf{w}_i)|i\in[1..\lambda]\}}{\operatorname{argmax}} f(\mathbf{N}). \tag{16}$$

As shown in the following proposition, given that $\mathbf{W}_t^i \sim (\mathcal{U}_{[0,1]}, \mathcal{N}(0,1))$ and $\mathcal{W}_t = (\mathbf{W}_t^i)_{i\in[1..\lambda]}$, the function $\mathcal{G}(\delta_t, \mathcal{W}_t)$ is distributed as the selected step $\tilde{\mathbf{N}}_t^\star$.

PROPOSITION 1: *Let a $(1, \lambda)$-ES with resampling optimize the problem defined in Eq. (7), and let $(\mathbf{W}_t^i)_{i\in[1..\lambda],t\in\mathbb{N}}$ be an i.i.d. sequence of random vectors with $\mathbf{W}_t^i \sim (\mathcal{U}_{[0,1]}, \mathcal{N}(0,1))$, and $\mathcal{W}_t = (\mathbf{W}_t^i)_{i\in[1..\lambda]}$. Then*

$$\tilde{\mathbf{N}}_t^\star \overset{d}{=} \mathcal{G}(\delta_t, \mathcal{W}_t), \tag{17}$$

where the function $\mathcal{G}$ is defined in Eq. (16).

PROOF: Since f is a linear function $f(\tilde{\mathbf{Y}}_t^i) = f(\mathbf{X}_t) + \sigma_t f(\tilde{\mathbf{N}}_t^i)$, so $f(\tilde{\mathbf{Y}}_t^i) \leq f(\tilde{\mathbf{Y}}_t^j)$ is equivalent to $f(\tilde{\mathbf{N}}_t^i) \leq f(\tilde{\mathbf{N}}_t^j)$. Hence $\star = \operatorname{argmax}_{i\in[1..\lambda]} f(\tilde{\mathbf{N}}_t^i)$ and therefore $\tilde{\mathbf{N}}_t^\star = \operatorname{argmax}_{\mathbf{N}\in\{\tilde{\mathbf{N}}_t^i|i\in[1..\lambda]\}} f(\mathbf{N})$. From Lemma 2 and Eq. (15), $\tilde{\mathbf{N}}_t^i \overset{d}{=} \tilde{\mathcal{G}}(\delta_t, \mathbf{W}_t^i)$, so $\tilde{\mathbf{N}}_t^\star \overset{d}{=} \operatorname{argmax}_{\mathbf{N}\in\{\tilde{\mathcal{G}}(\delta_t,\mathbf{W}_t^i)|i\in[1..\lambda]\}} f(\mathbf{N})$, which from Eq. (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 $\varphi^* := \mathbf{E}([\mathbf{X}_{t+1}]_1 - [\mathbf{X}_t]_1) = \sigma\mathbf{E}([\tilde{\mathbf{N}}_t^\star]_1)$ (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, that is, $\varphi^* = \mathbf{E}(f(\mathbf{X}_{t+1}) - f(\mathbf{X}_t))$.

As suggested by Arnold (2011a), the sequence $(\delta_t)_{t\in\mathbb{N}}$ plays a central role for the analysis, and we 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*

$$\delta_{t+1} = \delta_t - \tilde{\mathbf{N}}_t^\star.\boldsymbol{n} \overset{d}{=} \delta_t - \mathcal{G}(\delta_t, \mathcal{W}_t).\boldsymbol{n}, \tag{18}$$

where $\mathcal{G}$ is the function defined in Eq. (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 $\delta_{t+1} = \frac{g(\mathbf{X}_{t+1})}{\sigma_{t+1}} = \frac{-(\mathbf{X}_t + \sigma\tilde{\mathbf{N}}_t^\star).\mathbf{n}}{\sigma} = \delta_t - \tilde{\mathbf{N}}_t^\star.\mathbf{n}$, and Proposition 1 states 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:

$$\begin{aligned}\frac{[\mathbf{X}_t-\mathbf{X}_0]_1}{t} &= \frac{1}{t}\sum_{k=0}^{t-1}([\mathbf{X}_{k+1}]_1-[\mathbf{X}_k]_1)\\ &= \frac{\sigma}{t}\sum_{k=0}^{t-1}[\tilde{\mathbf{N}}_k^\star]_1 \overset{d}{=} \frac{\sigma}{t}\sum_{k=0}^{t-1}[\mathcal{G}(\delta_k,\mathcal{W}_k)]_1. \end{aligned} \tag{19}$$

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 in turn implies—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, assuming the LLN can be applied, the following limit will hold

$$\frac{[\mathbf{X}_t-\mathbf{X}_0]_1}{t} \xrightarrow[t\to\infty]{a.s.} \sigma\int_{\mathbb{R}_+^*}\mathbf{E}\left([\mathcal{G}(\delta,\mathcal{W})]_1\right)\pi(d\delta). \tag{20}$$

If the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is also V-ergodic with $|\mathbf{E}([G(\delta,\mathcal{W})]_1)| \le V(\delta)$, 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 also underlies the study carried out by Arnold (2011a, Section 4), where 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 provide a formal proof that indeed $(\delta_t)_{t\in\mathbb{N}}$ admits a stationary limit distribution π, and prove some other useful properties that allow us to conclude 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 Eq. (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 now derive the first properties on the Markov chain $(\delta_t)_{t\in\mathbb{N}}$. First 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 nontrivial 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 by $\mathcal{B}^+(\mathbb{R}_+^*)$ 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 nontrivial 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 a probability measure α on $\mathbb{N}$ and a nontrivial measure μ_α such that for all sets $A \in \mathcal{B}(\mathbb{R}_+^*)$ and all $\delta \in C$,

$$K_\alpha(x, A) := \sum_{m \in \mathbb{N}} P^m(\mathbf{x}, A)\alpha(m) \geq \mu_\alpha(A). \tag{24}$$

A small set is therefore also a petite set. The existence of a small set combined with a control of the Markov chain outside the small set allows deducing powerful stability properties of the Markov chain. If there exists a ν_1-small set C such that $\nu_1(C) > 0$, then the Markov chain is called 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 Eq. (18). Then $(\delta_t)_{t\in\mathbb{N}}$ is $\mu_{\mathbb{R}_+^*}$-irreducible, strongly aperiodic, and compact sets of $\mathbb{R}_+^*$ and sets of the form $(0, M]$ with $M > 0$ are small sets.*

PROOF: Take $\delta \in \mathbb{R}_+^*$ and $A \in \mathcal{B}(\mathbb{R}_+^*)$. 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 $\mu_{\mathbb{R}_+^*}(A) > 0$, $P(\delta, A) > 0$. 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 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 $\delta \in C$,

$$P(\delta, A) \geq \underbrace{g_C \mu_{\mathbb{R}_+^*}(A \cap [0, 1])}_{:=\nu_C(A)}.$$

The measure ν_C being nontrivial, the previous equation shows that compact sets of $\mathbb{R}_+^*$ are small and that for C a compact such that $\mu_{\mathbb{R}_+^*}(C \cap [0, 1]) > 0$, 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, that is, $\eta_A = \sum_{t=1}^{\infty} 1_A(\delta_t)$. We 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

$$\Delta V(\delta) = \mathbf{E}[V(\delta_{t+1})|\delta_t = \delta] - V(\delta). \tag{27}$$

Drift conditions translate that outside a small set, the drift operator is negative. We 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

$$\sum_{t\in\mathbb{N}} r_f^t \| P^t(\delta, \cdot) - \pi \|_f < \infty, \forall \delta \in \mathbb{R}_+^*, \tag{28}$$

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 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 $\delta \in \mathbb{R}_+^*$,

$$\Delta V(\delta) \leq -\epsilon V(\delta) + b\mathbf{1}_C(\delta). \tag{29}$$

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.

