Markov Chain Analysis of Evolution Strategies on a Linear Constraint Optimization Problem

Alexandre Chotard, Anne Auger, Nikolaus Hansen

Markov Chain Analysis of Evolution Strategies on a Linear Constraint Optimization Problem

Alexandre Chotard, Anne Auger and Nikolaus Hansen

***Abstract*— This paper analyses a $(1, \lambda)$-Evolution Strategy, a randomised comparison-based adaptive search algorithm, on a simple constraint optimization problem. The algorithm uses resampling to handle the constraint and optimizes a linear function with a linear 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 path length control. We exhibit for each case a Markov chain whose stability analysis would allow us to deduce the divergence of the algorithm depending on its internal parameters. We show divergence at a constant rate when the step-size is constant. We sketch that with step-size adaptation geometric divergence takes place. Our results complement previous studies where stability was assumed.**

I. Introduction

Derivative Free Optimization (DFO) methods are tailored for the optimization of numerical problems in a black-box context, where the algorithms can only query the objective function to optimize $f : \mathbb{R}^n \to \mathbb{R}$, and no properties on f, such as convexity or differentiability, is exploited.

Evolution Strategies (ES) are comparison-based randomised DFO algorithms. At iteration t, solutions are sampled from a multivariate normal distribution centered in a vector $\mathbf{X}_t$. The candidate solutions are ranked according to f, and update of $\mathbf{X}_t$ and other parameters of the distribution (usually a step-size σ_t and a covariance matrix) is performed using the ranking information given by the candidate solutions. Since ES do not directly use the function values of the new points, but only how f ranks the different samples, they are invariant to the composition of the objective function by a strictly increasing function $h : \mathbb{R} \to \mathbb{R}$.

This property and the black-box scenario make Evolution Strategies suited for a wide class of real-world problems, where constraints on the variables are often given. Different techniques for handling constraints in randomised algorithms have been proposed, see [6] for a survey. For ES, common techniques are resampling, i.e. resample a solution till it lies in the feasible domain, repair of solutions that project unfeasible points onto the feasible domain (e.g. [1]), penalty methods where unfeasible solutions are penalised either by a quantity that depends on the distance to the constraint (e.g. [7] with adaptive penalty weights) (if this latter one can be computed) or by the constraint value itself (e.g. stochastic ranking [11]) or methods inspired from multi-objective optimization (e.g. [10]).

Alexandre Chotard, Anne Auger and Nikolaus Hansen work in TAO, at INRIA-Saclay and LRI in University Paris-Sud, France (mail at name@lri.fr).

Our thanks to Dirk Arnold for suggesting this work during PPSN is Sicily.

In this paper we focus on the resampling method and study it on a simple constraint problem. More precisely, we study a $(1, \lambda)$-ES optimizing a linear function with a linear constraint and resampling any unfeasible solution until a feasible solution is sampled. The linear function models the situation where the current point is, relatively to the step-size, far from the optimum and "solving" this function means diverging. The linear constraint models being close to the constraint relatively to the step-size and far from other constraints. Due to the invariance of the algorithm to the composition of the objective function by a strictly increasing map, the linear function could be composed by a function without derivative and with many discontinuities without any impact on our analysis.

The problem we address was studied previously for different step-size adaptation mechanisms: with constant step-size, self-adaptation and cumulative step-size adaptation [2], [3]. The drawn conclusion is that when adapting the step-size the $(1, \lambda)$-ES fails to diverge unless some requirements on internal parameters of the algorithm are met. However, the approach followed in the aforementioned studies relies on finding simplified theoretical models to explain the behaviour of the algorithm: typically those models arise by doing some approximations (considering some random variables equal to their expected value, ...) and assuming some mathematical properties like the existence of stationary distributions of underlying Markov chains.

In contrast, our motivation is to study the real–in the sense not simplified–algorithm and prove rigorously different mathematical properties of the algorithm allowing to deduce the exact behaviour of the algorithm, as well as to provide tools and methodology for such studies. Our theoretical studies need to be complemented by simulations of the convergence/divergence rates. The mathematical properties that we derive show that these numerical simulations converge fast.

As for the step-size adaptation mechanism, our aim is to study the cumulative step-size adaptation (CSA), default step-size mechanism for the CMA-ES algorithm [8]. The mathematical object to study for this purpose is a discrete time, continuous state space Markov chain that is defined as the couple: evolution path and normalized distance to the constraint. More precisely stability properties like irreducibility, existence of a stationary distribution of this Markov chain need to be studied to deduce the geometric divergence of the CSA and have a rigorous mathematical framework to perform Monte Carlo simulations allowing to study the influence of different parameters of the algorithm. We start however by

illustrating in details the methodology on the simpler case where the step-size is constant. We deduce in this case the divergence at a constant speed. We keep–due to some space limitation–the details of the generalization to the CSA study for a future publication and give only a sketch of the results.

This paper is organized as follows. In Section II we define the $(1, \lambda)$-ES using resampling and the problem. In Section III we provide some preliminary derivations on the distributions that come into play for the analysis. In Section IV we analyze the constant step-size case: exhibit the Markov chain, prove its stability and deduce the divergence of the $(1, \lambda)$-ES on the constraint problem. In Section V we sketch out our results when the step-size is adapted using cumulative step-size adaptation. Finally we discuss our results and our methodology in Section VI.

Notations

Throughout this article, we denote by φ the density function of the standard multivariate normal distribution, and Φ the cumulative distribution function of a standard univariate normal distribution. The standard (unidimensional) normal distribution is denoted $\mathcal{N}(0, 1)$, the (n-dimensional) multivariate normal distribution with covariance matrix identity is denoted $\mathcal{N}(\mathbf{0}, \mathrm{Id}_n)$ and the i^{th} order statistic of λ i.i.d. standard normal random variables is denoted $\mathcal{N}_{i:\lambda}$. The uniform distribution on an interval I is denoted $\mathcal{U}_I$. We denote μ_{Leb} the Lebesgue measure. The set of natural numbers (including 0) is denoted $\mathbb{N}$, and the set of real numbers $\mathbb{R}$. We denote $\mathbb{R}_+$ the set $\{x \in \mathbb{R} | x \geq 0\}$, and for $A \subset \mathbb{R}^n$, the set A^* denotes $A \backslash \{\mathbf{0}\}$ and $\mathbf{1}_A$ denotes the indicator function of A. For two vectors $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^n$, we denote $[\mathbf{x}]_i$ the i^{th}-coordinate of $\mathbf{x}$, and $\mathbf{x}.\mathbf{y}$ the scalar product of $\mathbf{x}$ and $\mathbf{y}$. Take $(a, b) \in \mathbb{N}^2$ with $a \geq b$, we denote $[a..b]$ the interval of integers between a and b. For a topological set $\mathcal{X}$, $\mathcal{B}(\mathcal{X})$ denotes the Borel algebra of $\mathcal{X}$. For $\mathbf{X}$ and $\mathbf{Y}$ two random vectors, we denote $\mathbf{X} \overset{d}{=} \mathbf{Y}$ if $\mathbf{X}$ and $\mathbf{Y}$ are equal in distribution. For $(X_t)_{t \in \mathbb{N}}$ a sequence of random variables and X a random variable we denote $X_t \overset{a.s.}{\to} X$ if X_t converges almost surely to X and $X_t \overset{P}{\to} X$ if X_t converges in probability to X.

II. Problem Statement and Algorithm Definition

A. $(1, \lambda)$-ES with resampling

In this paper, we study the behaviour 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 $X_{\text{feasible}} = \{\mathbf{x} \in \mathbb{R}^n | g(\mathbf{x}) \geq 0\}$. To handle the constraint, the algorithm resamples any unfeasible solution until a feasible solution is found.

From iteration $t \in \mathbb{N}$, given the vector $\mathbf{X}_t \in \mathbb{R}^n$ and step-size $\sigma_t \in \mathbb{R}_+^*$, the 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 $g(\mathbf{Y}_t^i) < 0$, then it is resampled until it lies within the feasible domain. The first feasible i^{th} candidate solution is denoted $\tilde{\mathbf{Y}}_t^i$ and the realization of the multivariate normal distribution giving $\tilde{\mathbf{Y}}_t^i$ is $\tilde{\mathbf{N}}_t^i$, which is called a feasible step. Note that $\tilde{\mathbf{N}}_t^i$ is not distributed as a multivariate normal distribution, further details on its distribution are given later on.

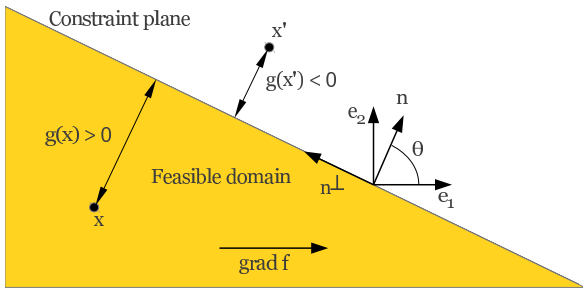

Fig. 1. Linear function with a linear constraint, in the plane generated by ∇f and $\mathbf{n}$, a normal vector to the constraint hyperplane with angle $\theta \in (0, \pi/2)$ with ∇f. The point $\mathbf{x}$ is at distance $g(\mathbf{x})$ from the constraint.

We define $\star = \underset{i \in [1..\lambda]}{\operatorname{argmax}} f(\tilde{\mathbf{Y}}_t^i)$ as the index realizing the maximum objective function, and call $\tilde{\mathbf{N}}_t^\star$ the selected step. The vector $\mathbf{X}_t$ is then updated as the solution realizing the maximum value of the objective function, i.e.

$$\mathbf{X}_{t+1} = \tilde{\mathbf{Y}}_t^\star = \mathbf{X}_t + \sigma_t \tilde{\mathbf{N}}_t^\star \ . \tag{2}$$

The step-size and other internal parameters are then adapted. We denote for the moment in a non specific manner the adaptation as

$$\sigma_{t+1} = \sigma_t \xi_t \tag{3}$$

where ξ_t is a random variable whose distribution is a function of the selected steps $(\tilde{\mathbf{N}}_i^\star)_{i \leq t}$. We will define later on specific rules for this adaptation.

B. Linear fitness function with linear constraint

In this paper, we consider the case where f, the function that we optimize, and g, the constraint, are linear functions. W.l.o.g., we assume that $\|\nabla f\| = \|\nabla g\| = 1$. We denote $\mathbf{n} := -\nabla g$ a vector normal 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

$$f(\mathbf{x}) = [\mathbf{x}]_1 \tag{4}$$

and the vector $\mathbf{e}_2$ lives in the plane generated by ∇f and $\mathbf{n}$ and is such that the angle between $\mathbf{e}_2$ and $\mathbf{n}$ is positive. We define θ the angle between ∇f and $\mathbf{n}$, and restrict our study to $\theta \in (0, \pi/2)$. The function g can be seen as a signed distance to the linear constraint as

$$g(\mathbf{x}) = \mathbf{x}.\nabla g = -\mathbf{x}.\mathbf{n} = -[\mathbf{x}]_1 \cos\theta - [\mathbf{x}]_2 \sin\theta \ . \tag{5}$$

A point is feasible if and only if $g(\mathbf{x}) \geq 0$ (see Figure 1). Overall the problem reads

$$\begin{aligned} &\text{maximize} \quad f(\mathbf{x}) = [\mathbf{x}]_1 \quad \text{subject to} \\ &g(\mathbf{x}) = -[\mathbf{x}]_1 \cos\theta - [\mathbf{x}]_2 \sin\theta \geq 0 \ . \end{aligned} \tag{6}$$

Although $\tilde{\mathbf{N}}_t^i$ and $\tilde{\mathbf{N}}_t^\star$ are in $\mathbb{R}^n$, due to the choice of the coordinate system and the independence of the sequence $([\mathbf{N}_t^i]_k)_{k\in[1..n]}$, only the two first coordinates of these vectors are affected by the resampling implied by g and the selection according to f. Therefore $[\tilde{\mathbf{N}}_t^\star]_k \sim \mathcal{N}(0,1)$ for $k \in [3..n]$. With an abuse of notations, the vector $\tilde{\mathbf{N}}_t^i$ will denote the 2-dimensional vector $([\tilde{\mathbf{N}}_t^i]_1, [\tilde{\mathbf{N}}_t^i]_2)$, likewise $\tilde{\mathbf{N}}_t^\star$ will also denote the 2-dimensional vector $([\tilde{\mathbf{N}}_t^\star]_1, [\tilde{\mathbf{N}}_t^\star]_2)$, and $\mathbf{n}$ will denote the 2-dimensional vector $(\cos\theta, \sin\theta)$. The coordinate system will also be used as $(\mathbf{e}_1, \mathbf{e}_2)$ only.

Following [2], [3], [4], we denote the normalized signed distance to the constraint as δ_t, that is

$$\delta_t = \frac{g(\mathbf{X}_t)}{\sigma_t} \ . \tag{7}$$

We initialize the algorithm by choosing $\mathbf{X}_0 = -\mathbf{n}$ and $\sigma_0 = 1$, which implies that $\delta_0 = 1$.

III. 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.

A. Feasible steps

The random vector $\tilde{\mathbf{N}}_t^i$, the i^{th} feasible step, is distributed as the standard multivariate normal distribution truncated by the constraint, as stated in the following lemma.