LEMMA 4: *Consider the $(1, \lambda)$-ES with resampling optimizing the constrained problem (7), and let $\mathcal{G}$ be the function defined in Eq. (16). We denote by 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 $(a, b) \in \mathbb{R}^2$, $lim_{\delta\to+\infty} \boldsymbol{E}(K) = \boldsymbol{E}(exp(a\mathcal{N}_{\lambda:\lambda}))\boldsymbol{E}(exp(b\mathcal{N}(0, 1))) < \infty$ and $lim_{\delta\to+\infty} \boldsymbol{E}(\bar{K}) < \infty$.*

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 Eq. (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.*

[2]The condition $\pi(V) < \infty$ is given by Meyn and Tweedie (1993, Theorem 14.0.1).

[3]The function V of (29) is unbounded off small sets (Meyn and Tweedie, 1993, Lemma 15.2.2; 1993, Prop. 55.7); hence the Markov chain is Harris-recurrent (Meyn and Tweedie, 1993, Theorem 9.1.8).

PROOF: Take the function $V : \delta \mapsto \exp(\alpha\delta)$; then

$$\Delta V(\delta) = \mathbf{E}\left(\exp\left(\alpha(\delta - \mathcal{G}(\delta, \mathcal{W}).\mathbf{n})\right)\right) - \exp(\alpha\delta)$$
$$\frac{\Delta V}{V}(\delta) = \mathbf{E}\left(\exp\left(-\alpha\mathcal{G}(\delta, \mathcal{W}).\mathbf{n}\right)\right) - 1.$$

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

$$\lim_{\delta\to+\infty} \mathbf{E}\left(\exp\left(-\alpha\mathcal{G}(\delta, \mathcal{W}).\mathbf{n}\right)\right) = \left(\sum_{i\in\mathbb{N}} \frac{(-\alpha\cos\theta)^i\, \mathbf{E}\left(\mathcal{N}^i_{\lambda:\lambda}\right)}{i!}\right)\left(\sum_{i\in\mathbb{N}} \frac{(-\alpha\sin\theta)^i\, \mathbf{E}\left(\mathcal{N}(0,1)^i\right)}{i!}\right),$$

which in turns yields

$$\lim_{\delta\to+\infty} \frac{\Delta V}{V}(\delta) = (1 - \alpha\mathbf{E}(\mathcal{N}_{\lambda:\lambda})\cos\theta + o(\alpha))\,(1 + o(\alpha)) - 1$$
$$= -\alpha\mathbf{E}(\mathcal{N}_{\lambda:\lambda})\cos\theta + o(\alpha).$$

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, Prop. 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 from Meyn and Tweedie (1993) proves the proposition. □

We have proved rigorously the existence (and unicity) of an invariant measure π for the Markov chain $(\delta_t)_{t\in\mathbb{N}}$, which provides the steady state behavior described by 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 Eq. (18). The sequence $([\boldsymbol{X}_t]_1)_{t\in\mathbb{N}}$ diverges in probability to $+\infty$ at constant speed, that is,*

$$\frac{[\boldsymbol{X}_t - \boldsymbol{X}_0]_1}{t} \xrightarrow[t\to+\infty]{P} \sigma\boldsymbol{E}_{\pi\otimes\mu_{\mathcal{W}}}\left([\mathcal{G}(\delta, \mathcal{W})]_1\right) > 0, \tag{30}$$

and the expected progress satisfies

$$\varphi^* = \boldsymbol{E}\left([\boldsymbol{X}_{t+1} - \boldsymbol{X}_t]_1\right) \xrightarrow[t\to+\infty]{} \sigma\boldsymbol{E}_{\pi\otimes\mu_{\mathcal{W}}}\left([\mathcal{G}(\delta, \mathcal{W})]_1\right) > 0, \tag{31}$$

where φ^* is the progress rate defined by Arnold (2011a, Eq. (2)), $\mathcal{G}$ is defined in Eq. (16), $\mathcal{W} = (\boldsymbol{W}^i)_{i\in[1..\lambda]}$ with $(\boldsymbol{W}^i)_{i\in[1..\lambda]}$ an i.i.d. sequence such that $\boldsymbol{W}^i \sim (\mathcal{U}_{[0,1]}, \mathcal{N}(0, 1))$, π is the stationary measure of $(\delta_t)_{t\in\mathbb{N}}$ whose existence was proved 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 $\mathbf{E}_{\pi}(\mathbf{E}_{\mu_{\mathcal{W}}}(|[\mathcal{G}(\delta, \mathcal{W})]_1|))$. 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 Eq. (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 Eq. (19) allows us to deduce the convergence in probability of the left-hand side of Eq. (19) to the right-hand side of the previous equation.

From Eq. (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 Eq. (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

$$\mathbf{E}_{\pi\otimes\mu_{\mathcal{W}}}\left([\mathcal{G}(\delta, \mathcal{W})]_1\right) = -\tan\theta\,\mathbf{E}_{\pi\otimes\mu_{\mathcal{W}}}\left([\mathcal{G}(\delta, \mathcal{W})]_2\right).$$

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 this 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, giving 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 expressions (30) and (31) and for 10^6 time steps gives the progress rate of $\varphi^* = \mathbf{E}([\mathbf{X}_{t+1} - \mathbf{X}_t]_1)$ (Arnold, 2011a), which once normalized by σ and λ yields Figure 2. We normalize per λ, as in evolution strategies the cost of the algorithm is assumed to be the number of

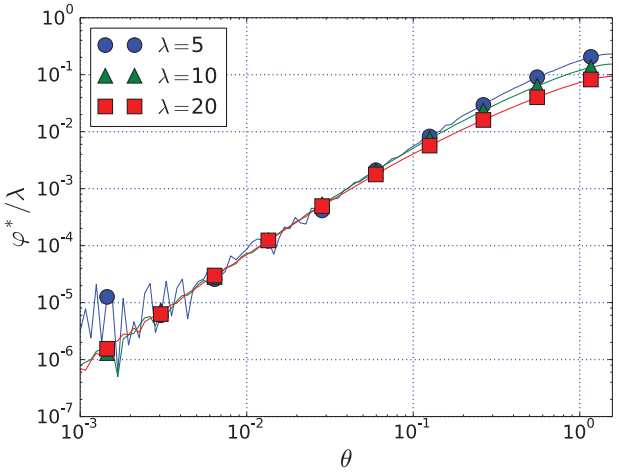

Figure 2: Normalized progress rate $\varphi^* = \mathbf{E}(f(\mathbf{X}_{t+1}) - f(\mathbf{X}_t))$ divided by λ for the $(1, \lambda)$-ES with constant step-size $\sigma = 1$ and resampling, plotted against the constraint angle θ, for $\lambda \in \{5, 10, 20\}$.

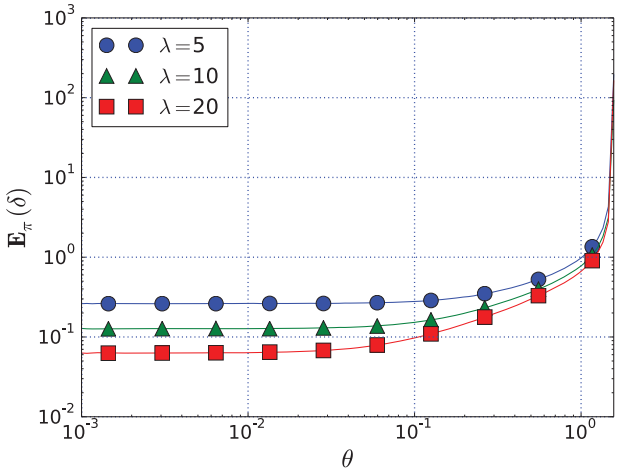

Figure 3: Average normalized distance δ from the constraint for the $(1, \lambda)$-ES with constant step-size and resampling, plotted against the constraint angle θ, for $\lambda \in \{5, 10, 20\}$.

f-calls. We see that for small values of θ, the normalized serial progress rate assumes roughly $\varphi^*/\lambda \approx \theta^2$. Only for larger constraint angles the serial progress rate depends on λ, where smaller λ are preferable.

Figure 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 $\mathbf{X}_t$ toward 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. This technique was studied on sphere functions by Arnold and Beyer (2004) and on ridge functions by Arnold and MacLeod (2008).

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

$$\mathbf{p}_{t+1} = (1-c)\mathbf{p}_t + \sqrt{c(2-c)}\tilde{\mathbf{N}}_t^\star. \tag{32}$$

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. (32) 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 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 by 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 (resp. lower) than n, then the step-size is increased (resp. decreased) following

$$\sigma_{t+1} = \sigma_t \exp\left(\frac{c}{2d_\sigma}\left(\frac{\|\mathbf{p}_{t+1}\|^2}{n} - 1\right)\right), \tag{33}$$

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 notation, Eq. (33) is rewritten as

$$\sigma_{t+1} = \sigma_t \exp\left(\frac{c}{2d_\sigma}\left(\frac{\|\mathbf{p}_{t+1}\|^2 + K_t}{n} - 1\right)\right), \tag{34}$$

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 denote by $\eta_c^\star$ the multiplicative step-size change σ_{t+1}/σ_t, which is the function

$$\eta_c^\star(\mathbf{p}_t, \delta_t, \mathcal{W}_t, K_t) = \exp\left(\frac{c}{2d_\sigma}\left(\frac{\|(1-c)\mathbf{p}_t + \sqrt{c(2-c)}\mathcal{G}(\delta_t, \mathcal{W}_t)\|^2 + K_t}{n} - 1\right)\right). \tag{35}$$

Note that for $c = 1$, $\eta_1^\star$ is a function only of δ_t, $\mathcal{W}_t$, and K_t, which we denote by $\eta_1^\star(\delta_t, \mathcal{W}_t, K_t)$.

We prove in the next proposition that for $c < 1$, the sequence $(\delta_t, \mathbf{p}_t)_{t\in\mathbb{N}}$ is a homogeneous Markov chain, and make 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 $\delta_t = g(\boldsymbol{X}_t)/\sigma_t$. The sequence $(\delta_t, \boldsymbol{p}_t)_{t\in\mathbb{N}}$ is a*

time-homogeneous Markov chain and

$$\delta_{t+1} \overset{d}{=} \frac{\delta_t - \mathcal{G}(\delta_t, \mathcal{W}_t).\boldsymbol{n}}{\eta_c^\star(\boldsymbol{p}_t, \delta_t, \mathcal{W}_t, K_t)}, \tag{36}$$

$$\boldsymbol{p}_{t+1} \overset{d}{=} (1-c)\boldsymbol{p}_t + \sqrt{c(2-c)}\mathcal{G}(\delta_t, \mathcal{W}_t), \tag{37}$$

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, $\mathcal{G}$ 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

$$\delta_{t+1} \overset{d}{=} \frac{\delta_t - \mathcal{G}(\delta_t, \mathcal{W}_t).\boldsymbol{n}}{exp\left(\frac{c}{2d_\sigma}\left(\frac{\|\mathcal{G}(\delta_t, \mathcal{W}_t)\|^2}{n} - 1\right)\right)}. \tag{38}$$

PROOF: With Eq. (32) and Eq. (17) we get Eq. (37).

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 Eqs. (36) and (37) immediately yields Eq. (38), 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. (34) we can express $\ln(\sigma_t/\sigma_0)/t$ as

$$\frac{1}{t}\ln\frac{\sigma_t}{\sigma_0} = \frac{c}{2d_\sigma}\left(\frac{\frac{1}{t}\left(\sum_{i=0}^{t-1}\|\mathbf{p}_{i+1}\|^2 + K_i\right)}{n} - 1\right). \tag{39}$$

The right-hand side suggests we 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 Eq. (39) becomes

$$\frac{1}{t}\ln\frac{\sigma_t}{\sigma_0} \overset{d}{=} \frac{c}{2d_\sigma}\left(\frac{\frac{1}{t}\sum_{i=0}^{t-1}\|\mathcal{G}(\delta_i, \mathcal{W}_i)\|^2 + K_i}{n} - 1\right). \tag{40}$$

To apply the LLN we need the Markov chain to be Harris-positive and the properties mentioned in the following lemma.

LEMMA 5 (CHOTARD AND AUGER 2015, PROP. 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 the set $(0, M] \times B(\mathbf{0}, r)$ for $M > 0$ and $r > 0$ small enough.

To show the Harris positivity of $(\delta_t)_{t\in\mathbb{N}}$ we use the drift function $V : \delta \in \mathbb{R}_+^* \mapsto \delta^\alpha + \delta^{-\alpha}$. From the definition of the drift operator ΔV in Eq. (27) and the update of δ_t in Eq. (38), we have

$$\Delta V(\delta) = \mathbf{E}\left(\frac{(\delta - \mathcal{G}(\delta, \mathcal{W}).\mathbf{n})^\alpha}{\eta_1^\star(\delta, \mathcal{W}, K)^\alpha}\right) + \mathbf{E}\left(\frac{\eta_1^\star(\delta, \mathcal{W}, K)^\alpha}{(\delta - \mathcal{G}(\delta, \mathcal{W}).\mathbf{n})^\alpha}\right) - \delta^\alpha - \delta^{-\alpha}. \tag{41}$$

To verify the drift condition (29), using the fact from Lemma 5 that for $0 < m < M$ the compact $[m, M]$ is a small set, it is sufficient to show that the limits of $\Delta V / V(\delta)$ in 0 and ∞ are negative. These limits result from the limits studied in the following lemma corresponding to the decomposition in Eq. (41).