Lemma 1: Let a $(1,\lambda)$-ES with resampling optimize a function f under a constraint function g. If g is a linear form determined by a vector $\mathbf{n}$ as in (5), then the distribution of the feasible step $\tilde{\mathbf{N}}_t^i$ only depends on the normalized distance to the constraint δ_t and its density given that δ_t equals δ reads

$$p_\delta(\mathbf{x}) = \frac{\varphi(\mathbf{x})\mathbf{1}_{\mathbb{R}_+}(\delta - \mathbf{x}.\mathbf{n})}{\Phi(\delta)} \ . \tag{8}$$

Proof: A solution $\mathbf{Y}_t^i$ is feasible if and only if $g(\mathbf{Y}_t^i) \geq 0$, which is equivalent to $-(\mathbf{X}_t + \sigma_t \mathbf{N}_t^i).\mathbf{n} \geq 0$. Hence dividing by σ_t, a solution is feasible if and only if $\delta_t = -\mathbf{X}_t.\mathbf{n}/\sigma_t \geq \mathbf{N}_t^i.\mathbf{n}$. 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

$$\Pr(\mathcal{N}(0,1) \leq \delta_t) = \Phi(\delta_t) \ .$$

Therefore the density probability 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. (8). ■

Then the marginal density function $p_{1,\delta}$ of $[\tilde{\mathbf{N}}_t^i]_1$ can be computed by integrating Eq. (8) over $[\mathbf{x}]_2$ and reads

$$p_{1,\delta}(x) = \varphi(x)\frac{\Phi\left(\frac{\delta - x\cos\theta}{\sin\theta}\right)}{\Phi(\delta)} \ , \tag{9}$$

(see [2, Eq. 4] for details) and we denote $F_{1,\delta}$ its cumulative distribution function.

It will be important in the sequel to be able to express the vector $\tilde{\mathbf{N}}_t^i$ as a function of δ_t and of a *finite* number of random samples. Hence we give an alternative way to sample $\tilde{\mathbf{N}}_t^i$ rather than the resampling technique that involves an unbounded number of samples.

Lemma 2: Let a $(1,\lambda)$-ES with resampling optimize a function f under a constraint function g, where g is a linear form determined by a vector $\mathbf{n}$ as in (5). Let the feasible step $\tilde{\mathbf{N}}_t^i$ be the random vector described in Lemma 1 and $\mathbf{Q}$ be the 2-dimensional rotation matrix of angle θ. Then

$$\tilde{\mathbf{N}}_t^i \overset{d}{=} \tilde{F}_{\delta_t}^{-1}(U_t^i)\mathbf{n} + \mathcal{N}_t^i\mathbf{n}^\perp = \mathbf{Q}^{-1}\begin{pmatrix}\tilde{F}_{\delta_t}^{-1}(U_t^i)\\ \mathcal{N}_t^i\end{pmatrix} \tag{10}$$

where $\tilde{F}_{\delta_t}^{-1}$ denotes the generalized inverse of the cumulative distribution of $\tilde{\mathbf{N}}_t^i.\mathbf{n}$ [1], $U_t^i \sim \mathcal{U}_{[0,1]}$, $\mathcal{N}_t^i \sim \mathcal{N}(0,1)$ with $(U_t^i)_{i\in[1..\lambda],t\in\mathbb{N}}$ i.i.d. and $(\mathcal{N}_t^i)_{i\in[1..\lambda],t\in\mathbb{N}}$ i.i.d. random variables.

Proof: We define a new coordinate system $(\mathbf{n}, \mathbf{n}^\perp)$ (see Figure 1). It is the image of $(\mathbf{e}_1, \mathbf{e}_2)$ by $\mathbf{Q}$. In the new basis $(\mathbf{n}, \mathbf{n}^\perp)$, only the coordinate along $\mathbf{n}$ is affected by the resampling. Hence the random variable $\tilde{\mathbf{N}}_t^i.\mathbf{n}$ follows a truncated normal distribution with cumulative distribution function $\tilde{F}_{\delta_t}$ equal to $\min(1, \Phi(x)/\Phi(\delta_t))$, while the random variable $\tilde{\mathbf{N}}_t^i.\mathbf{n}^\perp$ follows an independent standard normal distribution, hence $\tilde{\mathbf{N}}_t^i \overset{d}{=} (\tilde{\mathbf{N}}_t^i.\mathbf{n})\mathbf{n} + \mathcal{N}_t^i\mathbf{n}^\perp$. Using the fact that if a random variable has a cumulative distribution F, then for F^{-1} the generalized inverse of F, $F^{-1}(U)$ with $U \sim \mathcal{U}_{[0,1]}$ has the same distribution as this random variable, we get that $\tilde{F}_{\delta_t}^{-1}(U_t^i) \overset{d}{=} \tilde{\mathbf{N}}_t^i.\mathbf{n}$, so we obtain Eq. (10). ■

We now extend our study to the selected step $\tilde{\mathbf{N}}_t^\star$.

B. Selected step

The selected step $\tilde{\mathbf{N}}_t^\star$ is chosen among the different feasible steps $(\tilde{\mathbf{N}}_t^i)_{i\in[1..\lambda]}$ to maximize the function f, and has the density described in the following lemma.

Lemma 3: Let a $(1,\lambda)$-ES with resampling optimize the problem (6). Then the distribution of the selected step $\tilde{\mathbf{N}}_t^\star$ only depends on the normalized distance to the constraint δ_t and its density given that δ_t equals δ reads

$$\begin{aligned} p_\delta^\star(\mathbf{x}) &= \lambda p_\delta(\mathbf{x})\, F_{1,\delta}([\mathbf{x}]_1)^{\lambda-1} \ , \\ &= \lambda\frac{\varphi(\mathbf{x})\mathbf{1}_{\mathbb{R}_+}(\delta - \mathbf{x}.\mathbf{n})}{\Phi(\delta)}\left(\int_{-\infty}^{[\mathbf{x}]_1}\varphi(u)\frac{\Phi(\frac{\delta - u\cos\theta}{\sin\theta})}{\Phi(\delta)}\mathrm{d}u\right)^{\lambda-1} \end{aligned} \tag{11}$$

where p_δ is the density of $\tilde{\mathbf{N}}_t^i$ given that $\delta_t = \delta$ given in Eq. (8) and $F_{1,\delta}$ the cumulative distribution function of $[\tilde{\mathbf{N}}_t^i]_1$ whose density is given in Eq. (9) and $\mathbf{n}$ the vector $(\cos\theta, \sin\theta)$.