LEMMA 6: *For $\alpha > 0$ small enough*

$$\frac{1}{\delta^\alpha + \delta^{-\alpha}} \boldsymbol{E}\left(\frac{\left(\delta - \mathcal{G}(\delta, \mathcal{W}).\boldsymbol{n}\right)^\alpha}{\eta_1^\star(\delta, \mathcal{W}, K)^\alpha}\right) \underset{\delta\to+\infty}{\longrightarrow} E_1 E_2 E_3 < \infty \tag{42}$$

$$\frac{1}{\delta^\alpha + \delta^{-\alpha}} \boldsymbol{E}\left(\frac{\left(\delta - \mathcal{G}(\delta, \mathcal{W}).\boldsymbol{n}\right)^\alpha}{\eta_1^\star(\delta, \mathcal{W}, K)^\alpha}\right) \underset{\delta\to 0}{\longrightarrow} 0 \tag{43}$$

$$\frac{1}{\delta^\alpha + \delta^{-\alpha}} \boldsymbol{E}\left(\frac{\eta_1^\star(\delta, \mathcal{W}, K)^\alpha}{\left(\delta - \mathcal{G}(\delta, \mathcal{W}).\boldsymbol{n}\right)^\alpha}\right) \underset{\delta\to+\infty}{\longrightarrow} 0 \tag{44}$$

$$\frac{1}{\delta^\alpha + \delta^{-\alpha}} \boldsymbol{E}\left(\frac{\eta_1^\star(\delta, \mathcal{W}, K)^\alpha}{\left(\delta - \mathcal{G}(\delta, \mathcal{W}).\boldsymbol{n}\right)^\alpha}\right) \underset{\delta\to 0}{\longrightarrow} 0, \tag{45}$$

where $E_1 = \boldsymbol{E}(exp(-\frac{\alpha}{2d_\sigma n}(\mathcal{N}_{\lambda:\lambda}^2 - 1)))$, $E_2 = \boldsymbol{E}(exp(-\frac{\alpha}{2d_\sigma n}(\mathcal{N}(0,1)^2 - 1)))$, and $E_3 = \boldsymbol{E}(exp(-\frac{\alpha}{2d_\sigma n}(K - (n-2))))$; where $\mathcal{G}$ is the function defined in Eq. (16) and $\eta_1^\star$ is defined in Eq. (35) (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 (see appendix).

We now prove the Harris positivity of $(\delta_t)_{t\in\mathbb{N}}$ by proving a stronger property, namely, the geometric ergodicity, which we show using the drift inequality (29).

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 is 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. (41) we have the following drift quotient

$$\frac{\Delta V(\delta)}{V(\delta)} = \frac{1}{V(\delta)}\mathbf{E}\left(\frac{(\delta - \mathcal{G}(\delta, \mathcal{W}).\mathbf{n})^\alpha}{\eta_1^\star(\delta, \mathcal{W}, K)^\alpha}\right) + \frac{1}{V(\delta)}\mathbf{E}\left(\frac{\eta_1^\star(\delta, \mathcal{W}, K)^\alpha}{(\delta - \mathcal{G}(\delta, \mathcal{W}).\mathbf{n})^\alpha}\right) - 1, \tag{46}$$

with $\eta_1^\star$ defined in Eq. (35) and $\mathcal{G}$ in Eq. (16). From Lemma 6, following the same notation as 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 $\lambda\varphi(x)\Phi(x)^{\lambda-1}$ and $\exp|\alpha/(2d_\sigma n)(x^2 - 1)|\lambda\varphi(x)\Phi(x)^{\lambda-1} \leq \lambda\exp(1/(2d_\sigma n))\exp(\alpha/(2d_\sigma n)x^2 - x^2/2)$, which for α small enough is integrable, we get

$$\mathbf{E}\left(\sum_{k\in\mathbb{N}} \frac{\left|-\frac{\alpha}{2d_\sigma n}\left(\mathcal{N}_{\lambda:\lambda}^2 - 1\right)\right|^k}{k!} \right) = \int_{\mathbb{R}} \exp\left|\frac{\alpha}{2d_\sigma n}\left(x^2 - 1\right)\right| \lambda\varphi(x)\Phi(x)^{\lambda-1}\mathrm{d}x < \infty.$$

Therefore 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 $\alpha/(2d_\sigma n)x - x/2 \leq 0$ for all $x \geq 0$), 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 Eq. (46), $\lim_{\delta\to 0} \Delta V(\delta)/V(\delta) = -1$. 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 $b \in \mathbb{R}$ 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 because 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 on 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 into 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 Propositions 4 and 6 we showed 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

because they have different update rules (compare Eq. (18) and Eq. (36)). The chain $(\delta_t)_{t\in\mathbb{N}}$ being Harris-positive, we may now apply LLN to Eq. (40) 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*

$$\frac{1}{t} ln\left(\frac{\sigma_t}{\sigma_0}\right) \xrightarrow[t\to\infty]{P} \frac{1}{2d_\sigma n}\left(\boldsymbol{E}_{\pi_1\otimes\mu_{\mathcal{W}}}\left(\|\mathcal{G}(\delta,\mathcal{W})\|^2\right)-2\right), \tag{47}$$

and in expectation

$$\boldsymbol{E}\left(ln\left(\frac{\sigma_{t+1}}{\sigma_t}\right)\right) \xrightarrow[t\to+\infty]{} \frac{1}{2d_\sigma n}\left(\boldsymbol{E}_{\pi_1\otimes\mu_{\mathcal{W}}}\left(\|\mathcal{G}(\delta,\mathcal{W})\|^2\right)-2\right) \tag{48}$$

with $\mathcal{G}$ defined in Eq. (16) and $\mathcal{W} = (\boldsymbol{W}^i)_{i\in[1..\lambda]}$, where $(\boldsymbol{W}^i)_{i\in[1..\lambda]}$ is an i.i.d. sequence such that $\boldsymbol{W}^i \sim (\mathcal{U}_{[0,1]}, \mathcal{N}(0,1))$, $\mu_{\mathcal{W}}$ 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 $f(\boldsymbol{X}_{t+1}) - f(\boldsymbol{X}_t)$ diverges or converges geometrically in probability

$$\frac{1}{t} ln\left|\frac{f(\boldsymbol{X}_{t+1})-f(\boldsymbol{X}_t)}{\sigma_0}\right| \xrightarrow[t\to\infty]{P} \frac{1}{2d_\sigma n}\left(\boldsymbol{E}_{\pi_1\otimes\mu_{\mathcal{W}}}\left(\|\mathcal{G}(\delta,\mathcal{W})\|^2\right)-2\right). \tag{49}$$

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 (Meyn and Tweedie, 1993, Theorem 17.0.1) to Eq. (39) 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 $\mathbf{E}_{\pi_1\times\mu_{\mathcal{W}}}(\|G(\delta,\mathcal{W})\|^2) + n - 2$. 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 $\pi_1(\mathbb{R}) = 1$), $\mathbf{E}_{\pi_1}(\mathbf{E}_{\mu_{\mathcal{W}}}(\|\mathcal{G}(\delta,\mathcal{W})\|^2|\delta_t = \delta)) \le M_{\mathcal{G}^2} < \infty$. 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. (40) yields Eq. (47).

From Proposition 1, Eq. (32) for $c = 1$, and Eq. (34), $\ln(\sigma_{t+1}/\sigma_t) \overset{d}{=} 1/(2d_\sigma n)(\|\mathcal{G}(\delta_t,\mathcal{W}_t)\|^2 + \chi^2_{n-2} - n)$, 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 $\|P^t(\delta,\cdot) - \pi_1\|_V \le M_\delta r^{-t}$. We showed that the function $\delta \mapsto \mathbf{E}(\|\mathcal{G}(\delta,\mathcal{W})\|^2)$ is bounded, so since $V(\delta) \ge 1$ for all $\delta \in \mathbb{R}^*_+$, there exists k such that $\mathbf{E}_{\mu_{\mathcal{W}}}(\|\mathcal{G}(\delta,\mathcal{W})\|^2) \le kV(\delta)$ 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. (48).

For expression (49) 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 Eq. (13), since $1/2 \leq \Phi(x) \leq 1$ for all $x \geq 0$ and $F_{1,\delta}(x) \leq 1$, 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 converge 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 (49) we get (49). □

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 right-hand side of Eq. (47) is strictly positive, then the step-size diverges geometrically. The Law of Large Numbers entails that Monte Carlo simulations will converge to the right-hand side of Eq. (47), 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 addressed 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 $t \to +\infty$ appears to converge to 0 for low enough values of d_σ and θ.

As in Figure 3 we simulate the Markov chain $(\delta_t, \mathbf{p}_t)_{t\in\mathbb{N}}$ defined in Eq. (36) to obtain Figure 4 after an average of δ_t over 10^6 time steps. Assuming that the Markov chain $(\delta_t, \mathbf{p}_t)_{t\in\mathbb{N}}$ admits an invariant probability measure π_c, the expected value $\mathbf{E}_{\pi_c}(\delta)$ shows the same dependency in λ as in the constant case. With larger population size, the algorithm follows the constraint closer, because 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 that $\mathbf{E}_{\pi_c}(\ln(\sigma_{t+1}/\sigma_t)) > 0$ (see Figure 6).

In Figure 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 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 Figure 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 by Arnold (2011a), where π is used instead of π_c to estimate $\ln(\sigma_{t+1}/\sigma_t)$, is accurate for not too large values of the constraint angle.

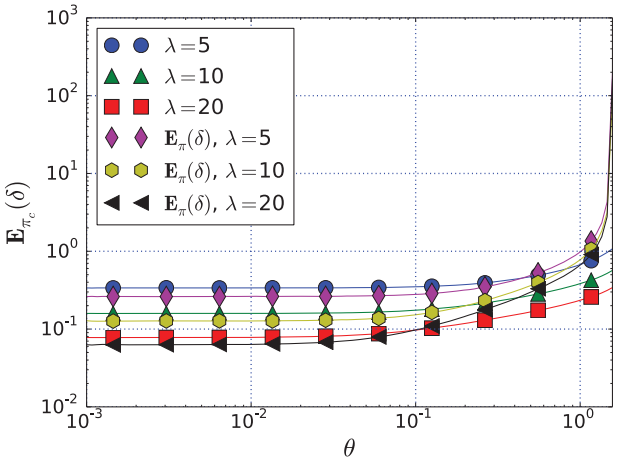

Figure 4: Average normalized distance δ from the constraint for the $(1, \lambda)$-CSA-ES, plotted against the constraint angle θ, for $\lambda \in \{5, 10, 20\}$, $c = 1/\sqrt{2}$, $d_\sigma = 1$, and dimension 2.

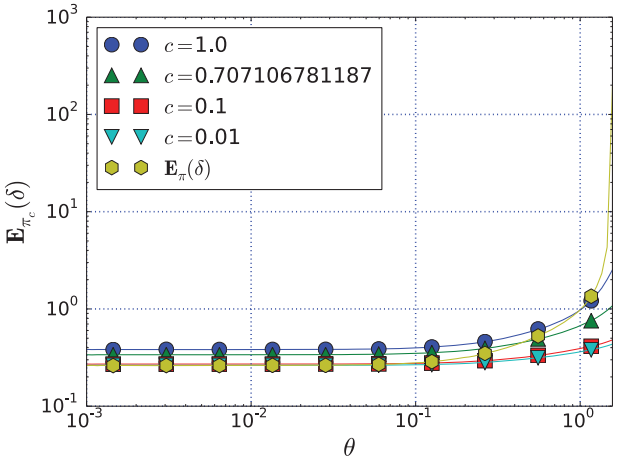

Figure 5: Average normalized distance δ from the constraint for the $(1, \lambda)$-CSA-ES, plotted against the constraint angle θ, with $c \in \{1, 1/\sqrt{2}, 0.1, 0.01\}$ and for constant step-size, where $\lambda = 5$, $d_\sigma = 1$, and dimension 2.

In Figure 6, corresponding to the left-hand side of Eq. (47), the adaptation response $\Delta_t := \ln(\sigma_{t+1}/\sigma_t)$ is averaged over 10^6 time steps and plotted against the constraint angle θ for different population sizes. If the 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 to make the step-size adaptation to succeed for a wider interval of values of θ.

In Figure 7, as in Figure 6, the adaptation response Δ_t is averaged for 10^6 time steps and plotted 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

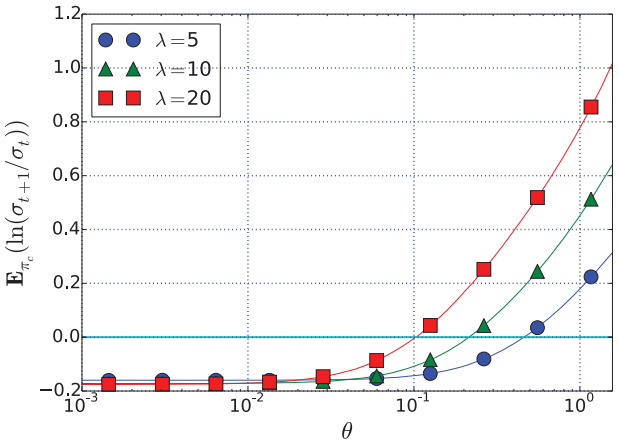

Figure 6: Average of the logarithmic adaptation response $\Delta_t = \ln(\sigma_{t+1}/\sigma_t)$ for the $(1, \lambda)$-CSA-ES, plotted against the constraint angle θ, for $\lambda \in \{5, 10, 20\}$, $c = 1/\sqrt{2}$, $d_\sigma = 1$, and dimension 2. Values below zero (straight line) indicate premature convergence.