Proof: The function f being linear, the rankings on $(\tilde{\mathbf{N}}_t^i)_{i\in[1..\lambda]}$ corresponds to the order statistic on

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

$([\tilde{\mathbf{N}}_t^i]_1)_{i\in[1..\lambda]}$. If we look at the joint cumulative distribution $F_\delta^\star$ of $\tilde{\mathbf{N}}_t^\star$

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

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

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

Deriving $F_\delta^\star$ on x and y yields the density of $\tilde{\mathbf{N}}_t^\star$ of Eq. (11). ■

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 the problem (6). Then the marginal distribution of $[\tilde{\mathbf{N}}_t^\star]_1$ only depends of δ_t and its density given that δ_t equals δ reads

$$\begin{aligned}p_{1,\delta}^\star(x) &= \lambda p_{1,\delta}(x) F_{1,\delta}(x)^{\lambda-1} \ , \qquad (12)\\ &= \lambda\varphi(x)\frac{\Phi\left(\frac{\delta - x\cos\theta}{\sin\theta}\right)}{\Phi(\delta)} F_{1,\delta}(x)^{\lambda-1} \ ,\end{aligned}$$

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

$$p_{2,\delta}^\star(y) = \lambda\frac{\varphi(y)}{\Phi(\delta)}\int_{-\infty}^{\frac{\delta - y\sin\theta}{\cos\theta}} \varphi(u) F_{1,\delta}(u)^{\lambda-1}\mathrm{d}u \ . \qquad (13)$$

Proof: Integrating Eq. (11) directly yields Eq. (12).

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. (11) 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. (8) 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 $\delta - x\cos\theta - y\sin\theta \geq 0$ is equivalent to $x \leq (\delta - y\sin\theta)/\cos\theta$, hence Eq. (13) holds. ■

We will need in the next sections 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. To do so, using notations of Lemma 2, we define the function $\tilde{\mathcal{G}} : \mathbb{R}_+ \times ([0,1]\times\mathbb{R}) \to \mathbb{R}^2$ as

$$\tilde{\mathcal{G}}(\delta, \mathbf{w}) = \mathbf{Q}^{-1}\begin{pmatrix} \tilde{F}_\delta^{-1}([\mathbf{w}]_1) \\ [\mathbf{w}]_2 \end{pmatrix} \ . \qquad (14)$$

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. $\mathcal{G}(\delta,(U,\mathcal{N}))$) 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

$$\mathcal{G}(\delta, (\mathbf{w}_i)_{i\in[1..\lambda]}) = \underset{\mathbf{N}\in\left\{\tilde{\mathcal{G}}(\delta,\mathbf{w}_i)|i\in[1..\lambda]\right\}}{\operatorname{argmax}} f(\mathbf{N}) \ . \qquad (15)$$

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 $\mathcal{G}(\delta, \mathcal{W}_t)$ 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. (6), 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

$$\tilde{\mathbf{N}}_t^\star \overset{d}{=} \mathcal{G}(\delta_t, \mathcal{W}_t) \ , \qquad (16)$$

where the function $\mathcal{G}$ is defined in Eq. (15).

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. (14), $\tilde{\mathbf{N}}_t^i \overset{d}{=} \tilde{\mathcal{G}}(\delta_t, \mathbf{W}_t^i)$, 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 (15) is $\mathcal{G}(\delta_t, \mathcal{W}_t)$. ■

IV. Constant step-size case

We illustrate in this section our methodology analysis on the simple case where the step-size is constantly equal to σ and prove that then $(\mathbf{X}_t)_{t\in\mathbb{N}}$ diverges almost surely at constant speed (Theorem 1). The analysis of the CSA will then be a generalisation of the results presented here, with a few more technical results to derive.

As suggested in [2], the sequence $(\delta_t)_{t\in\mathbb{N}}$ plays a central role for the analysis, and we will show that it admits a stationary measure. We first prove that this sequence is an homogeneous Markov chain.

Proposition 2: Consider the $(1,\lambda)$-ES with resampling and with constant step-size σ optimizing the constraint problem (6). Then the sequence $\delta_t = g(\mathbf{X}_t)/\sigma$ is an homogeneous Markov chain on $\mathbb{R}_+$ and

$$\delta_{t+1} = \delta_t - \tilde{\mathbf{N}}_t^\star.\mathbf{n} \overset{d}{=} \delta_t - \mathcal{G}(\delta_t, \mathcal{W}_t).\mathbf{n} \ , \qquad (17)$$

where $\mathcal{G}$ is the function defined in (15) and $(\mathcal{W}_t)_{t\in\mathbb{N}} = (\mathbf{W}_t^i)_{i\in[1..\lambda],t\in\mathbb{N}}$ is an i.i.d. sequence with $\mathbf{W}_t^i \sim (\mathcal{U}_{[0,1]}, \mathcal{N}(0,1))$ for all $(i,t) \in [1..\lambda]\times\mathbb{N}$.

Proof: It follows from the definition of δ_t that $\delta_{t+1} = \frac{g(\mathbf{X}_{t+1})}{\sigma_{t+1}} = \frac{-\left(\mathbf{X}_t + \sigma\tilde{\mathbf{N}}_t^\star\right).\mathbf{n}}{\sigma} = \delta_t - \tilde{\mathbf{N}}_t^\star.\mathbf{n}$, and in Proposition 1 we state that $\tilde{\mathbf{N}}_t^\star \overset{d}{=} \mathcal{G}(\delta_t, \mathcal{W}_t)$. Since δ_{t+1} has the same distribution as a time independent function of δ_t and of $\mathcal{W}_t$ where $(\mathcal{W}_t)_{t\in\mathbb{N}}$ are i.i.d., it is an 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:

$$\frac{[\mathbf{X}_t-\mathbf{X}_0]_1}{t} = \frac{1}{t}\sum_{k=0}^{t-1}[\mathbf{X}_{k+1}]_1 - [\mathbf{X}_k]_1$$
$$= \frac{\sigma}{t}\sum_{k=0}^{t-1}[\tilde{\mathbf{N}}_k^\star]_1 \overset{d}{=} \frac{\sigma}{t}\sum_{k=0}^{t-1}[\mathcal{G}(\delta_k,\mathcal{W}_k)]_1 \ . \quad (18)$$

The latter term suggests the use of a Law of Large Numbers (LLN) to prove the convergence of $\frac{[\mathbf{X}_t-\mathbf{X}_0]_1}{t}$ which will in turn imply–if the limit is positive–the divergence of $f(\mathbf{X}_t)$ at a constant rate. Sufficient conditions on a Markov chain to be able to apply the LLN include the existence of an invariant probability measure π. The limit term is then expressed as an expectation over the stationary distribution. More precisely, assume the LLN can be applied, the following limit will hold

$$\lim_{t\to\infty}\frac{[\mathbf{X}_t-\mathbf{X}_0]_1}{t} = \sigma\int_{\mathbb{R}_+}\mathbf{E}\left([\mathcal{G}(\delta,\mathcal{W})]_1\right)\pi(d\delta) \quad (19)$$
$$= \lim_{t\to\infty}\mathbf{E}_{\delta_0\sim\mu}\left([\mathbf{X}_{t+1}]_1-[\mathbf{X}_t]_1\right) \ , \quad (20)$$

with μ any initial distribution. The latter term corresponds to the limit of the progress rate (see [2, Eq. 2]). The invariant measure π is also underlying the study carried out in [2, Section 4] where more precisely it is stated: *"Assuming for now that the mutation strength σ is held constant, when the algorithm is iterated, the distribution of δ-values tends to a stationary limit distribution."*. We will now provide a formal proof that indeed $(\delta_t)_{t\in\mathbb{N}}$ admits a stationary limit distribution π, as well as prove some other useful properties that will allow us in the end to conclude to the divergence of $(f(\mathbf{X}_t))_{t\in\mathbb{N}}$.