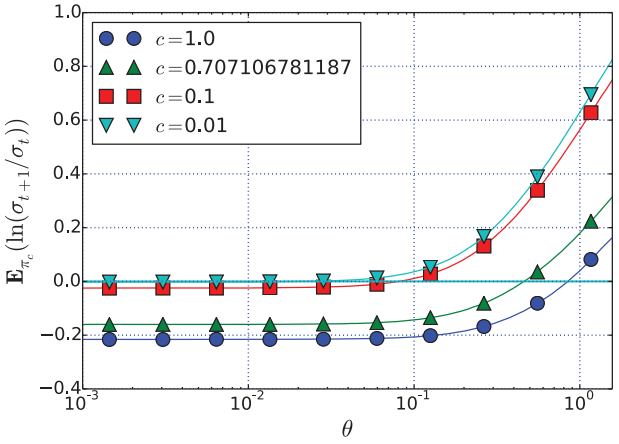

Figure 7: Average of the logarithmic adaptation response $\Delta_t = \ln(\sigma_{t+1}/\sigma_t)$ for the $(1, \lambda)$-CSA-ES, plotted against the constraint angle θ, for $\lambda = 5$, $c \in \{1, 1/\sqrt{2}, 0.1, 0.01\}$, $d_\sigma = 1$, and dimension 2. Values below zero (straight line) indicate premature convergence.

step-size, although $\mathbf{E}_{\pi_c}(\ln(\sigma_{t+1}/\sigma_t))$ appears to converge quickly to an asymptotic constant when $\ln(c) \to -\infty$. Lower values of c also allow success of the step-size adaptation for wider range values of θ, and in the case of premature convergence a lower value of c means a lower convergence rate.

In Figure 8 the adaptation response Δ_t is averaged 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 low enough value of d_σ implies geometric divergence of the step-size regardless of the constraint angle. However, simulations suggest that while for $d_\sigma \geq 1$ the probability that $f(\mathbf{X}_t) > f(\mathbf{X}_0)$ is close to 1, this probability decreases with

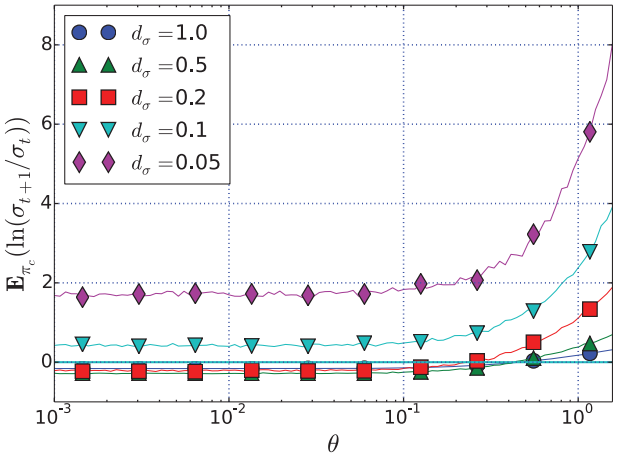

Figure 8: Average of the logarithmic adaptation response $\Delta_t = \ln(\sigma_{t+1}/\sigma_t)$ for the $(1, \lambda)$-CSA-ES, plotted against the constraint angle θ, for $\lambda = 5$, $c = 1/\sqrt{2}$, $d_\sigma \in \{1, 0.5, 0.2, 0.1, 0.05\}$, and dimension 2. Values below zero (straight line) indicate premature convergence.

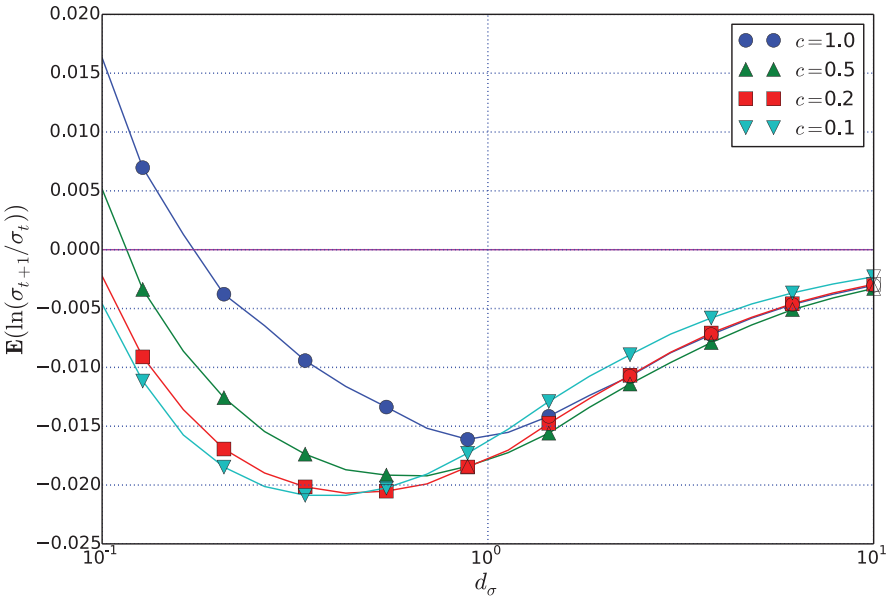

Figure 9: Average of the logarithmic adaptation response $\Delta_t = \ln(\sigma_{t+1}/\sigma_t)$ against d_σ for the $(1, \lambda)$-CSA-ES minimizing a sphere function for $\lambda = 5$, $c \in \{1, 0.5, 0.2, 0.1\}$, and dimension 30.

smaller values of d_σ. A low value of d_σ will also prevent convergence when it is desired, as shown in Figure 9.

In Figure 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

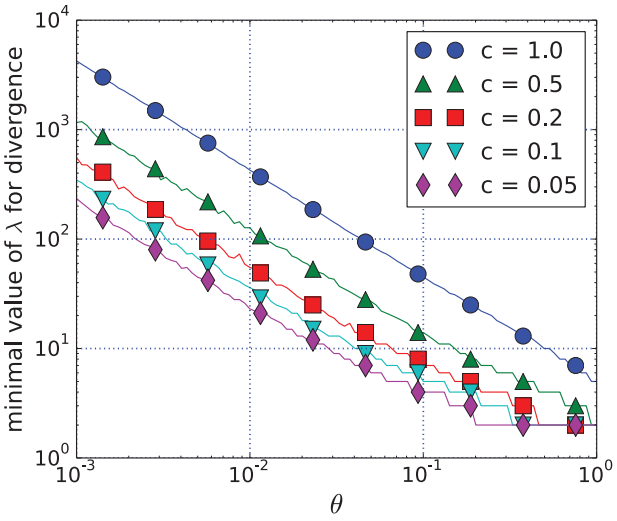

Figure 10: Minimal value of λ allowing geometric divergence for the (1, λ)-CSA-ES, plotted against the constraint angle θ, for $c \in \{1., 0.5, 0.2, 0.05\}$, $d_\sigma = 1$, and dimension 2.

dimension 30, averaged over 10 runs. Low values of d_σ make the algorithm diverge, while convergence is desired here.

In Figure 10 the smallest population size allowing geometric divergence on the linear constrained function 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 $c = 1$) or 20 (for the other values of c). If the step-size had decreased (resp. increased), this value of λ became a lower (resp. upper) bound for λ, and a larger (resp. smaller) value of λ was tested until the estimated upper and lower bounds for λ met. 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, as expected.

In Figure 11 the largest 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 as with Figure 10. For a certain θ the algorithm was run with a certain value of c, and when the logarithm of the step-size increased (resp. decreased) by more than $1{,}000\sqrt{c}$, the run was stopped, the value of c tested became the new lower (resp. upper) bound for c, and a new c taken between the lower and upper bounds was tested, until the lower and upper bounds were distant by less than the precision $\theta^2/10$. As with λ, simulations suggest that for small enough values of c, the probability that $\lim_{t\to+\infty} f(\mathbf{X}_t) > f(\mathbf{X}_0)$ is equal to 1, so small enough values of c appear to solve the linear function on this constrained problem.

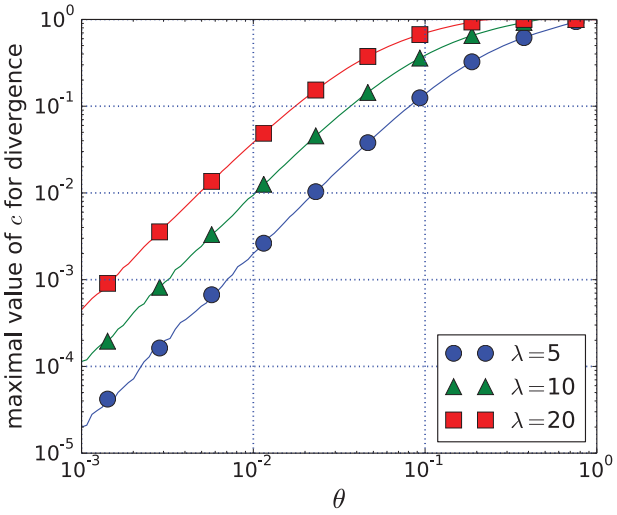

Figure 11: Transition boundary for c between convergence and divergence (lower value of c is divergence) for the $(1, \lambda)$-CSA-ES, plotted against the constraint angle θ, for $\lambda \in \{5, 10, 20\}$ and dimension 2.

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 infeasible 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 normalized distance to the constraint, which was presumed by 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 with constraint angle $\theta \to 0$ following $c \propto \theta^2$ or $\lambda \propto 1/\theta$. Smaller values of the constraint angle seem to increase the difficulty of the problem arbitrarily, that is, no given values for c and λ solve the problem for *every* $\theta \in (0, \pi/2)$. However, when using a repair method to handle the constraint instead of resampling with 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). Therefore, adaptation of the covariance matrix may render the problem arbitrarily close to the most simple 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 was done by

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 analyzing ES, complementing particularly well previous studies (Arnold, 2011a, 2012; Arnold and Brauer, 2008).

Acknowledgments

This work was supported by grants ANR-2010-COSI-002 (SIMINOLE) and ANR-2012-MONU-0009 (NumBBO) of the French National Research Agency.

References

Arnold, D. V. (2002). *Noisy optimization with evolution strategies*. Norwell, MA: Kluwer.

Arnold, D. V. (2011a). On the behaviour of the (1,λ)-ES for a simple constrained problem. In *Proceedings of the Conference on Foundations of Genetic Algorithms*, pp. 15–24.

Arnold, D. V. (2011b). Analysis of a repair mechanism for the $(1, \lambda)$-ES applied to a simple constrained problem. In *Proceedings of the Conference on Genetic and Evolutionary Computation (GECCO)*, pp. 853–860.

Arnold, D. V. (2012). On the behaviour of the $(1, \lambda)$-σSA-ES for a constrained linear problem. In *Parallel Problem Solving from Nature*, pp. 82–91.

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 Brauer, D. (2008). On the behaviour of the (1+1)-ES for a simple constrained problem. In *Parallel Problem Solving from Nature*, pp. 1–10.

Arnold, D. V., and Hansen, N. (2012). A (1+1)-CMA-ES for constrained optimisation. In *Proceedings of the Conference on Evolutionary Computation (GECCO)*, pp. 297–304.

Arnold, D. V., and MacLeod, A. (2008). Step length adaptation on ridge functions. *Evolutionary Computation*, 16(2): 151–184.

Arnold, D. V., and Porter, J. (2015). Towards an augmented lagrangian constraint handling approach for the (1+1)-es. In *Proceedings of the Conference on Genetic and Evolutionary Computation Conference (GECCO)*, pp. 249–256.

Chotard, A., and Auger, A. (2015). Verifiable conditions for irreducibility, aperiodicity and T-chain property of a general Markov chain. Retrieved from arXiv:1508.01644.

Chotard, A., Auger, A., and Hansen, N. (2012a). Cumulative step-size adaptation on linear functions. In *Parallel Problem Solving from Nature*, pp. 72–81.

Chotard, A., Auger, A., and Hansen, N. (2012b). Cumulative step-size adaptation on linear functions. Technical Report, Inria.

Chotard, A., Auger, A., and Hansen, N. (2014). Markov chain analysis of evolution strategies on a linear constraint optimization problem. In *Proceedings of the IEEE Congress on Evolutionary Computation*, pp. 159–166.

Chotard, A., and Holena, M. (2014). A generalized Markov-chain modelling approach to $(1, \lambda)$-es linear optimization. In *Parallel Problem Solving from Nature*, pp. 902–911.

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*, 2nd ed. Cambridge: Cambridge University Press.

Mezura-Montes, E., and Coello, C.A.C. (2008). Constrained optimization via multiobjective evolutionary algorithms. In J. Knowles, D. Corne, and K. Deb (Eds.), *Multiobjective problem solving from nature*, pp. 53–75. New York: 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. *IEEE Transactions on Evolutionary Computation*, 4(3): 284–294.

Appendix

PROOF OF LEMMA 4: From Proposition 1 and Lemma 3, the density probability function of $\mathcal{G}(\delta, \mathcal{W})$ is $p^\star_\delta$, and from Eq. (12),

$$p^\star_\delta\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) = \lambda \frac{\varphi(x)\varphi(y)\mathbf{1}_{\mathbb{R}^*_+}\left(\delta - \begin{pmatrix} x \\ y \end{pmatrix}.\mathbf{n}\right)}{\Phi(\delta)} F_{1,\delta}(x)^{\lambda-1},$$

where $F_{1,\delta}$ is the cumulative density function of $[\mathcal{G}(\delta, \mathcal{W})]_1$, whose probability density function is $p_{1,\delta}$. 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 $p_{1,\delta}(x) < 2\varphi(x)$. Thus $p_{1,\delta}(x)$ 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 $\mathbf{E}(\exp(a\mathcal{N}_{\lambda:\lambda}))\mathbf{E}(\exp(b\mathcal{N}(0, 1)))$; and this quantity is finite.