A. 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 $\Pr(\delta_{t+1}\in A|\delta_t=\delta) = \Pr(\delta_t - \tilde{\mathbf{N}}_t^\star.\mathbf{n}\in A|\delta_t=\delta)$,

$$P(\delta,A) = \int_{\mathbb{R}^2}\mathbf{1}_A\left(\delta-\mathbf{u}.\mathbf{n}\right)p_\delta^\star\left(\mathbf{u}\right)\mathrm{d}\mathbf{u} \quad (21)$$

where $p_\delta^\star$ is the density of $\tilde{\mathbf{N}}_t^\star$ given in (11). For $t\in\mathbb{N}^*$, the t-step transition kernel P^t is defined by $P^t(\delta,A) := \Pr(\delta_t\in A|\delta_0=\delta)$.

From the transition kernel, we will now derive the first properties on the Markov chain $(\delta_t)_{t\in\mathbb{N}}$. First of all we investigate the so-called ψ-irreducible property.

A Markov chain $(\delta_t)_{t\in\mathbb{N}}$ on a state space $\mathbb{R}_+$ is *ψ-irreducible* if there exists a non-trivial measure ψ such that for all set $A\in\mathcal{B}(\mathbb{R}_+)$ with $\psi(A)>0$ and for all $\delta\in\mathbb{R}_+$, there exists $t\in\mathbb{N}^*$ such that $P^t(\delta,A)>0$. We denote $\mathcal{B}^+(\mathbb{R}_+)$ the set of Borel sets of $\mathbb{R}_+$ with strictly positive ψ-measure.

We also need the notion of *small sets*: a set $C\in\mathcal{B}(\mathbb{R}_+)$ is called a small set if there exists $m\in\mathbb{N}^*$ and a non trivial measure ν_m such that for all set $A\in\mathcal{B}(\mathbb{R}_+)$ and all $\delta\in C$

$$P^m(\delta,A)\geq\nu_m(A) \ . \quad (22)$$

If there exists C a ν_1-small set such that $\nu_1(C)>0$ then the Markov chain is said *strongly aperiodic*.

Proposition 3: Consider a $(1,\lambda)$-ES with resampling and with constant step-size optimizing the constraint problem (6) and let $(\delta_t)_{t\in\mathbb{N}}$ be the Markov chain exhibited in (17). Then $(\delta_t)_{t\in\mathbb{N}}$ is μ_{Leb}-irreducible, strongly aperiodic, and compact sets are small sets.

Proof: Using Eq. (21) and Eq. (11) 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

$$P(\delta,A) = \int_{\mathbb{R}}\int_{\mathbb{R}}1_A(u)g(\delta,u,v)\mathrm{d}v\mathrm{d}u \ . \quad (23)$$

For all δ,u,v the function $g(\delta,u,v)$ is strictly positive hence for all A with $\mu_{Leb}(A)>0$, $P(\delta,A)>0$. Hence $(\delta_t)_{t\in\mathbb{N}}$ is irreducible with respect to the Lebesgue measure.

In addition, the function $(\delta,u,v)\mapsto g(\delta,u,v)$ is continuous as the composition of continuous functions (the continuity of $\delta\mapsto F_{1,\delta}(x)$ for all x coming from the dominated convergence theorem). Given a compact C we hence know that there exists $g_C>0$ such that for all $(\delta,u,v)\in C\times[0,1]^2$, $g(\delta,u,v)\geq g_C>0$. Hence for all $\delta\in C$,

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

The measure ν_C being non-trivial, the previous equation shows that compact sets are small and that for C a compact such that $\mu_{Leb}(C\cap[0,1])>0$, we have $\nu_C(C)>0$ hence the chain is strongly aperiodic. ∎

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 $A\in\mathcal{B}(\mathbb{R}_+)$

$$\pi(A) = \int_{\mathbb{R}_+}P(\delta,A)\pi(\mathrm{d}\delta) \ . \quad (24)$$

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 set $A\in\mathcal{B}^+(\mathbb{R}_+)$ and for all $\delta\in\mathbb{R}_+$, $\Pr(\eta_A=\infty|\delta_0=\delta)=1$ where η_A is the occupation time of A, i.e. $\eta_A=\sum_{t=1}^\infty 1_A(\delta_t)$. We will show that the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is positive and Harris-recurrent by using so-called Foster-Lyapunov drift conditions: define the *drift* operator for a positive function V as

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

Drift conditions translate that outside a small set, the drift operator is negative. We will show a drift condition for V-geometric ergodicity where given a function $f \geq 1$, a positive and Harris-recurrent chain $(\delta_t)_{t\in\mathbb{N}}$ with invariant measure π is called *f-geometrically ergodic* if $\pi(f) < \infty$ and there exists $r_f > 1$ such that

$$\sum_{t\in\mathbb{N}} r_f^t \|P^t(\delta,\cdot) - \pi\|_f < \infty \ , \forall \delta \in \mathbb{R}_+ \ , \qquad (25)$$

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 V-geometric ergodicity, we will prove that there exists a small set C, constants $b \in \mathbb{R}$, $\epsilon \in \mathbb{R}_+^*$ and a function $V \geq 1$ finite for at least some $\delta_0 \in \mathbb{R}_+$ such that for all $\delta \in \mathbb{R}_+$

$$\Delta V(\delta) \leq -\epsilon V(\delta) + b\mathbf{1}_C(\delta) \ . \qquad (26)$$

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 [9, Theorem 15.0.1][2] as well as positive and Harris-recurrent[3].

Because compacts 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 constraint problem (6), and let $\mathcal{G}$ be the function defined in (15). We denote K and $\bar{K}$ the random variables $\exp(\mathcal{G}(\delta,\mathcal{W}).(a,b))$ and $\exp(a|[\mathcal{G}(\delta,\mathcal{W})]_1| + b|[\mathcal{G}(\delta,\mathcal{W})]_2|)$. For $\mathcal{W} \sim (\mathcal{U}_{[0,1]},\mathcal{N}(0,1))^\lambda$ and any $(a,b) \in \mathbb{R}^2$ $\lim_{\delta\to+\infty} \mathbf{E}(K) = \mathbf{E}(\exp(a\mathcal{N}_{\lambda:\lambda}))\mathbf{E}(\exp(b\mathcal{N}(0,1))) < \infty$ and $\lim_{\delta\to+\infty} \mathbf{E}(\bar{K}) < \infty$

For the proof see the appendix. We are now ready to prove a drift condition for geometric ergodicity.

Proposition 4: Consider a $(1,\lambda)$-ES with resampling and with constant step-size optimizing the constraint problem (6) and let $(\delta_t)_{t\in\mathbb{N}}$ be the Markov chain exhibited in (17). The Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is V-geometrically ergodic with $V : \delta \mapsto \exp(\alpha\delta)$ for $\alpha > 0$ small enough, and is Harris-recurrent and positive with invariant probability measure π.

Proof: Take the function $V : \delta \mapsto \exp(\alpha\delta)$ then $\Delta V(\delta) = \mathbf{E}\left(\exp\left(\alpha\left(\delta - \mathcal{G}(\delta,\mathcal{W}).\mathbf{n}\right)\right)\right) - \exp(\alpha\delta)$, $\frac{\Delta V}{V}(\delta) = \mathbf{E}\left(\exp\left(-\alpha\mathcal{G}(\delta,\mathcal{W}).\mathbf{n}\right)\right) - 1$. With Lemma 4 we obtain $\lim_{\delta\to+\infty} \mathbf{E}\left(\exp\left(-\alpha\mathcal{G}(\delta,\mathcal{W}).\mathbf{n}\right)\right) =$
$\mathbf{E}\left(\exp(-\alpha\mathcal{N}_{\lambda:\lambda}\cos\theta)\right)\mathbf{E}(\exp(-\alpha\mathcal{N}(0,1)\sin\theta)) < \infty$. 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 the limit equals to

$$\left(\sum_{i\in\mathbb{N}} \frac{(-\alpha\cos\theta)^i \mathbf{E}\left(\mathcal{N}_{\lambda:\lambda}^i\right)}{i!}\right)\left(\sum_{i\in\mathbb{N}} \frac{(-\alpha\sin\theta)^i \mathbf{E}\left(\mathcal{N}(0,1)^i\right)}{i!}\right) \ ,$$

which in turns yields

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

Since for $\lambda \geq 2$, $\mathbf{E}(\mathcal{N}_{\lambda:\lambda}) > 0$, for $\alpha > 0$ and small enough we get $\lim_{\delta\to+\infty} \frac{\Delta V}{V}(\delta) < -\epsilon < 0$. Hence there exists $\epsilon > 0$, $M > 0$ and $b \in \mathbb{R}$ such that

$$\Delta V(\delta) \leq -\epsilon V(\delta) + b\mathbf{1}_{[0,M]}(\delta) \ .$$

According to Proposition 3, $[0, M]$ is a small set, hence it is petite [9, Proposition 5.5.3]. Furthermore $(\delta_t)_{t\in\mathbb{N}}$ is a ψ-irreducible aperiodic Markov chain so $(\delta_t)_{t\in\mathbb{N}}$ satisfies the conditions of Theorem 15.0.1 from [9], which with Lemma 15.2.2, Theorem 9.1.8 and Theorem 14.0.1 of [9] proves the proposition. ∎

We now proved rigorously the existence (and unicity) of an invariant measure π for the Markov chain $(\delta_t)_{t\in\mathbb{N}}$, which provides the so-called steady state behaviour in [2, 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 [9, Theorem 17.1.7] in Eq (18) 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 constraint problem (6) and let $(\delta_t)_{t\in\mathbb{N}}$ be the Markov chain exhibited in (17). The sequence $([\mathbf{X}_t]_1)_{t\in\mathbb{N}}$ diverges in probability to $+\infty$ at constant speed, that is

$$\frac{[\mathbf{X}_t - \mathbf{X}_0]_1}{t} \xrightarrow[t\to+\infty]{P} \sigma\mathbf{E}_{\pi\times\mu_\mathcal{W}}\left([\mathcal{G}\left(\delta,\mathcal{W}\right)]_1\right) > 0 \ , \qquad (27)$$

with $\mathcal{G}$ defined in (15) and $\mathcal{W} = (\mathbf{W}^i)_{i\in[1..\lambda]}$ where $(\mathbf{W}^i)_{i\in[1..\lambda]}$ is an i.i.d. sequence such that $\mathbf{W}^i \sim (\mathcal{U}_{[0,1]},\mathcal{N}(0,1))$ 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 [9, Theorem 17.0.1] to $[\mathcal{G}]_1$ we only need $[\mathcal{G}]_1$ to be $\pi\times\mu_\mathcal{W}$-integrable.

With Fubini-Tonelli's theorem $\mathbf{E}_{\pi\times\mu_\mathcal{W}}(|[\mathcal{G}(\delta,\mathcal{W})]_1|)$ equals to $\mathbf{E}_\pi(\mathbf{E}_{\mu_\mathcal{W}}(|[\mathcal{G}(\delta,\mathcal{W})]_1|))$. As $\delta \geq 0$, we have $\Phi(\delta) \geq \Phi(0) = 1/2$, and for all $x \in \mathbb{R}$ as $\Phi(x) \leq 1$, $F_{1,\delta}(x) \leq 1$ and $\varphi(x) \leq \exp(-x^2/2)$ with Eq. (12) 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 (12) $|x|p^\star_{1,\delta}(x) \leq 2\lambda|x|\varphi(x)$, 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

[2]The condition $\pi(V) < \infty$ is given by [9, Theorem 14.0.1].

[3]The function V of (26) is unbounded off petite sets [9, Lemma 15.2.2], hence with [9, Theorem 9.1.8] the Markov chain is Harris-recurrent.

$\mathbf{E}_\pi(\mathbf{E}_{\mu_\mathcal{W}}(|[\mathcal{G}(\delta,\mathcal{W})]_1|)) \leq M < \infty$, meaning $[\mathcal{G}]_1$ is $\pi \times \mu_\mathcal{W}$-integrable. Hence we may apply the LLN on Eq. (18)

$$\frac{\sigma}{t}\sum_{k=0}^{t-1}[\mathcal{G}(\delta_k,\mathcal{W}_k)]_1 \xrightarrow[t\to+\infty]{a.s.} \sigma\mathbf{E}_{\pi\times\mu_\mathcal{W}}\left([\mathcal{G}(\delta,\mathcal{W})]_1\right) < \infty \ .$$

The equality in distribution in (18) allows us to deduce the convergence in probability of the left hand side of (18) to the right hand side of the previous equation.