The same reasoning can be applied to $E(\bar{K})$. □

PROOF OF LEMMA 6: 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 $p^\star_\delta \varphi_\chi$. With the theorem of Fubini-Tonelli the expected value of the random variable $\frac{(\delta - \mathcal{G}(\delta, \mathcal{W}_t).\mathbf{n})^\alpha}{\eta_1^\star(\delta, \mathcal{W}, K)^\alpha}$, which we denote by 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 $x \in \mathbb{R}$, $F_{1,\delta}(x) \leq 1$, and with Eqs. (9) and (12),

$$p^\star_\delta(x, y) = \lambda \frac{\varphi(x)\varphi(y)\mathbf{1}_{\mathbb{R}^*_+}(\delta - ax - by)}{\Phi(\delta)} F_{1,\delta}(x)^{\lambda-1} \leq \lambda \frac{\varphi(x)\varphi(y)}{\Phi(0)}, \tag{50}$$

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

$$p^\star_\delta((x, y))\frac{h_\delta(x, y)}{\delta^\alpha} \leq \lambda \frac{\varphi(x)\varphi(y)}{\Phi(0)} \frac{(1 + |ax + by|)^\alpha}{\exp\left(\frac{\alpha}{2d_\sigma n}\left(x^2 + y^2 - 2\right)\right)} \quad \text{for } \delta \geq 1, \text{ and} \tag{51}$$

$$p^\star_\delta((x, y))\frac{h_\delta(x, y)}{\delta^\alpha} \underset{\delta\to+\infty}{\longrightarrow} \varphi_{\mathcal{N}_{\lambda:\lambda}}(x)\varphi(y) \frac{1}{\exp\left(\frac{\alpha}{2d_\sigma n}\left(x^2 + y^2 - 2\right)\right)}. \tag{52}$$

Since the right-hand side of (51) is integrable, we can use Lebesgue's dominated convergence theorem and deduce from (52) that

$$\frac{E_\delta}{\delta^\alpha} = \int_{\mathbb{R}}\int_{\mathbb{R}} \frac{h_\delta(x, y)}{\delta^\alpha} p^\star_\delta((x, y))\mathrm{d}y\mathrm{d}x E_3 \underset{\delta\to+\infty}{\longrightarrow} \int_{\mathbb{R}}\int_{\mathbb{R}} \frac{\varphi_{\mathcal{N}_{\lambda:\lambda}}(x)\varphi(y)}{\exp\left(\frac{\alpha}{2d_\sigma n}\left(x^2 + y^2 - 2\right)\right)} \mathrm{d}y\mathrm{d}x E_3$$

$$\text{and so} \quad \frac{E_\delta}{\delta^\alpha} \underset{\delta\to+\infty}{\longrightarrow} E_1 E_2 E_3 < \infty.$$

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 $\delta \to +\infty$.

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 Eq. (50) we have

$$\delta^\alpha h_\delta(x, y) p^\star_\delta((x, y)) \leq \lambda \frac{(1 + |ax + by|)^\alpha \varphi(x)\varphi(y)}{\Phi(0)\exp\left(\frac{\alpha}{2d_\sigma n}\left(x^2 + y^2 - 2\right)\right)} \quad \text{for } 0 < \delta \leq 1. \tag{53}$$

The right-hand side of (53) 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 the expected value of the random variable $\frac{\eta_1^\star(\delta,\mathcal{W},K)^\alpha}{(\delta - \mathcal{G}(\delta,\mathcal{W}_t).\mathbf{n})^\alpha}$; then

$$H_\delta = \int_{\mathbb{R}}\int_{\mathbb{R}}\int_{\mathbb{R}} \frac{p_\delta^\star((x,y))\varphi_\chi(z)\exp\left(\frac{\alpha}{2d_\sigma n}(z-(n-2))\right)}{h_\delta(x,y)} \mathrm{d}z\mathrm{d}y\mathrm{d}x.$$

Integrating over z yields $H_\delta = \int_{\mathbb{R}}\int_{\mathbb{R}} \frac{p_\delta^\star((x,y))}{h_\delta(x,y)} \mathrm{d}y\mathrm{d}x H_3$.

We now study the limit when $\delta \to +\infty$ of H_δ/δ^α. With Eq. (50), we have that

$$\frac{p_\delta^\star((x,y))}{\delta^\alpha h_\delta(x,y)} \leq \lambda \frac{\varphi(x)\varphi(y)}{\Phi(0)} \frac{\exp\left(\frac{\alpha}{2d_\sigma n}\left(x^2+y^2-2\right)\right)}{\delta^\alpha |\delta - ax - by|^\alpha}.$$

With the change of variables $\tilde{x} = x - \delta/a$ we get

$$\begin{aligned}
\frac{p_\delta^\star((\tilde{x}+\frac{\delta}{a},y))}{\delta^\alpha h_\delta(\tilde{x}+\frac{\delta}{a},y)} &\leq \lambda \frac{\exp\left(-\frac{(\tilde{x}+\frac{\delta}{a})^2}{2}\right)\varphi(y)}{\sqrt{2\pi}\Phi(0)} \frac{\exp\left(\frac{\alpha}{2d_\sigma n}\left(\left(\tilde{x}+\frac{\delta}{a}\right)^2+y^2-2\right)\right)}{\delta^\alpha |a\tilde{x}+by|^\alpha} \\
&\leq \lambda \frac{\varphi(\tilde{x})\varphi(y)}{\Phi(0)} \frac{\exp\left(\frac{\alpha}{2d_\sigma n}\left(\tilde{x}^2+y^2-2\right)\right)}{|a\tilde{x}+by|^\alpha} \frac{\exp\left(\left(\frac{\alpha}{2d_\sigma n}-\frac{1}{2}\right)\left(2\frac{\delta}{a}\tilde{x}+\frac{\delta^2}{a^2}\right)\right)}{\delta^\alpha} \\
&\leq \lambda \frac{\varphi(\tilde{x})\varphi(y)}{\Phi(0)} \frac{1}{h_0(\tilde{x},y)} \frac{\exp\left(\left(\frac{\alpha}{2d_\sigma n}-\frac{1}{2}\right)\left(2\frac{\delta}{a}\tilde{x}+\frac{\delta^2}{a^2}\right)\right)}{\exp(\alpha\ln(\delta))}.
\end{aligned}$$

An upper bound for all $\delta \in \mathbb{R}_+^*$ of the right-hand side of the previous inequation is a function of 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 $H_{\delta}/\delta^{\alpha}$ 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 $(0, 1]$. As 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 a function of an upper bound of the function $k : \delta \in (0, 1] \mapsto 2\delta\tilde{x}/a + \delta^2/a^2$. This upper bound 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 upper bound of k is realized 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 $\delta \to 0$. □