As the measure π is an invariant measure for the Markov chain $(\delta_t)_{t\in\mathbb{N}}$, using (17), $\mathbf{E}_{\pi\times\mu_\mathcal{W}}(\delta) = \mathbf{E}_{\pi\times\mu_\mathcal{W}}(\delta - \mathcal{G}(\delta,\mathcal{W}).\mathbf{n})$, hence $\mathbf{E}_{\pi\times\mu_\mathcal{W}}(\mathcal{G}(\delta,\mathcal{W}).\mathbf{n}) = 0$ and thus

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

We see from Eq. (13) that for $y > 0$, $p^\star_{2,\delta}(y) < p^\star_{2,\delta}(-y)$ hence the expected value $\mathbf{E}_{\pi\times\mu_\mathcal{W}}([\mathcal{G}(\delta,\mathcal{W})]_2)$ is strictly negative. With the previous equation it implies that $\mathbf{E}_{\pi\times\mu_\mathcal{W}}([\mathcal{G}(\delta,\mathcal{W})]_1)$ is strictly positive.

■

We showed rigorously the divergence of $[\mathbf{X}_t]_1$ and gave an exact expression of the divergence rate, which is the limit of the progress rate defined in [2, Eq. (2)]. The fact that the chain $(\delta_t)_{t\in\mathbb{N}}$ is V-geometrically ergodic gives that $\sum_t r_V^t \|P^t(\delta,\cdot) - \pi\|_V < \infty$. This implies that the distribution π can be simulated efficiently by a Monte Carlo simulation allowing to have precise estimations of the divergence rate of $[\mathbf{X}_t]_1$. Assuming a CLT could be applied, confidence intervals on the Monte Carlo simulations could also be obtained.

A Monte Carlo simulation of the right hand side of Eq. (27) for 10^6 time steps gives the progress rate $\varphi^\star = \mathbf{E}([\mathbf{X}_{t+1} - \mathbf{X}_t]_1)$, which once normalized by σ and λ yields Fig. 2. We normalize per λ as in evolution strategies the cost of the algorithm is assumed to be the number of f-calls. We see that for small values of θ, the normalized serial progress rate assumes roughly $\varphi^\star/\lambda \approx \theta^2$. Only for larger constraint angles the serial progress rate depends on λ where smaller λ are preferable.

Fig. 3 is obtained through simulations of the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ defined in Eq. (17) 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 $\mathbf{E}_\pi(\delta_t) \to +\infty$ since the selection does not attract $\mathbf{X}_t$ towards the constraint anymore, while the resampling still repels $\mathbf{X}_t$ from the constraint. 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.

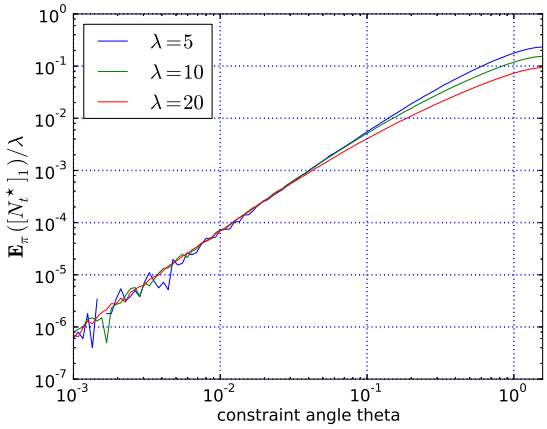

Fig. 2. Normalized progress rate $\varphi^\star = \mathbf{E}([\tilde{\mathbf{N}}_t^\star]_1)$ divided by λ for the $(1,\lambda)$-ES with constant step-size and resampling, plotted against the constraint angle θ, for $\lambda \in \{5, 10, 20\}$.

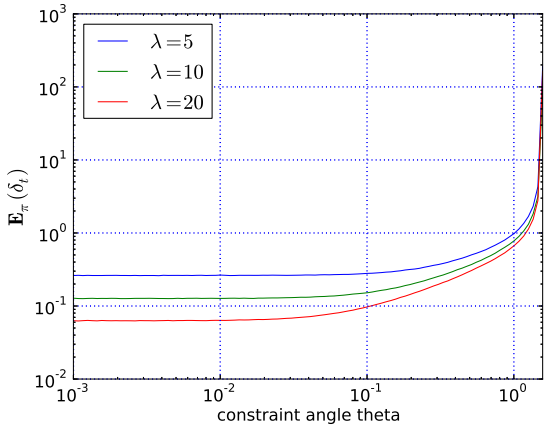

Fig. 3. Average normalized distance δ from the constraint for the $(1,\lambda)$-ES with constant step-size and resampling plotted against the constraint angle θ for $\lambda \in \{5, 10, 20\}$.

V. Cumulative Step-size Adaptation case

We generalise the previous results to the cumulative step-size adaptation mechanism. However due to space limitation we only sketch the results that we plan to present in details in an extended version of the paper. CSA introduces a new variable, $\mathbf{p}_t$, called the evolution path. It is a weighted recombination of the previous selected steps, where the weight of $\tilde{\mathbf{N}}_k^\star$ is proportional to $(1-c)^{t-1-k}$ with $c \in (0,1]$ being the cumulation parameter. For $c = 1$ the algorithm has "no memory" and the evolution path $\mathbf{p}_t$ is $\tilde{\mathbf{N}}_{t-1}^\star$. The step-size is adapted depending on the norm of $\mathbf{p}_t$ [8]. The Markov chain to study in this case is $(\delta_t, \mathbf{p}_t)_{t\in\mathbb{N}}$, except when $c = 1$ where it is $(\delta_t)_{t\in\mathbb{N}}$.

As in Section IV if the Markov chain is ψ-irreducible, aperiodic, and compact sets are small, then for $c = 1$ the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ is positive, Harris recurrent and V-geometrically ergodic, and a LLN can be applied on $\ln(\sigma_t/\sigma_0)$ to obtain that

$$\frac{1}{t}\ln\left(\frac{\sigma_t}{\sigma_0}\right) \xrightarrow[t\to\infty]{a.s.} \frac{\left(\mathbf{E}_{\pi_c\times\mu_\mathcal{W}}\left(\|\mathcal{G}\left(\delta,\mathcal{W}\right)\|^2\right) - 2\right)}{2d_\sigma n} \ , \qquad (28)$$

with π_c the stationary measure of $(\delta_t)_{t\in\mathbb{N}}$, $\mathcal{G}$ defined in (15), $\mathcal{W} = (\mathbf{W}^i)_{i\in[1..\lambda]}$ where $(\mathbf{W}^i)_{i\in[1..\lambda]}$ is an i.i.d. sequence such that $\mathbf{W}^i \sim (\mathcal{U}_{[0,1]}, \mathcal{N}(0,1))$ and $\mu_\mathcal{W}$ the probability measure of $\mathcal{W}$. So the step-size converges (resp. diverges) exponentially fast when the right hand side of Eq. (28) is strictly negative (resp. strictly positive).

VI. Discussion

We investigated the $(1, \lambda)$-ES with constant step-size optimizing a linear function under a linear constraint handled by resampling unfeasible solutions. We prove the stability (formally V-geometric ergodicity) of the Markov chain $(\delta_t)_{t\in\mathbb{N}}$ defined as the normalised distance to the constraint, which was *pressumed* in [2]. This property implies the divergence of the algorithm at a constant speed (see Theorem 1). In addition, it ensures (fast) convergence of Monte Carlo simulations of the divergence rate, justifying their use.

We believe that with the same approach, the CSA can be analysed. Simulations suggest that geometric divergence occurs for a small enough cumulation parameter, c, or large enough population size, λ. However, smaller values of the constraint angle seem to increase the difficulty of the problem arbitrarily, i.e. no given values for c and λ solve the problem for *every* $\theta \in (0, \pi/2)$.

Using a different covariance matrix to generate new samples can be interpreted as a change of the constraint angle. Therefore a correct adaptation of the covariance matrix will render the problem arbitrarily close to the one with $\theta = \pi/2$. The unconstrained linear function case has been shown to be solved by a $(1, \lambda)$-ES with cumulative step-size adaptation for a population size larger than 3, regardless of other internal parameters [5]. We believe this is a strong argument for using covariance matrix adaptation with ES when dealing with constraints, as pure step-size adaptation has been shown to be liable to fail on even a very basic problem.

This work provides a methodology that can be applied to many ES variants. It demonstrates that a rigorous analysis of the constrained problem can be achieved. It relies on the theory of Markov chains for a continuous state space that once again proves to be a natural theoretical tool for analysing ESs, complementing particularly well previous studies [2], [3], [4].

Acknowledgments

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

References

[1] Dirk V. Arnold. Analysis of a repair mechanism for the $(1, \lambda)$-ES applied to a simple constrained problem. In *Proceedings of the 13th annual conference on Genetic and evolutionary computation*, GECCO 2011, pages 853–860, New York, NY, USA, 2011. ACM.

[2] D.V. Arnold. On the behaviour of the (1,λ)-ES for a simple constrained problem. In *Foundations of Genetic Algorithms - FOGA 11*, pages 15–24. ACM, 2011.

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

[4] D.V. Arnold and D. Brauer. On the behaviour of the $(1 + 1)$-ES for a simple constrained problem. In G. Rudolph et al., editor, *Parallel Problem Solving from Nature - PPSN X*, pages 1–10. Springer, 2008.

[5] A. Chotard, A. Auger, and N. Hansen. Cumulative step-size adaptation on linear functions: Technical report. Technical report, Inria, 2012.

[6] Carlos A. Coello Coello. Constraint-handling techniques used with evolutionary algorithms. In *Proceedings of the 2008 GECCO conference companion on Genetic and evolutionary computation*, GECCO 2008, pages 2445–2466, New York, NY, USA, 2008. ACM.

[7] N. Hansen, S.P.N. Niederberger, L. Guzzella, and P. Koumoutsakos. A method for handling uncertainty in evolutionary optimization with an application to feedback control of combustion. *IEEE Transactions on Evolutionary Computation*, 13(1):180–197, 2009.

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

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

[10] Efrén Mezura Montes and Carlos A Coello Coello. A simple multimembered evolution strategy to solve constrained optimization problems. *Evolutionary Computation, IEEE Transactions on*, 9(1):1–17, 2005.

[11] Thomas P. Runarsson and Xin Yao. Stochastic ranking for constrained evolutionary optimization. *Evolutionary Computation, IEEE Transactions on*, 4(3):284–294, 2000.

Appendix

Proof of Lemma 4. *Proof:* From Proposition 1 the density probability function of $\mathcal{G}(\delta, \mathcal{W})$ is $p^\star_\delta$, and from Eq. (11)

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

From Eq. (9) $p_{1,\delta}(x) = \varphi(x)\Phi((\delta - x\cos\theta)/\sin\theta)/\Phi(\delta)$, so as $\delta \geq 0$ we have $1 \geq \Phi(\delta) \geq \Phi(0) = 1/2$, hence $p_{1,\delta}(x) \leq 2\varphi(x)$. So $p_{1,\delta}(x)$ converges when $\delta \to +\infty$ to $\varphi(x)$ while being bounded by $2\varphi(x)$ which is integrable. Therefore we can apply Lebesgue's dominated convergence theorem: $F_{1,\delta}$ converges to Φ when $\delta \to +\infty$ and is finite.

For $\delta \in \mathbb{R}_+$ and $(x, y) \in \mathbb{R}^2$ let $h_{\delta,y}(x)$ be $\exp(ax)p^\star_\delta((x, y))$. With Fubini-Tonelli's theorem $\mathbf{E}(\exp(\mathcal{G}(\delta, \mathcal{W}).(a, b))) = \int_{\mathbb{R}}\int_{\mathbb{R}} \exp(by)h_{\delta,y}(x)\mathrm{d}x\mathrm{d}y$. For $\delta \to +\infty$, $h_{\delta,y}(x)$ converges to $\exp(ax)\lambda\varphi(x)\varphi(y)\Phi(x)^{\lambda-1}$ while being dominated by $2\lambda\exp(ax)\varphi(x)\varphi(y)$, which is integrable. Therefore by the dominated convergence theorem and as the density of $\mathcal{N}_{\lambda:\lambda}$ is $x \mapsto \lambda\varphi(x)\Phi(x)^{\lambda-1}$, when $\delta \to +\infty$, $\int_{\mathbb{R}} h_{\delta,y}(x)\mathrm{d}x$ converges to $\varphi(y)\mathbf{E}(\exp(a\mathcal{N}_{\lambda:\lambda})) < \infty$.

So the function $y \mapsto \exp(by)\int_{\mathbb{R}} h_{\delta,y}(x)\mathrm{d}x$ converges to $y \mapsto \exp(by)\varphi(y)\mathbf{E}(\exp(a\mathcal{N}_{\lambda:\lambda}))$ while being dominated by $y \mapsto 2\lambda\varphi(y)\exp(by)\int_{\mathbb{R}}\exp(ax)\varphi(x)\mathrm{d}x$ which is integrable. Therefore we may apply the dominated convergence theorem: $\mathbf{E}(\exp(\mathcal{G}(\delta, \mathcal{W}).(a, b)))$ converges to $\int_{\mathbb{R}}\exp(by)\varphi(y)\mathbf{E}(\exp(a\mathcal{N}_{\lambda:\lambda}))\mathrm{d}y$ which equals to $\mathbf{E}(\exp(a\mathcal{N}_{\lambda:\lambda}))\mathbf{E}(\exp(b\mathcal{N}(0, 1)))$; and this quantity is finite.

The same reasoning gives that $\lim_{\delta\to\infty}\mathbf{E}(\bar{K}) < \infty$. ■